Beyond Social Media Analytics: Understanding Human Behaviour and Deep Emotion using Self Structuring Incremental Machine Learning

by

Tharindu Rukshan Bandaragoda

MPhil (Monash University, 2015)

BSc. Eng. (Honours) (University of Moratuwa, 2011)

A thesis submitted in total fulfilment of the requirements

for the degree of

Doctor of Philosophy

Research Centre for Data Analytics and Cognition

La Trobe Business School

College of Arts, Social Sciences and Commerce

La Trobe University

Victoria, Australia

February, 2020

© Copyright

by

Tharindu Rukshan Bandaragoda

2020

To Amma and Thaththa

Abstract

Recent advances in mobile and social media technologies has led to a rapid increase in online social interactions among individuals from all walks of life. For instance, by 2019, 48% of the world population were active on at least one online social media platform. Social media platforms archive all aspects of online social interactions. The data representing social interactions (or social data) can be used to analyse social behaviours and underlying causalities. Unlike the data collected conventionally from censuses or controlled experiments, this data is neither retrospectively collected nor controlled, and accumulated in large volumes leading to more accurate insights.

Existing techniques for gaining insights from social data are mainly adaptations of standard machine learning and natural language processing techniques. The performance of such techniques on social data is suboptimal as social data is highly unstructured, due to brevity and out of vocabulary terms, while also dynamic and bursty. Specialised techniques are required to handle issues related to the unstructured and dynamic nature of these data streams. This thesis aims to develop such techniques that can be used to gain in-depth insights from social data.

In the first phase of the thesis, a conceptual framework has been developed considering social data as representing the surface layer of a hierarchy of human social behaviours, needs and cognition. This framework is subsequently employed to transform social data into representations that preserve social behaviours and their causalities. In the second phase existing machine learning and natural language processing techniques have been extended to overcome the challenges of social data. Two platforms were built to capture insights from fast-paced and slow-paced social data. For fast-paced, a self structuring and incremental learning technique was developed to automatically capture salient topics, and corresponding dynamics over time. An event detection technique was developed to automatically monitor those identified topic pathways for significant fluctuations in social behaviours using multiple indicators such as volume and sentiment. The capabilities of this platform are demonstrated using two large datasets with over 1 million tweets each. The separated topic pathways were representative of the key topics/discussions of each entity and coherent against topic coherence measures. The identified events were validated against contemporary events reported in news.

Secondly for the slow-paced social data, a suite of new machine learning and natural language processing techniques were developed to automatically capture self-disclosed information of the individuals such as demographics, emotions and timeline of personal events of significance. This platform was trialled on a large text corpus of over four million posts collected from online support groups. This was further extended to transform prostate cancer related online support group discussions into a multidimensional representation and investigated the self-disclosed quality of life of patients (and partners) against time, demographics and clinical factors. The capabilities of this extended platform have been demonstrated using a text corpus collected from 10 prostate cancer online support groups comprising of 609,960 prostate cancer discussions and 22,233 patients.

Beyond Social Media Analytics: Understanding Human Behaviour and Deep Emotion using Self Structuring Incremental Machine Learning

Statement of Authorship

Except where reference is made in the text of the thesis, this thesis contains no material published elsewhere or extracted in whole or in part from a thesis accepted for the award of any other degree or diploma. No other person's work has been used without due acknowledgement in the main text of the thesis. This thesis has not been submitted for the award of any degree or diploma in any other tertiary institution.

Tharindu Rukshan Bandaragoda
February 25, 2020

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my primary supervisor Prof. Damminda Alahakoon for his continues support and guidance throughout my PhD journey. Thank you for all the inspiration, motivation, enthusiasm and immense wisdom shared with me. I could not have imagined having a better supervisor and a mentor to support my PhD journey.
Secondly, I would like to thank my co-supervisor Dr Daswin De Silva for his continues support guidance and feedback. Thank you for always encouraging me to write and publish; and keeping my research journey on track until the very end. Also thank you for all the collaboration opportunities provided.
This work was supported by an Australian Government Research Training Program Scholarship, a La Trobe University Full Fee Research Scholarship and a La Trobe University Postgraduate Scholarship. Also, I would like to thank Data to Decisions Cooperative Research Centre (D2D CRC) for providing the top-up scholarships.

I would like to thank Dr Buddhi Jayathilaka for introducing me to this research group and providing guidance and support during the initial phase of my PhD. Moreover, I would like to make this an opportunity to thank Dr Weranja Ranasinghe, A/Prof. Nathan Lawrentschuk and Prof. Damien Bolton for research collaboration and providing clinical expertise to explain the analytical findings of my research. Also, I would like to thank Dr Evgeny Osipov and Lule University of Technology, Sweden for research collaboration and supporting my research visit.
A big thank to Latrobe Business school and its support staff for all kinds of support.
I would like to thank all my colleagues at the Center for Data Analytics and Cognition for all the support extended and the cherish memories.
I would like to express my sincer gratitude to my parents who always encouraged me to take this PhD journey. Thank you amma and thaththa for all the support and encouragement throughout. None of this would be possible without you.
I was so fortunate to have the unconditional love and support of my wife Chammi throughout my PhD journey. Thank you all your motivation, encouragements and tolerance. Also, I would like to thank my little daughter, Amelie, for her cuddles and smiles that kept me going during the latter part of my PhD.
Finally, I would like to thank everyone who supported me during my PhD journey.

Tharindu Rukshan Bandaragoda

La Trobe University
Victoria, Australia
February 2020

Contents

Abstract . . . . . **iv**

Acknowledgments . . . . . **vi**

List of Tables . . . . . **xi**

List of Figures . . . . . **xiii**

1 Introduction **1**
1.1 Background 1
1.2 Social media 2
1.3 Motivation 7
1.4 Research objectives 9
1.5 Research questions 9
1.6 Research contributions 10
1.7 Thesis outline 11

2 Literature review and reflections **13**
2.1 Manual assessment of social data 13
2.2 Topic extraction 14
2.2.1 Text clustering 14
2.2.2 Topic modelling 15
2.2.3 Topic extraction from social data 16
2.3 Event detection 18
2.3.1 Unspecified event detection 18
2.3.2 Specified event detection 20
2.4 Emotion extraction from text 23
2.4.1 Theories of emotion 24
2.4.2 Modelling emotions 24
2.4.3 Sentiment detection from text 28
2.4.4 Emotion detection from text 31
2.5 Self structuring techniques 32
2.5.1 Biological inspiration 32
2.5.2 Self organising maps 33
2.5.3 Dynamically expanding self organising maps 35

2.6 Unsupervised incremental learning techniques . . . 36
2.7 The formidable challenges to social data analysis . . . 38
2.7.1 Scale of data . . . 38
2.7.2 Unlabelled data . . . 38
2.7.3 Time sensitivity . . . 39
2.7.4 Diversity . . . 39
2.7.5 Unstructured nature of data . . . 40
2.8 Chapter Summary . . . 41

3 The Intellection of a Conceptual Framework . . . 43
3.1 On the importance of social data . . . 44
3.2 The proposed conceptual framework . . . 47
3.2.1 Cognition . . . 47
3.2.2 Social needs . . . 48
3.2.3 Social capital . . . 50
3.2.4 Social behaviour . . . 51
3.2.5 Social actions . . . 60
3.2.6 Social data . . . 62
3.3 An embodiment in self-structuring artificial intelligence . . . 63
3.4 Chapter Summary . . . 65

4 Ideation to creation: A new self-structuring artificial intelligence algorithm . . . 67
4.1 The design of the algorithm . . . 67
4.2 Unsupervised incremental learning for topic separation . . . 70
4.3 IKASL algorithm . . . 71
4.3.1 Initialisation . . . 71
4.3.2 Self structuring . . . 71
4.3.3 Generalisation . . . 73
4.3.4 Learning from generalisation . . . 74
4.3.5 Topic separation with IKASL . . . 75
4.4 IKASL*text* . . . 75
4.4.1 Handling dynamic vocabularies . . . 75
4.4.2 Addressing the brevity issue . . . 77
4.4.3 Identifying new topic pathways . . . 78
4.5 Event detection . . . 80
4.5.1 Volume based event indicator . . . 81
4.5.2 Sentiment based event indicator . . . 81
4.6 Illustrative Example . . . 82
4.7 Experiments . . . 83
4.7.1 Datasets . . . 83
4.7.2 Pre-processing . . . 84
4.7.3 Feature extraction . . . 85

4.7.4 Topic separation . . . 85
4.7.5 Evolution of topic pathways . . . 89
4.7.6 Emergence of new topic pathways . . . 95
4.7.7 Automatic event detection . . . 96
4.8 Chapter Summary . . . 99

5 Expatiation from algorithm to technology platform . . . 101
5.1 Online Support Groups (OSG) . . . 102
5.2 Users of Online Support Groups . . . 102
5.2.1 Motivations to participate in OSG . . . 103
5.2.2 Role of professionals in OSG . . . 106
5.2.3 Opportunities for researchers in OSG . . . 106
5.3 Limitations in finding relevant information . . . 107
5.4 Proposed technology platform . . . 110
5.5 Information retrieval techniques . . . 112
5.5.1 Narrative type extraction . . . 112
5.5.2 Age extraction . . . 114
5.5.3 Gender extraction . . . 118
5.5.4 Medical concept extraction . . . 121
5.5.5 Emotion extraction . . . 121
5.6 OSG user timeline construction . . . 125
5.6.1 Clinical event timeline . . . 126
5.6.2 Emotion timeline . . . 128
5.7 Evaluation of the information extraction techniques . . . 129
5.7.1 Narrative type classification performance . . . 129
5.7.2 Age and gender resolution performance . . . 130
5.8 Demonstration . . . 130
5.8.1 OSG data collection . . . 131
5.8.2 Information extraction . . . 131
5.8.3 Personalised search to extract relevant information . . . 132
5.8.4 OSG analytics for researchers . . . 133
5.9 Chapter Summary . . . 135

6 A digital health platform . . . 137
6.1 Cancer, cancer burden and cancer care . . . 138
6.1.1 Unmet needs of cancer . . . 138
6.1.2 Understanding the psychological burden of cancer . . . 139
6.1.3 OSG as a complementary approach to understand the psychological burden of cancer . . . 140
6.2 Prostate cancer and related OSG usage . . . 140
6.2.1 Incidence . . . 140
6.2.2 Diagnose . . . 141
6.2.3 Treatment and side-effects . . . 141

6.2.4 Prostate cancer OSG . . . 142
6.3 PRIME . . . 143
6.3.1 Aggregating OSG posts by the patient . . . 145
6.3.2 Patient profile based extractions . . . 145
6.3.3 Patient timeline based extractions . . . 148
6.4 Prostate related OSG data collection . . . 149
6.4.1 Ethical considerations . . . 149
6.5 Analysis of the quality of life of prostate cancer patients . . . 150
6.5.1 Inclusion criteria . . . 151
6.5.2 Participant characteristics . . . 151
6.5.3 Emotions by age . . . 151
6.5.4 Emotions by modality . . . 153
6.5.5 Emotions by patient and partner . . . 153
6.5.6 Significant emotion terms . . . 154
6.6 Analysis of decision making behaviour and decision factors . . . 156
6.6.1 Inclusion criteria . . . 156
6.6.2 Decision making behaviour groups . . . 156
6.6.3 Decision factors . . . 158
6.6.4 Patient timeline analysis . . . 161
6.7 Discussion . . . 163
6.8 Chapter Summary . . . 164

7 Conclusion . . . 165
7.1 Summary of contributions . . . 165
7.2 Addressing the research questions . . . 168
7.3 Future directions . . . 170

List of Tables

1.1 Applications of social data based on recent literature. 8

3.1 A summary of the key theories related to social needs discussed in this section. 51
3.2 A summary of the key theories related to social behaviour discussed in this section. 60
3.3 Comparison of social behaviours in physical world and online social media platforms. 61

4.1 Notations used for the functional details of IKASL algorithm 71
4.2 Characteristics of the twitter datasets. 84
4.3 Frequent terms in the prominent topic pathways of #Obama dataset 86
4.4 Distribution (%) of top five frequent terms of six prominent topic pathways in #Obama dataset 87
4.5 Frequent terms in the prominent topic pathways of #Microsoft dataset . . . 88
4.6 Further examination of topic segments illustrated as word clouds in Figure 4.9 92
4.7 Two topic pathways from #Microsoft Dataset represented using high frequent terms of the topic pathway and high frequent terms of top five most voluminous topic segments 93
4.8 Two topic pathways from #Obama Dataset represented using high frequent terms of the topic pathway and high frequent terms of top five most voluminous topic segments 94
4.9 Top 10 events detected based on event score $\mathcal{I}$ for #Microsoft Dataset . . . 98
4.10 Two topic pathways from #Obama Dataset represented using high frequent terms of the topic pathway and high frequent terms of top five most voluminous topic segments 100

5.1 Population and volume statistics of three large OSG. Note that this information are as of 02/04/2018. 108
5.2 Examples for narrative type resolution from OSG posts 113
5.3 Positive and negative samples of age related phrases found in forum posts . 114
5.4 A Sample of features extracted from a text chunk 116
5.5 Few examples of narration based gender resolution process 120
5.6 Emotion categories and a sample of representative terms used for each emotion. 123

5.7 A sample of representative terms of each clinical event type. Note that some of the terms are common misspelled words. 128
5.8 Evaluation results of the narrative type extraction technique 129
5.9 Evaluation results of the age extraction technique 130
5.10 Evaluation results of the gender extraction technique 130
5.11 Counts and percentages of extracted age, gender and narrative type 132
5.12 Excerpts from the top five results obtained using the three search approaches (including the proposed method) to retrieve similar experiences for the query "I'm a 40 year old woman taking Nexium for heartburn" . . . 134
5.13 Top five symptoms co-exists with the symptom 'heartburn' in different age-groups 135

6.1 A sample of candidate terms for each template term used in the template sentences of decision making behaviour. Note that some of the terms are abbreviations or shorten versions of the actual terms or phrases. 147
6.2 Population and volume statistics of the ten selected OSG on prostate cancer discussions. 150
6.3 Demographics and clinical characteristics of the identified low-intermediate risk prostate cancer patients and partners of patients from the ten selected OCSGs. Note that, the percentages in each column is determined relative to the total participants in each cohort (total or each modality). The *p-value* presented in the last column is from chi-square tests that evaluated the statistical significance across the three modalities. 152
6.4 Frequently-used emotion terms for three selected cohorts: (i) patients aged <40, (ii) patients aged >70 and (iii) partners of patients. 154
6.5 The decision factors employed for the treatment decision against the primary treatment modality of the patient. 159
6.6 The decision factors employed for the treatment decision against the decision making behaviour group. 160

List of Figures

1.1 Recent trends in internet and social media platform usage 3
1.2 Chapter structure of the thesis 11

2.1 The process that generates emotions, according to several prominent emotion theories. 25
2.2 A graphical representation of the Circumplex model consisting of valence and arousal dimensions. Multiple emotions are placed in the axis system based on their valance and arousal. Image adapted from Posner et al. (2005). 26
2.3 Plutchiks wheel of emotions adopted from (Plutchik, 2001) 27
2.4 Self Organising Map(SOM) creates a low dimensional structure from a high dimensional input space while preserving the topological relations exist in high dimensional input space. Source: (Haykin, 1994) 33

3.1 The approaches of transforming social data: a. the conventional approach of using machine learning techniques on social data which yields data-driven insights, b. combining machine learning techniques with social theories to transform social data into insights representative of underlying social behaviours, c. using insights representative of underlying social behaviours to better understand human social behaviours. 46
3.2 Incorporating social science theories and machine learning techniques to generate insights from social data. 47
3.3 The conceptual framework for understanding social data using social theories of human social behaviours, social needs and cognition as the foundation. 48
3.4 The onion like nature of personality (Altman and Taylor, 1973), where outer layers contain more visible information and inner layers contain more intimate information. *image source: Wikipedia* 56
3.5 Accumulation of the traces of social behaviours in online social media platforms as social data. 62
3.6 The proposed self-structuring artificial intelligence framework for generating insights from social data in online social media platforms. 64

4.1 High-level design of the proposed platform for topic separation and event detection from a social text stream 69
4.2 Two node expansions scenarios 73

4.3 Layered structure of IKASL algorithm 74
4.4 Local *term frequency* of terms A and B across social media message batches $B^0, \ldots, B^8$. τ_V is the local *term frequency* threshold that determines inclusion of exclusion in the respective vocabulary 77
4.5 An illustrative example showing topic pathway separation and event detection 83
4.6 Volume (%) of tweets in six prominent topic pathways of #Obama Dataset 88
4.7 Volume (%) of tweets in six prominent topic pathways of #Microsoft Dataset 89
4.8 Topic coherence scores for prominent topic pathways and whole dataset (for baseline) in (a) #Obama Dataset and (b) #Microsoft Dataset. 90
4.9 Word clouds generated for nine topic segments in topic pathway $TP^1_{Microsoft}$, which represents Microsoft-Google relationship. Note that most frequent terms *Microsoft* and *Google* were removed from word clouds for improved clarity. 91
4.10 A word cloud and a word frequency graph generated for the topic pathway $TP^1_{Microsoft}$. Note that most frequent terms *Microsoft* and *Google* were removed from word clouds for improved clarity. 93
4.11 Four new topic pathways emerged in #Microsoft Dataset. Each topic pathway is represented using a chain of bubbles where each bubble is a topic segment. Diameter of a bubble is set to number of tweets in the topic segment. 95
4.12 Topic coherence of newly emerged topic pathways. 96

5.1 a. Home page of patient.info OSG, which shows the popular discussion groups and the simple search interface, b. A sample OSG discussion thread. 103
5.2 Stakeholders of Online Support Groups (OSG). 104
5.3 Topic hierarchies in (a) healingwell and (b) healthboards OSG. 108
5.4 The proposed multi-stage self-structuring AI platform and its benefits for the three stakeholder groups in OSG. 110
5.5 Age resolution process 115
5.6 A text chunk and its segments L, M, and R 116
5.7 Gender resolution process 119
5.8 The proposed technique for emotion extraction 122
5.9 Demographic distribution of the posts that mention symptom 'heartburn' . 135
5.10 Temporal distribution of the posts that mention symptom 'heartburn' over a three-year period 136

6.1 An anonymised sample (parts are omitted and rephrased to preserve privacy) of five posts by a prostate cancer patient. The highlighted excerpts are demographic, clinical, emotion expressions and decision making process related information, stated in the form of free-text. 143
6.2 Patient Reported Information Multidimensional Exploration (PRIME) platform, which contains a suite of machine learning and natural language processing based information extraction modules to process cancer related OSG data. 144

6.3 Positive and negative emotions by age group. 153
6.4 Positive and negative emotions by the treatment modality (RP, EBRT and AS). 154
6.5 Positive and negative emotions by patient and partner. 155
6.6 The distribution of decision making behaviour groups across: (a) Treatment modality, (b) Age group, and (c) Gleason score. Note that the percentage value for a behaviour group in a particular cohort is determined using the total (across the three behaviour groups) in each cohort as the denominator. 157
6.7 The OSG participation activities by the three decision making behaviour groups over the patient trajectory from 3 months pre-treatment to 12 months post-treatment. Note that an active participant during a particular month is an individual who has participated in OSG discussions during that month. The OSG posts are classified as an expression of experience or providing advice by the classifier presented in Section 5.5.1. 158
6.8 The positive and negative emotion trajectories of the three decision making behaviour groups over the patient trajectory from 3 months pre-treatment to 12 months post-treatment. 161
6.9 The self-disclosed side effect timeline aggregated monthly across the three decision making behaviour groups from 1st month post-treatment to 12 months post-treatment. 162

List of Publications

Journals

1. T. Bandaragoda, D. De Silva, and D. Alahakoon. Automatic event detection in microblogs using incremental machine learning. *Journal of the Association for Information Science and Technology*, 68(10):2394–2411, 2017

2. T. Bandaragoda, D. De Silva, D. Alahakoon, W. Ranasinghe, and D. Bolton. Text mining for personalized knowledge extraction from online support groups. *Journal of the Association for Information Science and Technology*, 69(12):1446–1459, 2018

3. T. Bandaragoda, W. Ranasinghe, A. Adikari, D. de Silva, N. Lawrentschuk, D. Alahakoon, R. Persad, and D. Bolton. The patient-reported information multidimensional exploration (PRIME) framework for investigating emotions and other factors of prostate cancer patients with low intermediate risk based on online cancer support group discussions. *Annals of surgical oncology*, 25(6):1737–1745, 2018

4. W. Ranasinghe, D. de Silva, T. Bandaragoda, A. Adikari, D. Alahakoon, R. Persad, N. Lawrentschuk, and D. Bolton. Robotic-assisted vs. open radical prostatectomy: A machine learning framework for intelligent analysis of patient-reported outcomes from online cancer support groups. *Urologic Oncology: Seminars and Original Investigations*, 36(12):529.e1–529.e9, 2018

5. D. De Silva, W. Ranasinghe, T. Bandaragoda, A. Adikari, N. Mills, L. Iddamalgoda, D. Alahakoon, N. Lawrentschuk, R. Persad, E. Osipov, et al. Machine learning to support social media empowered patients in cancer care and cancer treatment decisions. *PLoS One*, 13(10):e0205855, 2018

6. W. Ranasinghe, T. Bandaragoda, D. De Silva, and D. Alahakoon. A novel framework for automated, intelligent extraction and analysis of online support group discussions for cancer related outcomes. *BJU International*, 120:59–61, 2017

7. D. Nallaperuma, R. Nawaratne, T. Bandaragoda, A. Adikari, S. Nguyen, T. Kempitiya, D. De Silva, D. Alahakoon, and D. Pothuhera. Online incremental machine learning platform for big data-driven smart traffic management. *IEEE Transactions on Intelligent Transportation Systems*, 20(12):4679–4690, 2019

8. M. Sherwood, M. Lordanic, T. Bandaragoda, E. Sherry, and D. Alahakoon. A new league, new coverage? comparing tweets and media coverage from the first season of AFLW. Media International Australia, 172(1):114–130, 2019

9. T. Bandaragoda, A. Adikari,R. Nawaratne, D. Nallaperuma, A. K. Luhach, T. Kempitiya, S. Nguyen, D. Alahakoon, D. De Silva. Artificial intelligence based commuter behaviour profiling framework using Internet of things for real-time decision-making. *Neural Computing and Applications*, 1-15, 2020.

Conference proceedings

1. T. Bandaragoda, D. De Silva, D. Kleyko, E. Osipov, U. Wiklund, and D. Alahakoon. Trajectory clustering of road traffic in urban environments using incremental machine learning in combination with hyperdimensional computing. In 2019 IEEE Intelligent Transportation Systems Conference (ITSC), pages 1664–1670. IEEE, 2019

2. R. Nawaratne, T. Bandaragoda, A. Adikari, D. Alahakoon, D. De Silva, and X. Yu. Incremental knowledge acquisition and self-learning for autonomous video surveillance. In IECON 2017-43rd Annual Conference of the IEEE Industrial Electronics Society, pages 4790–4795. IEEE, 2017

Chapter 1

Introduction

> *Man is by nature a social animal; an individual who is unsocial naturally and not accidentally is either beneath our notice or more than human. Society is something that precedes the individual. Anyone who either cannot lead the common life or is so self-sufficient as not to need to, and therefore does not partake of society, is either a beast or a god.*
>
> Aristotle, Politics (˜384 BC)

This chapter initiates the thesis by postulating an expansive context for its research questions, objectives and contribution. From the early origins of human societies to contemporary habituation of digital environments, the manifestation of social behaviours and interactions has eventuated in diverse schools of thought and distinct domains of knowledge. The persistence of social data in digital environments has significantly transformed this scholarly pursuit. This thesis is an endorsement, empowerment and advancement of that pursuit. Let us begin.

1.1 Background

Humans are social animals. Social interactions continue to be paramount for modern humans, just as it was the determinant of survival of our first ancestors against natural selection, the Hominin hunter-gatherer groups (1.8-1.3 million years ago) (Darwin, 1859). Over the course of human civilisation, the level of sophistication of social interactions, as well as the number of social interactions among humans have advanced from simple survival to complex social needs. In modern humans, social interactions are controlled by cognitive processes in the brain, which perceive the social environment (social actions of others) based on sensory information and existing knowledge. Evolution has led to biological adaptations in brain, sensory and motor neurons to optimise our social interactions in the natural world (Kock, 2004, 2005). For instance, human ears are more sensitive to human voices than any other acoustic stimuli (Nass and Brave, 2005) and the human brain has a high volume ratio for the neocortex (relative to the entire brain) in order to handle the computational demand of social interactions (Dunbar, 1998). Moreover, a study on cognitive skills among chimpanzees, orangutans, and two year-old children uncovered that

although there is no significant difference in cognitive skills for physical activities, human children possess superior cognitive skills for social activities (Herrmann et al., 2007).

Social behaviour is a goal directed action to achieve a perceived goal (Ajzen, 1985). Social behaviours are formally defined as abstractions of different types of human social interactions that are executed to achieve similar perceived goals (Ajzen, 1985). Some examples of social behaviours are interpersonal communication, self disclosure, cooperation and social comparison. For instance, expressing emotions and discussion of political views are distinct social interactions. However, both interactions disclose different types of information i.e., emotions and political views about oneself to others in society. Therefore, such interactions can be abstracted the behaviour of self-disclosure. Developing an intricate understanding of social behaviours is crucial for all domains of knowledge and academic disciplines. This expands across humanities, social sciences and biological sciences more directly, and indirectly into physical sciences, formal sciences and applied sciences.

Conventional studies of human social behaviours rely on *surface data*: limited information on a large number of individuals or *deep data*: extensive information on few individuals (Manovich, 2011). *Surface data* is the population level data collected using censuses which collects data about individuals, households and businesses. Such datasets are well structured and often analysed using statistical methods. *Deep data* is collected from a small group of individuals using controlled and natural experiments (Neuman, 2013). Such datasets are detailed and unstructured, and analysed using qualitative methods. Both these data collection approaches have merits and long served as the key approaches to study social behaviours. However, the population level *surface data* is often too shallow with few attributes about each individual, thus leads to over-generalised insights that are only valid at the population level. Also, such studies are costly and time consuming, hence, not frequent (e.g., US census is every 10 years).*Deep data* from social experiments are limited to few individuals and mostly obtained using controlled environments. As pointed-out by Meshi et al. (2015), the major issues of *deep data* are lack of generalisability to a larger population, lack external validity due to differences in controlled experiment environment and natural environment, and various biases in data collection such as *recall bias*.

Given these inherent limitations in conventional methods of studying human social behaviours, it is timely that recent technological advancements have collectively led to the invention and development of new digital environments for enriched social interactions, that also encapsulate the full spectrum of social behaviours. It is opportune that each interaction and behaviour is digitally represented, captured and archived to facilitate much more meticulous approaches to this field of study.

1.2 Social media

These (established and emerging) digital environments are widely referred to as social media or social media platforms. Technological developments in affordability, availability

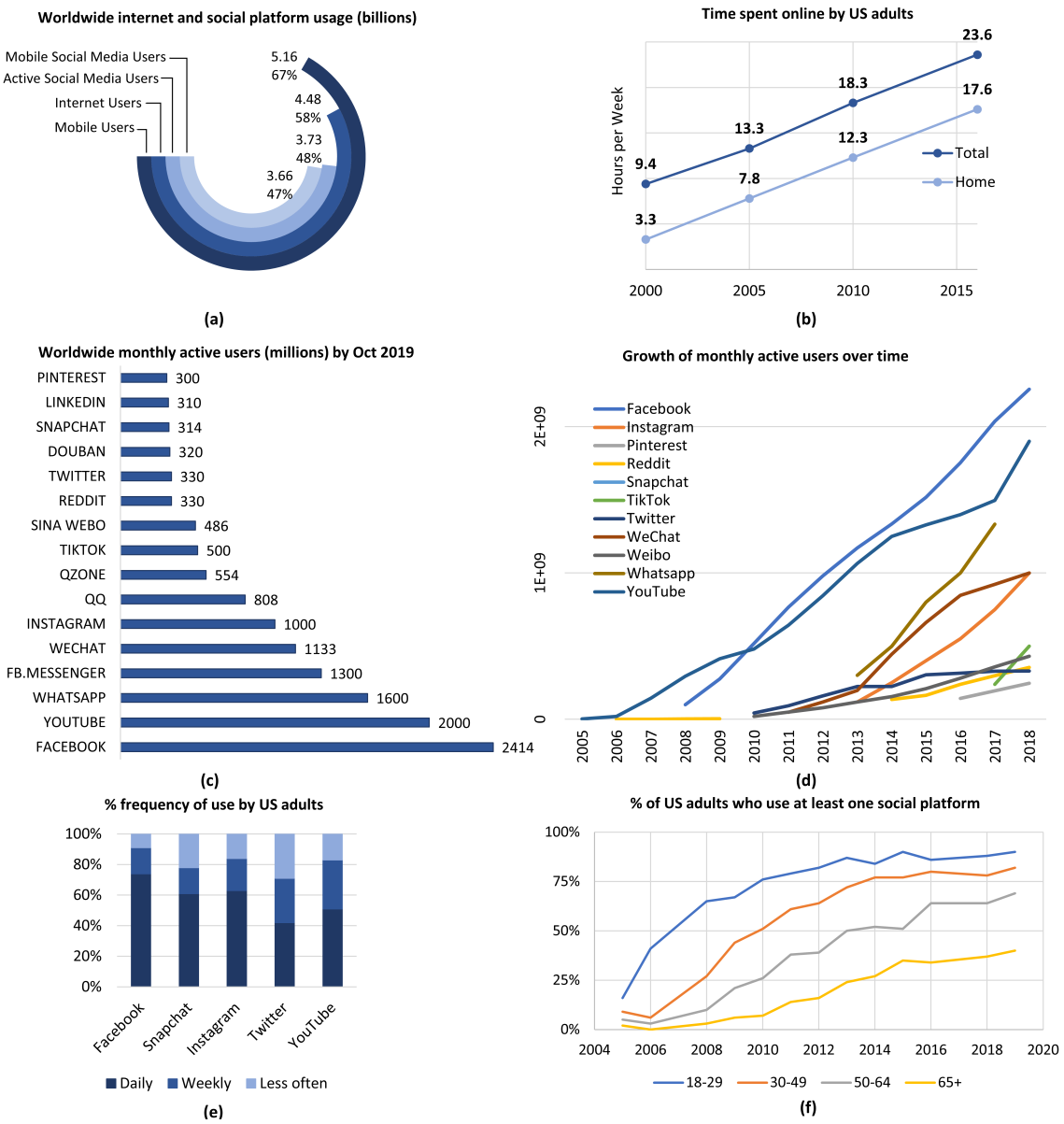


Figure 1.1: Recent trends in internet and social media platform usage:
(a) World internet and social media platform usage by fourth quarter of 2019 (source: (Kemp, 2019)),
(b) Time spent online by US adults from 2000 to 2016 (source: (Cole et al., 2017)),
(c) Worldwide monthly active users (MAU) in top online social media platforms by fourth quarter of 2019 (source: (Kemp, 2019)),
(d) Growth of worldwide monthly active users (MAU) in top online social media platforms from 2005 to 2018 (source: (Kemp, 2019)),
(e) Percentage frequency use of different online social media platforms by US adults in 2019 (source: (Pew Research Center, 2019), and
(f) Percentages of US adults by age group who use at least one social media platform from 2005 to 2019 (source: (Pew Research Center, 2019)))

and ease of use have reduced the digital divide, leading to an exponential increase in social media platforms for social interactions. As shown in Figure 1.1.a in 2019, worldwide mobile usage was 67% and internet usage was 58% which highlights worldwide adaptation

of mobile and internet technologies (Kemp, 2019). A recent survey of US adults (Cole et al., 2017) shows that the average weekly internet usage has risen from 9.4 hours in 2000 to 23.6 hours in 2016 (Figure 1.1.b). This increase was mainly due to domestic internet use for leisure and social activities, which has risen five-fold over the past 16 years (Cole et al., 2017). Although the functionality of modern online social media platforms are diverse, and often specialised to facilitate the demands of different socio-demographic segments, they are in general web and mobile based platforms that facilitate users to create, share and co-create multimodal (text, image, audio and video) user generated content as well as engage in online social discussions with other members on the platform (Kaplan and Haenlein, 2010; Kietzmann et al., 2011).

A vibrant history

Usenet (an online discussion system started in 1979) and Bulletin Board Systems (a dial-up internet based computer resource sharing system introduced in the late 1970s) are the likely precursors of modern day online social media platforms. The first recognisable online social media platform (SixDegrees.com) launched in 1997 (Boyd and Ellison, 2007) which allows the users to create profiles, friend lists and browse the profiles of their friends. As people started to flock into the internet in the early 2000s, there has been a rapid increase in the launch of diverse online social media platforms. As shown in Figure 1.1.a in 2019, out of the 7.73 billion of world population, 3.73 billion individuals are active users of at least one social media platform, which accounts for 48% of the world population. Figure 1.1.c depicts the worldwide Monthly Active User (MAU) counts in October 2019 of the top social media platforms. It shows that Facebook has 2.414 billion MAU which means 31% of the world population are active Facebook users. Similarly, video sharing platform YouTube has 2.0 billion MAU (25.9% of world population). Most popular messenger applications are WhatsApp (1.6 billion MAU), FB.Messenger (1.3 billion MAU), and WeChat (1.1 billion MAU). Popular microblogging platforms are QQ (808 million MAU), Sina-Weibo (486 million MAU) and Twitter (330 million MAU). This massive scale of active users highlights the popularity of online social media platforms and the increasing need for online social interactions, across the world. Moreover, Figure 1.1.d shows that all those online social media platforms have been launched in last 15 years and managed to get a massive number of active users within a very short period of time. For instance, TikTok a video sharing platform launched in 2017 has reached 500 million MAU by end of 2019. Furthermore, as shown in Figure 1.1.e, most of those active users engage with the social media platforms on daily basis. For instance in US, percentage of daily active users are close to 75% for Facebook, over 60% for Instagram, over 50% for YouTube and over 40% for Twitter. This phenomenon highlights that most of the active users use social media platforms frequently to participate in various online social activities.

A diversity of social behaviours

The social interactions on these online social media platforms are highly diverse representing the variety of interests of the users. In addition to those common online social

media platforms, social forums (or fora), which are also known as online support groups, self-help groups are a further innovation in this space. These are organised as discussion threads, where each thread starts with a question, comment or experience and other users of the forum respond to it. Such forums are mainly devoted for specialised areas such as health (WebMed, Patient.info), recreation (TripAdvisor, Yelp, Zomato), academia (Academia.edu), data science (Kaggle), and photography (EyeEm). Moreover, similar specialised social groups are often formed within mainstream online platforms such as Facebook groups, sub-reddits (in Reddit), LinkedIn groups to social interact with individuals with specific criteria (e.g, occupation, hobby, medical condition).

As online social media platforms continue to expand and develop, individuals and societies have become increasingly intrigued by the unprecedented opportunities provided to interact, share and collaborate globally. Even older individuals are increasingly using online social media platforms. For example, according to an annual survey in US by Pew Research (Pew Research Center, 2019) (Figure 1.1.f) use of online social media platforms are increasing rapidly among older age groups, and as of 2019 40% 65+ adults (was 2% in 2005 a 20 fold increase) and 69% among age 50-64 (was 5% in 2005, a 14 fold increase) use any form of social media platforms. In contrast, the increase was less for 18-29 and 30-49 age groups. It was a 5.6 fold increase (16% in 2005 to 90% in 2019) for 18-29 age group, and a 9 fold increase (9% in 2005 to 82% in 2019) for 30-49 age group.

Face-to-face interactions in the physical world that served as the main outlet of social behaviours during most of human history is gradually giving way to increased online interactions that occur on online social media platforms. This trend was apparent in a recent study (Wallsten, 2013) based on data from the American Time Use Survey (ATUS), which shows that the time spent socialising online has increased at the expense of the time spent on other activities such as face to face interactions, sleep, and travel. Initially, this trend of increased online social interactions was seen as a negative impact on the social life of individuals which lead to social isolation by decreasing physical social interactions (Kraut et al., 1998; Nie et al., 2002). However, later research points that although offline socialisation activities are degraded, online socialisation increased which leads to similar or even better overall social outcomes (Valenzuela et al., 2009; Ellison et al., 2007; Wellman et al., 2001). In fact, online social interactions overcomes some barriers of physical social interactions such as (i) *geographic proximity* where the individuals have to be in close proximity, (ii) *synchronicity* where interactions have to timely, and (iii) *accessibility* where some individuals and groups are not accessible to others (McFarland and Ployhart, 2015).

However, online social media are *lean mediums* in contrast to the *rich medium* of face-to-face communication which allows the transmission of multiple communication cues (Daft et al., 1987). In particular, non-verbal communication cues are not transferable through

most of the online social media platforms, excepts for those that allow video based interactions. This claim is disputed by other studies which argue that without non-verbal communication cues, humans are capable of adapting to utilise the available cues in *learner mediums* more effectively with prolonged use (Walther, 1992; Kock, 2004; Tidwell and Walther, 2002). For instance, in text-based mediums emoticons are used to expresses specific emotions and elongated terms (e.g., 'soooo') are used to emphasise the meaning of the base term.

Although a variety of social interactions (that represent a number of social behaviours) take place on online social media platforms, the baseline profile of salient social interactions is generally focused on the characteristics of each social media platform. Such characteristics of the social media platform are sometimes *by design* to intentionally encourage certain type of interactions, or may have emerged as a consequence of frequently having certain type of user interactions over others. For instance, mainstream social media platforms such as Twitter are fast-paced social media platforms, that are used for rapid dissemination of current and trending information through the social network using functions such as *re-tweeting* and *sharing* (Kwak et al., 2010; Lotan et al., 2011; Lovejoy et al., 2012), which may even spread faster than seismic waves during an earthquake (Sakaki et al., 2010a). Due to this focus of information dissemination rather than having continuing discussions, the ties between dyads are relatively weaker (unless known by other means), and often lacks homophily, reciprocity and interpersonal self-disclosure. Also, the emotions expressed in *tweets* are often shallow and intense, clearly leaning towards being positive or negative. In contrast, forums/online support groups such as online health support groups are slow-paced with relatively smaller groups where the participants has high degree of homophily (e.g., the patients of a specific medical condition). These social media platforms are used to seek and provide, informational and emotional support (Preece, 1999). Thus, users tend to publish longer posts providing high degree of self-disclosure and express complex emotions.

Digital representations of social data

Social data can be broadly defined as any form of information that represents and characterises social interactions. Every social interaction releases information into the environment in which it occurs, which can be both explicit or implicit as well as intentional or unintentional to the participants of the interaction. For instance, in a face-to-face conversation in the physical environment, some of the information that characterised the interaction are the theme of the conversation, duration, location, emotions/opinions expressed, facial expressions of the participants, tone variation of the conversation, and the appearance of participants (e.g., what they wear). In addition, there can be pre- and post-interaction information that characterises what leads to the interaction and what follows.

Social data from social interactions in the physical world (mainly face-to-face communications) is seldom recorded. The interactions in the physical environment are generated

by different forms of energy transformations which eventually dissipate into the environment without leaving any traces. A cognitive representation of social interaction exists in the memory of the participants, which gradually fades with time and often leads to inaccurate accounts (Schacter et al., 1998). Also, the recollection of such interactions are subjected to various forms of recall biases yielding a distorted recollection (Kahneman et al., 1999).

In contrast, almost all aspects of online social interactions are recorded. For instance, in a social conversation on Facebook or Twitter, its content (text, emojis, shared images or videos), duration, geolocation of participants and other actions (e.g., shares, likes, retweets) are recorded in the digitised form. In addition, pre- and post- online interactions of the individuals are recorded and each participant has a self-authored profile which often contains their socio-demographic information. All these information are stored in massive databases of online social media platforms, and often backed up as multiple copies. Theoretically, those digitised records of the online social interactions live indefinitely on the Internet. For instance, there are currently over 500 billion *tweets* archived in Twitter.

1.3 Motivation

Building upon this digest of social interactions, social behaviours, digital environments, and the social data accumulating on social media platforms, it is now pertinent to delineate the motivation of this thesis.

Conventional methods of studying human social behaviours are impacted by the challenges of collecting relevant data, in sufficient volumes. A distinctive opportunity lies within the massive amount of social data accumulated in online social media platforms. As noted earlier, social data are the archived records of online social interactions, thus, provides an unparalleled aperture into understanding various human social behaviours as well as underlying causalities. This data is contributed by a large number of individuals often in the scale of millions, thus they are in par or even larger than the population level datasets collected in censuses. Also, prolonged social interactions of individuals in online social media platforms generate sufficient information to conduct an in-depth analysis of their social behaviour.

Existing computational approaches of using these large volumes of social data are mainly adaptations of standard machine learning and natural language processing techniques. Table 1.1 shows some example uses of social data in different applications alongside the techniques employed and social media platforms used for data collection. These applications mainly use sentiment analysis techniques (Liu and Zhang, 2012) to evaluate public opinion, topic extraction techniques (Aggarwal and Zhai, 2012b) to identify salient discussion topics and entity extraction techniques (Ritter et al., 2011) to identify related entities (e.g., people, brands, locations) involved.

These adaptations of standard machine learning and natural language processing techniques demonstrate two high-level fundamental limitations when utilised for the exploration and understanding of social behaviours. Firstly, the constrained focus on basic insights of social data, when there are deep insights on human social behaviour that are encapsulated but largely untapped due to the challenging nature of pre-processing, analysis and synthesis. For instance, the shallow focus on surface-level meaning, frequency, and basic sentiment, and lack of consideration for deeper insights such as emotions, psycholinguistics, and personal traits, that provides a better representation of the underlying social behaviours in social data. Secondly, sub-optimal computational performance (Eisenstein, 2013) due to the highly unstructured nature of social data (Hu and Liu, 2012). The dynamic and bursty nature of social data (Barabasi, 2005), requires techniques that are time sensitive and can learn incrementally from a dynamic data stream, in contrast to conventional machine learning techniques which are mostly designed for static datasets.

Table 1.1: Applications of social data based on recent literature.

Application	Technique (social media platform)
Natural disaster detection	Sakaki et al. (2010a)- event detection (Twitter) Middleton et al. (2014)- event detection, clustering (Twitter) Abel et al. (2012)- event detection, entity extraction (Twitter)
Planned protest prediction	Ramakrishnan et al. (2014)- event detection, entity extraction (multiple sources) Becker et al. (2012)- event detection, entity extraction (multiple sources)
First story detection	Osborne et al. (2012)- event detection, entity extraction (Twitter, Wikipedia) Lau, Collier and Baldwin (2012a)- topic extraction (Twitter)
Traffic incident detection	Pan et al. (2013)- anomaly detection, graph analysis (Weibo) D'Andrea et al. (2015)- event detection, classification (Twitter)
Sales forecasting	Rui et al. (2013)- sentiment analysis, regression (Twitter) Mishne et al. (2006)- sentiment analysis (Blogs)
Stock market prediction	Si et al. (2013)- sentiment analysis, topic extraction (Twitter) Das and Chen (2007)- sentiment analysis, classification (stock message boards)
Public opinion prediction	Tumasjan et al. (2010)- sentiment analysis (Twitter) Bollen, Mao and Pepe (2011)- sentiment analysis (Twitter)
Public health insights	Huber et al. (2017)- thematic analysis (online support groups) Paul and Dredze (2011)- topic extraction (Twitter)

On motivation, this thesis is firstly inspired by the unprecedented opportunity to comprehensively study human social behaviours represented in social data accumulated on online digital social media platforms. Secondly, it is stimulated by the limitations of current computational and artificial intelligence approaches in addressing the scientific,

semantic and technical challenges of online social data. And finally, the ambition of common good and contribution to human society through the transformation of novel artificial intelligence approaches into technology platforms that deliver actionable insights for societal advancement.

1.4 Research objectives

This thesis develops an understanding of social behaviours through the study of online social data using novel computational and artificial intelligence approaches that model, transform, analyse and generate insights on social behaviours and underlying causalities. The research objectives are:

The first objective is to explore existing theories on human social behaviour, social needs and cognition; and develop a conceptual framework to understand social behaviours through representative online social data. This conceptual framework considers social data as representing the surface layer of a hierarchy of human social behaviour, needs and cognition.

The second objective is to advance new self-structuring artificial intelligence approaches to address the challenges and limitations of existing machine learning and natural language processing techniques in the study of social behaviours based online social data. This exploration will focus on self-structuring and incremental learning techniques to represent, adapt to and evolve with text based social data streams, and automatically monitor those structures for changes in social behaviours and causes.

The third objective is to transform the aforementioned approaches into technology platforms that capture insights from text based social data of different online social media streams and use such insights to gain descriptive understandings and predictive insights for decision-making.

1.5 Research questions

Based on the above research objectives, the following research questions will be addressed in this thesis.

1. *What are the limitations of existing artificial intelligence algorithms and natural language processing techniques in the study of social interactions and social behaviours using representative online social data in digital environments?*

2. *How can theories of social behaviour from social sciences contribute towards a conceptual model of enhanced understanding of social interactions in digital environments, as well as the representative online social data?*

3. *How can new incremental machine learning algorithms, founded on the principles of self-structuring artificial intelligence, address the challenges of using social data to understand social behaviours?*

4. *How can the research contributions that address research questions 2. and 3. be formulated into a technology platform that delivers actionable insights for societal advancement?*

1.6 Research contributions

Based on the above research objectives and research questions, this thesis yields the following contributions as the outcomes:

1. A comprehensive investigation on current state-of-the-art machine learning and natural language processing techniques employed to generate insights from social data followed by an exhaustive analysis on their limitations due to distinct challenges present in social data.

2. A conceptual framework based on existing social theories to explicate the complex relationships between social behaviours and online social data.

3. Development of two novel self-structuring artificial intelligence algorithms for generating insights from online social data. (1) A new unsupervised, self-structuring and incremental learning technique to structure a dynamic and diverse text based social data stream to automatically capture salient topics, and their dynamics over time. (2) An automated, intelligent event detector to monitor the identified topic pathways for significant fluctuations in social behaviours using multiple indicators such as volume and sentiment to detect social events of interest.

4. Demonstration of the two novel self-structuring artificial intelligence algorithms on two large microblogging data streams.

5. Transformation of the two new algorithms, along with an ensemble of related machine learning and natural language processing techniques, into a technology platform. This platform was further adapted for slow-paced online social data, to automatically capture and incrementally learn self-disclosed information.

6. Extending the above mentioned platform into an application domain with high social impact, patient-centred healthcare. The platform was demonstrated to better facilitate the diverse information needs of its stakeholder groups: consumers, researchers and health professionals.

7. Further materialising the patient-centred healthcare technology platform for prostate cancer related online support groups, with real-life outcomes that have advanced clinical knowledge of patient needs and expectations. This was demonstrated using a text corpus collected from 10 prostate cancer online support groups comprising of 609,960 prostate cancer discussions and 22,233 participants.

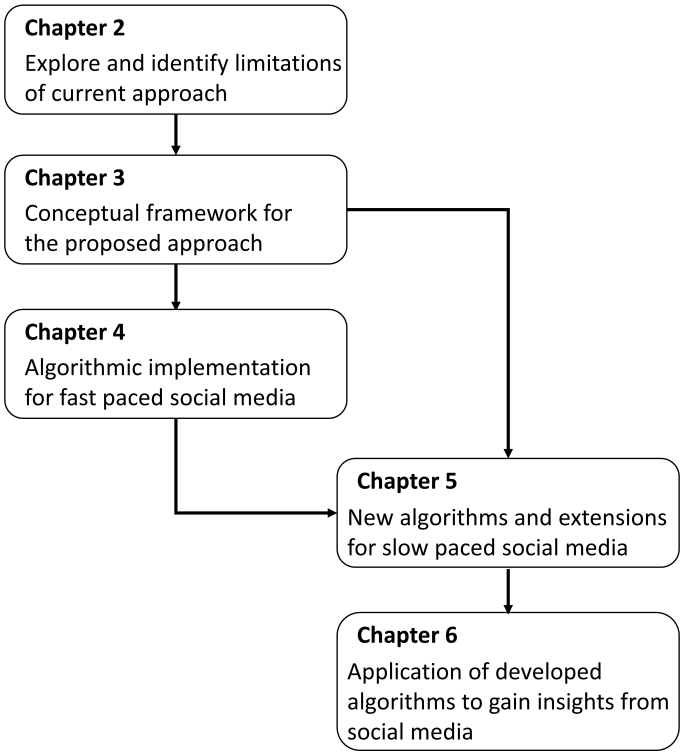


Figure 1.2: Chapter structure of the thesis

1.7 Thesis outline

Figure 1.2 shows how this thesis is organised.

Chapter two discusses the key literature on machine learning and natural language processing techniques that are used to capture insights from textual sources which includes text clustering, topic modelling, event detection, sentiment and emotion capturing, self-structuring and incremental learning. Subsequently, it presents the major challenges such techniques would encounter when applied to social data.

Chapter three presents the proposed conceptual framework for understanding the underlying mechanisms that generate social data using the existing physiological and social theories on cognition, social needs and social behaviour as the foundation. This conceptual framework is then extended to a platform that can be employed with artificial intelligence techniques to represent, self-structure and learn insights from social data.

Chapter four presents two novel techniques (i) a self structuring and incremental learning technique that is capable of automatically structuring a text based social data stream into topic pathways over time, and (ii) an automatic event detection technique that automatically monitors the above topic pathways for behavioural changes using multiple change detectors. These techniques were trialled using two large Twitter datasets and demonstrated their capabilities to identify salient topic pathways, track their evolution over time, detect new topic pathways, and detect significant events based on sentiment and volume based event indicators.

Chapter five presents the development of information structuring platform for online support group platforms to better facilitate the diverse information needs of its stakeholder

groups: consumers, researchers and health professionals. It presents the details of a set machine learning and natural language processing techniques to capture self disclosed information such as demographics, emotions, psycholinguistics. The capabilities of this platform are demonstrated using a large text corpus extracted from two large online health support groups.

Chapter six presents the application of the techniques developed in chapters four and five to gain insights from prostate cancer related online support group discussions. In addition it discusses the development of a specialised set of technique to capture treatment type, side effects, treatment decision making behaviour and numeric pathological information specific to prostate cancer from the self disclosed free-text. It then showcases the capabilities of the captured multidimensional representation (prostate cancer related information, topics/topic pathways, demographics, emotions) to analyse the self disclosed quality of life against the other factors.

Chapter seven concludes the thesis by providing a summary of work presented in the preceding chapters, how the research questions formulated above were addressed in the body of work and ideas for future research based on this thesis.

Chapter 2

Literature review and reflections

If I have seen further it is only by standing on the shoulders of giants.
Sir Isaac Newton (1675)

This chapter reviews existing machine learning and natural language processing techniques that are relevant to harness insights from social data. Also, it discusses the limitations and gaps of such techniques to effectively capture insights from social data.

This chapter is organised as follows: Section 2.1 reports on manual assessment of social data and its limitations; Section 2.2 reviews the techniques for topic extraction and text clustering; Section 2.3 reviews event detection from text data focussing on both specified and unspecified event detection techniques; Section 2.4 reviews emotions theories, models and sentiment/emotion extraction techniques from text; Section 2.5 discusses the self-structuring techniques covering biological inspiration and its computational adaptations; Section 2.6 reviews techniques developed for unsupervised incremental learning; and Section 2.7 discusses the challenges and limitations of existing techniques in relation to gaining insights from social data.

2.1 Manual assessment of social data

Social research using social data is a relatively emerging field of study that was developed in conjunction with the rapid growth of social media platforms and their use. During the initial stages, social data was analysed manually by the researchers. Such attempts include (i) manually coding social messages based on the topic discussed in the content, (ii) manually assessing of the sentiment polarity and strength, and (iii) manual inference of socio-demographic information of the authors of social data.

The key advantage of manual assessment of social data is that it does not require challenging technical skills. Also, humans are skilled at decoding the complex social expressions such as sarcasm which are often present in social data. However, personal judgement on social data varies across assessors as such judgements are impacted by the personal experience of each assessor. Such variations specifically occur in the assessment of sentiment or emotions expressed in social data. These variations were often minimised

by using multiple assessors to assess the same social data and producing an aggregated assessment. Also, statistical agreement measures such as kappa coefficient (κ) are measured to evaluates the inter assessors agreement.

However, the key issue with the manual assessment of social data is the associated human cost that limits its applicability to small social datasets often in the scale of hundreds of data points. Use of crowd sourced manual assessment techniques such as Amazon Mechanical Turk (Buhrmester et al., 2011) enables relatively inexpensive manual assessment possible. However, even such approaches can only assess social datasets in the scale of thousands. One of the key utility of the manual assessment is for the generation of reference assessment results which are often known as *gold standard records* that are used as the reference social data points for training and evaluation of the machine learning based models to automatically learn the underlying patterns of inferences made by human assessors.

2.2 Topic extraction

Topic extraction (or mining) is one of the key areas in natural language processing which attempts to organise a large set of documents (e.g., news articles or tweets) into semantically meaningful groups which are often denoted as *topics* or *themes*. Each identified topic is represented as a set of keywords that semantically represents a particular topic.

Topic extraction is mainly handled as an unsupervised task as in most of the scenarios underlying topics in a document corpus is not a priori knowledge. Two key unsupervised approaches for the topic extraction are (i) text clustering and (ii) topic modelling.

2.2.1 Text clustering

Clustering is the classic data mining task of finding groups of similar records in the data based on a similarity measure. Similarly, text clustering intends to separate the text corpus into groups with similar meaning (Aggarwal and Zhai, 2012a). It is assumed that a set of documents clustered together contains a common topic/theme/subject which is represented by the salient terms of that cluster. Text clustering techniques have been applied to organise documents (Cutting et al., 1992; Kohonen et al., 2000), organise web search results (Zamir et al., 1997), and generate an abstractive summary of a given corpus (Schütze and Silverstein, 1997).

Text clustering consists of two key phases (Larsen and Aone, 1999). First the documents are converted into representative feature vectors and then clustering techniques are employed to cluster those feature vectors into groups based on an appropriate distance function. These two phases are required because the clustering techniques are designed for datasets with real-valued features. Hence, the text corpus has to be transformed into representative real-valued feature vectors.

Document feature vectors are often obtained based on the bag-of-words (BOW) concept (Jurafsky, 2000) which simplifies a document as a bag of words, relaxing word order based relationships. Based on BOW, documents are transformed into feature vectors using

vector space model (VSM) (Salton et al., 1975) where each term in the corpus vocabulary is a feature and the feature value is the term weight. Feature values of a document could be simply set to binary indicating presence or absence of a respective term in the document. However, term weighting techniques based on statistical specificity of a term are employed to apply higher feature values to more specific terms which have higher information value. A popular such weighting scheme is *term frequency inverse document frequency* tf-idf (Salton and Buckley, 1988a; Sparck Jones, 1972) which gives high feature value to terms that exist more on a given document (proportional to term frequency) and less across the corpus (inverse document frequency). The size of the vocabulary often grows with the corpus size resulting in very high-dimensional feature vectors. Therefore, document frequency based pruning techniques are used to remove less informative terms which are either common terms among most of the documents or rare terms that exist in a small set of documents (Salton et al., 1975).

The techniques employed in text clustering are either extended or adapted from the clustering techniques used for quantitative data such as k-means (Lloyd, 1982), self-organising-maps (SOM) (Kohonen, 1990), and hierarchical clustering (Ward Jr, 1963). Scatter/Gather (Cutting et al., 1992) one of the well-known text clustering algorithm which uses hierarchical clustering to determine the cluster centres and further refines them using k-means clustering technique. (Larsen and Aone, 1999) proposed a variant of k-means which uses a damping technique during the centre adjustment that improves the cluster quality. (Steinbach et al., 2000) proposed a *bisecting* k-means variant that iteratively clusters the largest cluster (initially the entire dataset) into two clusters using k-means until k clusters are generated. (Kohonen et al., 2000) employs self-organising-maps to cluster a large collection of patent abstracts which uses randomly projected document vectors to reduce dimensions.

2.2.2 Topic modelling

Topic modelling is another approach widely employed to extract salient topics from a document corpus. It employs a probabilistic generative model to represent documents in a corpus. Topic models estimate probabilities for each document being generated by the identified topics, where the most salient topics of a document can be chosen based on a threshold. It means a document could belong to multiple topic clusters. This is in contrast to the previously discussed clustering approaches which exclusively partition a set of documents into topic clusters resulting in a single topic for a document. Having multiple topics is more intuitive for larger documents which may contain multiple topics.

Topic models assume that the corpus is generated by N number of *latent* topics, and each document d_i has a probability $P(t_j|d_i)$ for belonging to each topic t_j. Similarly, each term w_k in the corpus vocabulary V being part of a topic t_j is given by the probability $P(w_k|t_j)$. Both topic-document $P(t_j|d_i)$ and term-topic $P(w_k|t_j)$ cannot be directly estimated as topics are *latent* and the only relationship that can be estimated is the term-document probability $P(w_k|d_i)$ for a term w_k and document d_i which is given by the term frequency of w_k in d_i divided by number of terms in d_i. Therefore, the topic-document

and term-topic probabilities are estimated based on its relationship with term-document probability given bellow:

$$P(w_k|d_i) = \sum_{j=1}^{N} P(w_k|t_j) \times P(t_j|d_i)$$

Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999, 2001) is one of the initial generative topic models, which used expectation maximization (EM) algorithm to estimate the probabilities of the topic-document and term-topic relationships. Initially, the probabilities are randomly initialised and then iteratively conduct *expectation* (E) and maximization (M) steps. E-step estimates the posterior probabilities of latent variables given the current values of them and term-document probabilities. Then M-step estimate new values for the probabilities based on the probabilities of latent variables estimated in E-step. These two steps alternatively carried out until the likelihood converges to a local maximum. The key parameter to PLSA is the number of topics N that the generative model is optimised. The key limitation of PLSA is its tendency to overfit when estimating large number of parameters.

The most popular variant of probabilistic topic models is the Latent Dirichlet Allocation (LDA) (Blei et al., 2003) algorithm. LDA closely follows PLSA but, is a complete generative model which models topic-document and term-topic probabilities using Dirichlet distribution as priors using hyperparameters α and β respectively. α controls the mixture of topics in a document and higher values lead to higher topic-document probabilities i.e. more topics are assigned to a document. β controls the term-topic probabilities and high values lead to higher term-topic probabilities i.e., more terms are assigned to a topic. In contrast to PLSA, LDA has less number of parameters, thus less susceptible to overfitting problem. Moreover, LDA is capable of estimating the topic distribution of a new document that was not present during learning.

2.2.3 Topic extraction from social data

The two key challenge for topic extraction techniques on social data is the high sparsity and time sensitivity. Text data is often sparse, however, social data is overly sparse mainly due to brevity and diversity. Social media posts are often left brief as it allows authors to construct them efficiently and with less concern about the flow of it. On the other hand, social data is constructed by many individuals who use diverse writing styles and discuss diverse topics resulting a large vocabulary across a corpus of social data. Both these factors contribute to sparse document feature vectors (from *vector space model*) containing only a handful of activated features. This increased sparsity makes topic extraction techniques susceptible to noise, degrading their performance.

The topics discussed in social data changes over time as the interests of the author changes. Existing topics drop in popularity while new unseen topics emerge over time. Therefore, the topic mixture changes dynamically over time. However, both text clustering and topic modelling techniques assume a static mixture of topics thus have to be extended to facilitate changes in topic mixture.

Sparsity issue is often addressed using various aggregation strategies to increase the document size used for topic extraction. Rosen-Zvi et al. (2004, 2010) proposed the Author-Topic model. It is a generative model similar to LDA, except that, instead of topic-document relationship modelled in LDA, topic-author relationship is modelled in Author-Topic model. Hence, it estimates the topic-author association probabilities yielding a set of topics discussed by each author. This approach addresses the issue of sparsity as it learns a topic mixture for an aggregated corpus of social data of each author, thus has been utilised for topic mining in social data (Tang et al., 2008; Hong and Davison, 2010). However, Xu, Ru, Xiang and Yang (2011) pointed out that this strategy would only work for authors who are consistently discussing about a set of topics, and also it does not account for changes of authors interest over time.

Yan et al. (2013) present *biterm topic model* which builds a generative model to represent topic to word-pair (biterm) relationship. The assumption here is that if two words co-exists frequently it is likely that those two words are part of the same topic. This model is claimed to more robust to brevity in social data as it does not model the topic-document relationship which tends to suffer more from sparsity issues.

Mehrotra et al. (2013) attempted different aggregation strategies such as aggregate tweets based on hashtags, authors and time-window (aggregate hourly tweets). They empirically showed that hashtag based aggregating is the most effective for LDA based topic extraction. However, the granularity of hashtag based aggregation varies widely based on the usage of each hashtag. For example, #obama is used to discuss a wide spectrum of topics, so aggregating tweets of that hashtag might not yield meaningful topics. Another approach to the sparsity problem is to use relevant external resources to extend social data. Banerjee et al. (2007) and Hu et al. (2009) incorporate text in related Wikipedia articles with the short text to generate joint document feature vectors to reduce sparsity before applying text clustering techniques. Similarly, Petrović et al. (2012) employed Wordnet (Miller, 1995) ontology to enrich short text with paraphrases. These approaches are significantly time and resource extensive as they have to query a big knowledge graph for each text document, making them expensive to use in the scales of social data.

Topic modelling techniques are also extended to incorporate time sensitivity of text data. Wang and McCallum (2006) present *topic over time* (ToT) which is an extension to the LDA topic modelling algorithms to consider time in the generative model. The time-topic relationship added to the generative model of LDA with document-topic and term-topic relationships. The time-topic relationship is modelled using a beta distribution as the prior in contrast to document-topic and term-topic relationships modelled using Dirichlet priors. However, in ToT time is employed only to get a better fit to a fixed topic mixture and the topic trends such as topic evolution and new topic are not captured. Moreover, it is not trained in incremental fashion to use in a text stream.

AlSumait et al. (2008) and Lau, Collier and Baldwin (2012b) present *online* versions of LDA (OLDA) which are incremental versions of LDA, to facilitate topic mining from a text stream. These online variants take time-sliced input of a text stream and incrementally

update a LDA topic model at each time-slice. They use the LDA model generated from text corpus at $t-1$ time-slice as the prior to learn the LDA model at t. This process generates an evolutionary matrix over each topic to capture the evolution of that topic over time. Based on this evolutionary matrix, the topics that are significantly different at t in contrast to $t-1$ are marked as an outlier or new topics.

2.3 Event detection

An *event* can be broadly defined as an occurrence that can be bounded by space and time (Allan, Papka and Lavrenko, 1998). Events could happen anytime and anywhere which could impact a single person, a handful of individuals, or a large number of individuals (e.g., natural disaster, election). Events alter the behaviour of the individuals impacted.

An event in a social media platform is due to either a real world event (e.g., natural disaster, election) discussed in the social media platform or it can be originated from the social media platform itself without a link to a real world event (e.g., a controversial tweet from an influential author). An event often results a change of online social behaviour of the impacted individuals for the duration of the event (e.g., change of discussion topics, changes of emotions). Such changes due to an event often happen in near-realtime as a result of frequent use and availability of mobile devices. For instance, during natural disasters such as earthquakes and floods, related discussions instantly start to flow in social media platforms with firsthand individuals accounts, opinions, emotions and arguments (Java et al., 2007; Zhao and Rosson, 2009).

Identifying these events and also analysing the related online social behaviour yield valuable information for individuals, corporations, and law enforcement agencies. Corporations increasingly monitor social media platforms to understand customer opinions, concerns and sentiment on their product portfolio; and to provide insights for business decisions (Jansen et al., 2009; Pak and Paroubek, 2010). Social media platforms can be further monitored to identify breaking news on events reported in realtime by firsthand individual accounts (Phuvipadawat and Murata, 2010; Sankaranarayanan et al., 2009). Also, social media platforms are monitored for events that may lead to public unrest (Ramakrishnan et al., 2014).

Event detection techniques can be broadly separated into *unspecified* and *specified* events (Atefeh and Khreich, 2015) on the event type. Unspecified event detection is employed when prior information about the event is not available, and specified event detection is employed when details about the event looking for is partially known.

2.3.1 Unspecified event detection

Unspecified events are events with no prior information available. Such events can range from extreme events such as natural disasters or terrorist act, unplanned political events such as a controversial statement from a key politician, and unplanned corporate events such as product malfunction or service outage. Since no prior information is available

unspecified events can only be detected by monitoring the relevant social data streams for significant temporal changes. Therefore, unspecified events are mostly detected using unsupervised techniques.

Detecting and tracking unspecified events in a stream of news stories has been comprehensively studied during the DARPA funded Topic Detection and Tracking (TDT) research program (Allan, Carbonell, Doddington, Yamron, Yang et al., 1998). Both Allan, Papka and Lavrenko (1998) and Yang et al. (1998) employed single pass text clustering techniques to cluster a time-ordered news article stream where each cluster is denoted as an event. Each news article is checked for its membership in an existing cluster and if it significantly different such articles forms a new cluster i.e., new event.

However, event detection in social data streams is more challenging than the detection tasks in traditional media. As discussed in Section 2.2.3, in contrast to the professionally composed news articles, social data is unstructured, short and written with diverse styles. In addition, often social data streams are several magnitudes higher in scales compared to news steams.

Events with a significant public interest results increased number of related discussions by the individuals which lead to appear as sudden bursts (Kleinberg, 2003; Yang et al., 1998) in social data streams. Those discussion starts as bursts and then diminished over time with a heavy tail as public loses the interest on that. The discussions in social data streams are driven by events, thus can be considered as a mixture of event related discussions happening at different time windows. Detecting such bursts of activity is often the key to detecting events from a social data stream.

Initial burst detection techniques look for bursty keywords or tag words (hashtags in Twitter) (Atefeh and Khreich, 2015) with the hypothesis that such burst keywords are indicative of an event related to the topic inferred by the meaning of those keywords. Mathioudakis and Koudas (2010) developed a queuing technique to detect *bursty* terms from tweets and subsequently clustered the identified such terms into topics/events. Fung et al. (2005) modelled the individual word distributions over time as binomial distributions and identified the bursty terms. Those terms with sufficiently overlapping active time windows are grouped together and identified as events. He et al. (2007) applied Discrete Fourier Transform (DFT) and searched for bursty terms in the frequency domain which appears as a spike in the frequency spectrum. However, the time window of the event cannot be retrieved from the frequency domain. Weng et al. (2011) developed *Event Detection with Clustering of Wavelet-based Signals* (EDCoW) which applied the Wavelet Transform to word distributions, filters out the trivial terms based on autocorrelation, and then employed a graph clustering technique to group the event related terms into events. Sankaranarayanan et al. (2009) trained a Naive Bayes classifier to classify tweets as news or junk, which was trained on term vectors of a hand-labelled tweet dataset as junk or news. The identified *news* tweets are then online clustered to identify the salient topics among the news tweets.

Although change of volume based signals are widely employed for event detection, there are other changes that could happen to a social data stream due to an event. An event

can significantly alter the opinion or emotions expressed in a discussion (Tan et al., 2014; Thelwall et al., 2011). For example, a social media post by a politician on a certain issue could lead to a backslash with heavy criticism resulting significant increase in negative opinion and expression of emotions such as *anger* or *sarcasm*. Detecting such opinion or emotions related signals, supplement the volume based events detections, as well as provides more detailed insights about the genre of the event.

Paltoglou (2016) developed a sentiment based event detection technique that monitored the temporal variation of sentiment polarity of tweets with particular keywords (e.g., hashtags) and identified instances with significant changes of sentiment as events that are associated with that keyword. The events captured from this technique was shown to be comparable to the events captured from a frequency based approach. One key limitation of this technique is the assumption that there is a one-to-one mapping between keywords and topics, whereas in social data streams there can be multiple keywords associated with topics and vice-versa.

2.3.2 Specified event detection

Specified event detection is the detection of known events and its related information such as location, time, participants, outcomes etc. (Atefeh and Khreich, 2015). In contrast to unspecified events, the key terms that denote or associated with the event(s) are *a priori* information and can be employed to detect them. Hence, in contrast to the unsupervised change detection techniques used for unspecified events, specified events are mainly detected using supervised techniques which uses either an engineered feature dictionary or labelled datasets to train a classifier. Also, unlike the change detection techniques which requires a substantial change in social data stream from large scale events, specified event detection techniques can be employed to detect events at different scales, from large events such as riots, earthquakes to individual events such as report of an illness from a personal social data stream.

Specified event detection on different types of text sources has been studied in different application domains. For instance, in financial domain, financial documents and business news are mined for finance related events such as catastrophic financial events (Cecchini et al., 2010), mergers and acquisitions (Lau, Liao, Wong and Chiu, 2012), and competitor relations (Ma et al., 2011). Similarly, social data streams are mined for mentions of relevant entities (e.g., organisation names, brands, movie titles) and their associated mood (or sentiment) which can be used to determine market volatility such as stock price movements (Mao et al., 2012; Bollen, Mao and Zeng, 2011) and movie revenues (Rui et al., 2013; Asur and Huberman, 2010). In the health domain, electronic health records (EHR) such as admission notes, treatment plans, patient summaries, radiology reports, pathology reports and discharge notes are mined using natural language processing techniques to capture different clinical events of patients such as diagnosis, treatments, symptoms and side effects (Murdoch and Detsky, 2013; Meystre et al., 2008). These events are used with machine learning techniques to gain further insights. For example, association rules such as symptom-disease, drug-disease (Chen et al., 2008), drug-side effects (adverse drug

events) (Wang et al., 2009) and disease co-occurrence (Jensen et al., 2012) are automatically extracted from EHR text. These associations are further employed to build machine learning models to predict outcomes such as cancer survival (Miotto et al., 2016), and hospital readmissions (Rajkomar et al., 2018). These insights can be employed enhance the capabilities of clinical decision support systems which are used to assist clinicians by providing relevant and reliable information as well as automatically monitor, alert and even predict clinical conditions of patients (Jensen et al., 2012; Demner-Fushman et al., 2009).

In law enforcement domain social data streams are often monitored to capture events that could impact social stability (e.g., protests, riots) since social media are often used to organise and propagate information about such events (Lotan et al., 2011). (Popescu and Pennacchiotti, 2010) developed a supervised model to detect controversial events based on the expressed public opinion in Twitter. Similarly, Sakaki et al. (2010b) developed a model to detect natural disaster related event such as earthquakes. These techniques capture the events retrospectively while it is more beneficial to capture any planned events before it occurs. To capture future planned events Ramakrishnan et al. (2014) developed EMBERS platform that mines online free-text data (social data streams, websites, news and Wikipedia) to capture events that could lead to civil-unrest in 10 Latin American countries.

Irrespective of the problem domain, the key steps of specified event extraction can be abstracted to (i) identify the relevant documents/web pages/posts that contain event related information, and (ii) identify relevant entities and their relations to the event.

Identifying documents that contain event information

The techniques that mainly employed to extract event related documents are (i) thesaurus based techniques and (ii) classification model based techniques.

Thesaurus based techniques assumes that presence of certain terms in documents indicate that such documents contain information about the events of interest. Therefore, such thesaurus is engineered with event related terms and use it to identify the documents which contain such terms. For instance, EMBERS planned protest model (Ramakrishnan et al., 2014) employs a multi-lingual lexicon of planned protest related terms (e.g., *plan to strike* in English or *preparaci´on huelga* in Spanish) to capture the documents (e.g., social media posts or web pages) that discuss planned protests. These thesauruses are developed based on the domain knowledge and often based on knowledge bases or glossaries that exists in each domain. For example, in the health domain, the specified event detection techniques often employ the health related knowledge bases such as UMLS Metathesaurus (Bodenreider, 2004) which contains multiple health thesauruses including SNOMED CT which contains normalised health concepts, synonyms and descriptions; and RxNORM (Nelson et al., 2011) which is a drug thesaurus all drugs approved in USA.

The classification model based techniques use a labelled document set where positives discuss the event of interest and negatives do not, to train a machine learning classification model. The documents are first feature transformed, in order to convert

text into real-values features that are needed for machine learning techniques. Generally employed feature transformation techniques include one-hot-encoding which transform presence/absence of a term into a binary feature and *term frequency inverse document frequency* tf-idf (Salton and Buckley, 1988a; Sparck Jones, 1972) techniques discussed in Section 2.2.1. However, recent advances in semantic text-representations lead to the development of word-embedding techniques (Mikolov, Sutskever, Chen, Corrado and Dean, 2013) that are often used as a feature transformation in deep neural network based techniques. The features transformed positive and negative samples are then used to train a classification model which learns to differentiate positives from the negatives. This trained model is then used to identify the event related documents from an unlabelled document collection or stream. For instance, Sakaki et al. (2010b) trained an SVM (Support Vector Machine) (Cortes and Vapnik, 1995) based classifier using feature transformed tweets where positives are tweets that refer to an actual earthquake. This technique is mainly employed when reliable domain knowledge bases do not exist for the events of interest.

Identifying relevant entities of an event

Once an event is identified the next step is to capture more information about the event. This task is generally achieved by capturing the relevant named entities that are related to the identified event.

Named entities refer to terms or phrases that consistently stands for the same referent. This definition is derived based on *rigid designator* (Kripke, 1972): a designator is rigid when it denotes the same thing in all possible worlds. The relevant named entities defer across application domains. For instance, in a public event, the named entities are often location, time and people/organisations involved, whereas in the biomedical/health domain the named entities could be health related entities such as drugs, illnesses, symptoms, treatments, genes etc.

Recognition of named entities is achieved through detection of entity terms (words or phrases) and subsequently classifying them into meaningful categories. The recognition part is often handled using text chunking, regular expressions and noun phrase extraction techniques (Nadeau and Sekine, 2007) to extract potential entity phrases.

The categorisation of the identified entities is either achieved based on relevant ontologies or using supervised learning techniques. The ontologies contain term dictionaries for high-level categories and the identified entities present in those dictionaries are assigned to the relevant categories. Those ontologies are often engineered for the requirements of different application domains. WordNet (Miller, 1995) and YAGO (Mahdisoltani et al., 2013) are popular examples of such ontologies available for general purposes which contain a hierarchy of high-level categories for each entity in the ontology. Similarly, there are specialised domain specific ontologies i.e., UMLS Metathesaurus (Bodenreider, 2004) for health domain. These ontologies are employed by the state-of-the-art health text-processing tools such as Apache cTAKES (Savova et al., 2010) and MetaMap (Aronson et al., 2010).

The supervised learning techniques attempt to learn the semantic patterns around the entities using a labelled corpus and incorporate those patterns to identify the entities from an unlabelled corpus. The early attempts of supervised learning techniques are mainly based on a set of hand-crafted features to captures named entities of different categories. For example, rules like "Mr. XXX" can be used to identify person names (Nadeau and Sekine, 2007) in text. More recent attempts are based on machine learning techniques such as Conditional Random Fields (CRF) (Lafferty et al., 2001; Finkel et al., 2005) to learn generalised semantic patterns of the entities from the training corpus. One of the popular training corpus for supervised NER is MUC-6 (Grishman and Sundheim, 1996) which categorised the named entities into *person*, *organisation*, *location*, *time/date* and *money*. These categories are formulated for information extraction tasks form news articles. This categorisation has been followed by most of the current state-of-the-art NER tools in Standford NLP (Finkel et al., 2005) and Open NLP (Baldridge, 2005).

2.4 Emotion extraction from text

Emotions are an important aspect of human social life and social behaviour. In psychology, emotions are defined as a complex state of mind that influence thought process and behaviour. Emotions play an integral part in human social interactions, especially in the interpersonal communication process. In face-to-face communication, humans express emotions verbally (emotional words and tone) as well as using non-verbal cues such as facial expressions and hand gestures. In written communication expression of emotions is limited to the use of emotional terms (words, phrases and emoticons).

Understanding the emotions conveyed is an integral part of understanding the communicated message, during the communication process. The emotions conveyed in a message supplements its discussion topic by adding opinions/feelings of the communicator towards that discussion topic. Such expression of emotion is ubiquitous in social data during social conversations where the individuals freely express their opinions and feelings on different discussion topics. Capturing these emotions encapsulated in social data provides invaluable insights about the conversations happens in social media platforms, which is important to understand the opinions/feelings of individuals towards different topics.

In contrast to topic extraction and event detection, techniques on emotion extraction from text are still in the early stages. This is mainly because emotions extraction from text only became relevant and useful recently with the rise of social media platforms, while topic extraction and event detection techniques were started to develop much earlier to capture topics/events from news articles.

The rest of this section is organised as follows: firstly, emotion theories and relevant emotion models were discussed which are being used as the basis for the emotion extraction techniques. Subsequently, sentiment extraction techniques (two dimensional emotions extraction) were discussed and finally, multi-granular emotion extraction techniques were discussed.

2.4.1 Theories of emotion

The key theories of emotion can be categorised into physiological and cognitive. Physiological theories suggest that emotions are due to physical changes in the body. A prominent physiological theory is James-Lange theory of emotions (Cannon, 1927) which was formulated by two 19th-century scholars, William James and Carl Lange. This theory states that emotions are nothing more than conscious feelings about bodily changes. This idea that emotions occur solely due to physiological reactions to events is been criticised by others. Another similar theory is Cannon-Bard theory (Cannon, 1927) which states that after an event physiological reactions and emotions occurs at the same time. Two-factor theory (Schachter and Singer, 1962) a more recent theory of similar genre which states that emotions are generated by a two factor process. First, the physiological event occurs and then the individuals refer to his past experience or the immediate environment to find emotional cues to provide an emotion label to that event. A key commonality in those theories is that the stimulus for emotions is physical changes.

Cognitive theories, on the other hand, argue that emotions occur due to conscious cognitive activities such as thoughts, judgements, and evaluations. One of the key theories in this line of thought is Lazarus cognitive theory (Lazarus and Lazarus, 1991) proposed by Richard Lazarus. This theory states that emotions are determined by the cognitive appraisal of an event or stimuli in the environment. It argues that the quality and intensity of the emotions are controlled by the cognitive process. Therefore, cognitive appraisal mediates the effect of the event in the environment. Since the cognitive appraisal process is personalised (based on their past experience), same event often yield different emotional responses from different individuals. Figure 2.1 presents how emotions are generated following an event in the environment based on different theories of emotion discussed above.

The term *emotion* is used alongside and interchangeably with terms *sentiment*, *affect*, and *feeling*. Shouse (2005) points out that *affect* is the most general and abstract of the above terms. Affect represents the intensity of experience as either positive or negative. It is often not conscious and not explainable using language (Munezero et al., 2014). In contrast, feelings are conscious and labelled sensations that are checked against the past experience. Hence, feelings are personal to each individual as the labelling of sensations is based on the past experience of the individual. Moreover, Shouse (2005) states that emotions are the display or projection of feelings. Friedenberg and Silverman (2011) state that emotions are short episodes that involves both brain and body to respond to an event of interest. The sentiment is defined as the positive or negative mental predisposition in an individual towards an object (e.g., another person, event). Sentiment is often formed when an individual constantly perceive or think about an object, which leads to building up a dispositional idea about the object.

2.4.2 Modelling emotions

Characterising or assessing emotions is a challenging task. This is because, firstly, as discussed in above theories emotions originated by a cognitive process, which makes emotions

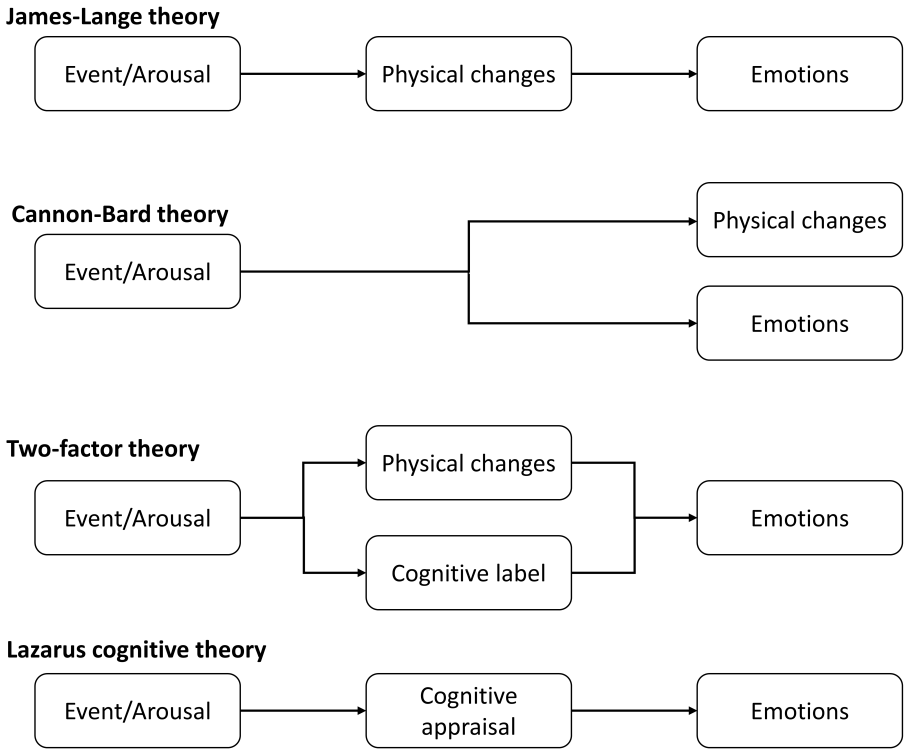


Figure 2.1: The process that generates emotions, according to several prominent emotion theories.

harder to measure as an entity. Secondly, emotions are personalised to individuals based on their personal experiences, thus the same event/experience may often yield different emotions on different individuals. These challenges lead to the development of emotional models to conceptualise human emotions. In contrast to emotional theories which describes origin of emotions, emotion models provide a conceptual model for characterising emotions. Such conceptual models are often employed in machine learning applications as the basis for capturing insights from emotion expressions.

The development of emotion models has taken two distinct approaches:

1. Dimensional models, which model emotions as a continuous n dimensional space.

2. Discrete models, which model all emotions into a number of basic categories.

Dimensional models argue that emotions are overlapping and interrelated affective states generated by a common neurophysiological system (Posner et al., 2005). Therefore, all human emotions can be represented in a conceptual continuous n dimensional space (often two or three dimensional). Each of these dimensions is modelled to represents a different aspect of the underlying neurophysiological system. The key dimensional models are (i) positive activation negative activation (PANA) model (Watson and Tellegen, 1985), (ii) Vector model (Bradley et al., 1992) and (iii) Circumplex model (Russell, 1980).

PANA model (Watson and Tellegen, 1985) represents emotions in two dimensions where the horizontal dimension is the strength of positive affect and the vertical dimension is the strength of negative affect. Among those models, the most popular dimensional model for emotions is the Circumplex model (Russell, 1980). It argues that all emotions

originate from two neurophysiological systems, *valence* (pleasant to unpleasant continuum) and *arousal* (active to passive continuum). According to Circumplex model, each emotion can be represented as a linear combination of valance and arousal. For instance, *excited* is a pleasant and active emotion while *relaxed* is a pleasant and passive emotion. Figure 2.2 shows a graphical representation of the Circumplex model and how different emotions can be represented along the valance and arousal dimensions of the model.

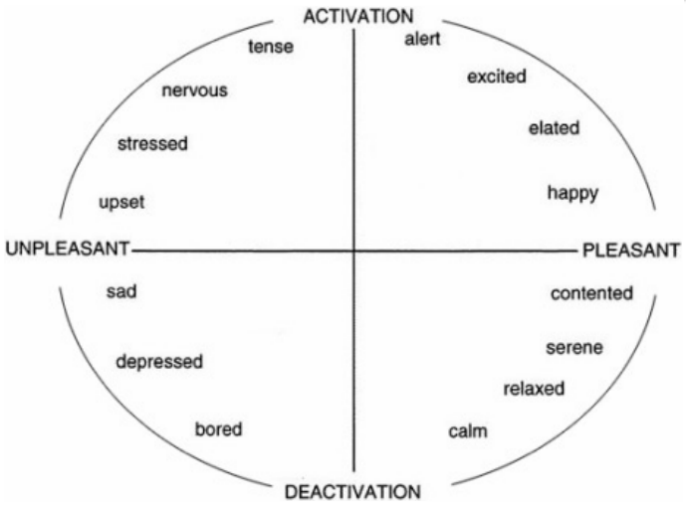


Figure 2.2: A graphical representation of the Circumplex model consisting of valence and arousal dimensions. Multiple emotions are placed in the axis system based on their valance and arousal. Image adapted from Posner et al. (2005).

In contrast to dimensional models, discrete emotion models argue that human emotional expression consists of a set of distinct basic emotions. These basics emotions are common to all humans and even recognisable across different cultures. The hypothesis of distinct basic emotions stems from Darwin (1872) that different emotions are distinct neuropsychological phenomena shaped by the natural selection process to provide an organised physiological and cognitive responses to challenges and opportunities in the environment (Plutchik, 1980a). In other words, emotions are formulated as part of the evolution process and thus, it is argued that there should be a set of basic emotions common to all humans.

This hypothesis of a universal set of basic emotions was initially experimented by Darwin (Darwin, 1872) in which he showed an array of photographs with different facial expressions to people and asked them to recognise them. This experiment had been advanced by (Ekman and Oster, 1979) where they showed photographs of Europeans with different facial expressions to two isolated tribal communities (have never met Europeans) in Borneo and New Guinea. Those people (including children) from the tribal communities were able to understand the emotions expressed in the facial expressions of the photographs with significant accuracy. Also, photographs with facial expression taken from the tribes were recognised by US college students with a similar level of success. These experiments lead to the assertion that there exists a basic set of discrete emotions present in all humans. Ekman (1992) further argues that basic emotions are distinguishable as they have distinct universal signals such as facial expressions common across all humans and also often

common across other primates as well. Ekman (1992) identified that there are six such distinct basic emotions which are: anger, disgust, fear, happiness, sadness, and surprise.

Similar findings on discrete basic emotions were published in other contemporary studies (Izard, 1993; Lazarus and Lazarus, 1991; Tomkins, 1984; Plutchik, 1980a). Out of these studies, emotion wheel proposed by (Plutchik, 1980a) is of particular interest. Plutchik presented a psychoevolutionary theory of basic emotions which argues that basic emotions evolved with the evolution thus, applies to all animals. The basic emotions provide support to all animals against the challenges from the environment. The theory further postulates that emotions can be conceptualised into pairs of polar opposites, emotions have a varying degree of similarity to others, and emotions have different levels of intensity in arousal. Based on these postulates, Plutchik (1980a) presented a set of eight basic emotions anger, fear, sadness, disgust, surprise, anticipation, trust, and joy. These emotions exist as four bipolar pairs such that anger vs fear, joy vs sadness, trust vs disgust and surprise vs anticipation. Also, as shown in Figure 2.3 those emotions were placed in a circumplex model based on the degree of similarity among emotions creating a wheel of emotions known as *Plutchiks emotion wheel*.

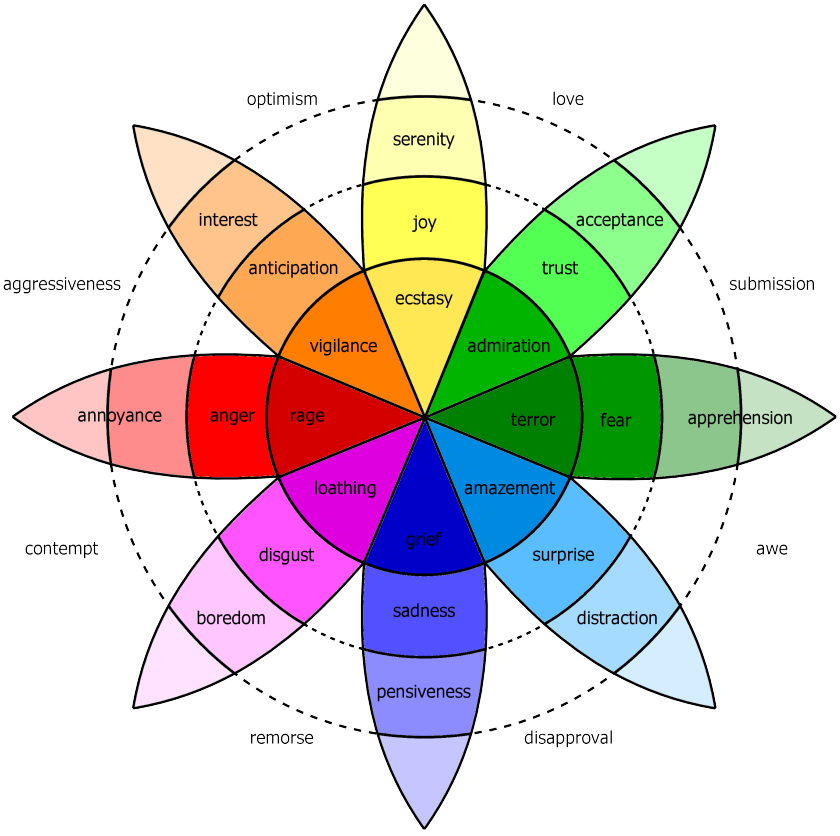


Figure 2.3: Plutchiks wheel of emotions adopted from (Plutchik, 2001)

The next sections discuss how those emotions models are utilised to detect sentiment and emotions from text.

2.4.3 Sentiment detection from text

As discussed in Section 2.4.1 sentiment is often defined as affect (or emotion) towards an object or topic. Sentiment detection (or opinion mining) is the computerised analysis of discourse to capture emotions/opinions expressed towards an object (or topic) by the author. Most of the sentiment extraction techniques provide a positive or negative intensity score for a given text based on the emotional expression in that text.

Sentiment detection techniques are often considered as the computational implementation of the Circumplex model (Russell, 1980) discussed in the previous section, which models emotions in a two dimensional space as linear combinations of valance and arousal.

Most of the early sentiment detection techniques were limited to detecting valance (positive/negative) dimension of the text (Pang et al., 2008). Those techniques consider the task as a binary classification task and classify a given text as positive or negative based on the sentiment expressed in that text. The key issue with these techniques is the lack of arousal or strength of the sentiment expressed. For instance, both "food is good" and "food is exceptional" would be just rated as positive overlooking that the latter is relatively more positive.

Recent sentiment detection techniques attempt to capture both valence (positive/negative) and arousal (strength of the sentiment) and provide a polarised sentiment score (Mohammad, 2016) where the polarity represents valance and the number represents the amount of arousal. These approaches consider the problem as either a regression problem or a multi-class classification problem (sentiment in a Likert-type scale).

Sentiment detection techniques can be broadly categorised into two approaches (Medhat et al., 2014): (i) rule based classifiers based on sentiment lexicons and (ii) machine learning models learned using a text corpus with human annotated sentiment.

Rule based classifiers depend on a sentiment lexicon, which is a dictionary of lexical features (e.g., words, phrases) that has a semantic orientation as either positive or negative. Some lexicons only provide a list of positive and negative terms. Two prominent such lexicons are General Inquirer (Stone et al., 1966) and LIWC (Pennebaker et al., 2001). General Inquirer (Stone et al., 1966) is the oldest sentiment lexicon that is still in use, which has been manually constructed by social scientists, political scientists, and psychologists to capture different aspects of text messages. It is a thesaurus of more than 11,000 terms in 183 categories, in which the terms categorised as positive and negative can be used as a sentiment lexicon. LIWC (Pennebaker et al., 2001) which stands for *Linguistic Inquiry and Word Count* is a thesaurus constructed by sociologists and linguists. It contains more than 4,500 terms organized into 76 categories, in which categories *positive emotion* and *negative emotion* can be used as a sentiment lexicon.

There are also sentiment lexicons with both polarity and valence scores for each term. SentiWordNet (Esuli and Sebastiani, 2006) is sentiment lexicon which is created by annotating over 115,000 synsets in WordNet (Miller, 1995) lexical database. Each synset is annotated with three numeric scores indicating its *positivity*, *negativity*, and *objectivity* where the sum of the three scores equals to 1.00 for each synset. Affective Norms for English Words (ANEW) (Bradley and Lang, 1999) is another lexicon with valance scores.

It has 1,034 English terms which were given numeric scores ranging from 1-9. These scores indicate both valence and arousal where < 5 is negative, $= 5$ is neutral and > 5 is positive. Valance score is given by $\|score - 5.0\|$, where a score of 1.0 provides the highest negative valance and score of 9.0 gives the highest positive valence.

A simple rule based classifier would search a given text for the presence of sentiment related terms based on the employed sentiment lexicon and assign the polarity based on the most prominent sentiment polarity present in the text. Similarly, sentiment valence can be determined by aggregating the valence scores of the terms present in the given text. Advances versions of the rules based sentiment classifiers are equipped with rules to capture linguistic patterns when used with sentiment related terms, would alter valence and/or arousal of that term given in the lexicon. The state-of-the-art rule based techniques such as SentiStrength (Thelwall et al., 2012) and VADER (Hutto and Gilbert, 2014) uses a set of engineered rules with a thesaurus of terms relevant to each rule to alter the designated sentiment score of sentiment related terms.

For instance, VADER (Hutto and Gilbert, 2014) uses a sentiment lexicon constructed by combining lexicons General Inquirer (Stone et al., 1966), LIWC (Pennebaker et al., 2001) and ANEW (Bradley and Lang, 1999) as well as social media related emotion artefacts such as emoticons and slang (e.g., LOL). Valence scores for this lexicon were determined by crowdsourcing. Moreover, a set of four rules were used to alter the sentiment score of a term as follows:

1. Degree modifiers: those terms when used in conjunction with a sentiment related term alter (increase or decrease) the valance. For instance, the term *really* increases the valence of *good* when used together (e.g., good vs really good). Similarly, *somewhat* decreases the valence (e.g., good vs somewhat good). A dictionary of degree modifiers was employed or this rule.

2. Negation: negations terms like *not*, *hardly* reverse the polarity (arousal) of a sentiment term (e.g., not good). A negation term dictionary is used to capture such occurrences.

3. Capitalisation: increase the valence of a term (e.g., good vs GOOD)

4. Exclamation mark: increases the valence (e.g., good vs good!)

Similarly, sentiment lexicon in SentiStrength (Thelwall et al., 2010, 2012) is constructed based on General Inquirer (Stone et al., 1966), LIWC (Pennebaker et al., 2001); as well as a set of slang, emoticons and idioms commonly used in social data. The initial valence scores for each term is determined by a pool of raters. Moreover, since SentiStrength is mainly optimised for brevity in social data, the valence scores were further refined for that domain using a supervised machine learning technique which was based on a corpus of sentiment labelled MySpace posts. Similar to the previous technique, SentiStrength uses rules such as negation, degree modifiers, punctuations, capitalisation and repeated letters (good vs goooood) to alter the sentiment scores.

The rule based classifier approach involves layers of engineering efforts which includes the construction of sentiment lexicons, assigning valence scores to each sentiment term

and setting up linguistic rules to alter the sentiment score of sentiment terms. This process result a set of sentiment scores for a given text based on its content, which is often aggregated to produce the final sentiment score of that text.

In contrast, the techniques that train a machine learning model mainly dependent on a sufficiently large text corpus with sentiment labels/scores. Using sentiment label/scores as the target, machine learning models were trained on the features extracted from the text corpus. The features vectors often constructed based on words and n-grams in the document using vector-space model. In addition, the syntactic structure of the document is often captured by using part of the speech tags and punctuations as features. Moreover, the classifiers designed for social data often capture features that represent hashtags, user mentions, and emoticons related features.

In an early work, Pang et al. (2002) employed several machine learning techniques to train a sentiment classifier on a labelled movie review dataset using unigram, bigram and POS features. Labels were automatically generated as positive, neutral and negative based on the star rating provides with a movie review. It was found that Support Vector Machine (SVM) technique performs better than other techniques. Other contemporary works employed similar machine learning techniques such as Naive Bayes (Melville et al., 2009) and Random Forest (Da Silva et al., 2014). Some approaches used feature selection techniques such as PCA and information gain (Riloff et al., 2006) to reduce the sparsity by selecting important features.

Socher et al. (2013) build the Stanford Sentiment Treebank, a corpus of movie reviews (a subset of (Pang et al., 2002)) with fully labelled parse trees. Such parse trees can be used to capture the context of the sentiment language in contrast to assessing the sentiment of individual words or phrases. Socher et al. (2013) further present a recurrent neural network (RNN) based sentiment classifier learned from the above corpus which is learned to represent sentiment arousal and valence. It was shown that RNN is capable of capturing the context specifics of sentiment.

Lexicon based approaches are more generalisable to different application domains as they are based on sentiment lexicons constructed using emotional terms that are generally employed in any domain. However, such approaches may not sensitive to the sentiment expressions that are specific to a particular domain. For instance, some sentiment lexicons do not include sentiment terms that are widely used in social data (e.g., LOL), but not part of the standard English vocabulary. In contrast, machine learning models learn specific sentiment expressions encapsulated in their training corpus. Such techniques also often capture more complex sentiment expressions that may span across multiple words and semantic patterns that are used to alter the sentiment expression. Hence, machine learning models which learned from a sufficiently large corpus often yield better performance on a similar corpus in comparison to lexicon based methods. However, they are less generalisable to capturing sentiment from text corpora in a different domain, as the patterns learned by the machine learning model may not be generally applicable.

There are attempts on hybrid approaches which employs both sentiment lexicons and machine learning models to build better sentiment detection techniques. Lexicon based

techniques employ machine learning models to optimise for a given application domain. For instance, SentiStrength Thelwall et al. (2012) optimise the valence scorers of its sentiment lexicon to social data using a machine learning model trained on MySpace text corpus. Similarly, SocialSent Hamilton et al. (2016) employed to build a domain specific sentiment lexicon from a word-embedding learned from a corpus of a given application domain. This technique first builds a word embedding from a large corpus of text from the relevant application domain. Subsequently, it starts with a generic seed lexicon of positive and negative terms and uses a random walk technique to expand the lexicon by looking for similar terms in the word embedding. This approach result a sentiment lexicon that better represent the sentiment expressed in the given application domain. Machine learning model based techniques also employed sentiment lexicons to improve its generalisability. For instance, Chikersal et al. (2015) incorporated some features based on sentiment lexicons in their machine learning model. These lexicon based features consist of the number of negative and number of positive terms of each sentiment lexicon present in the given text. Such inclusions make the model more generalisable as it considers sentiment polarity assessment from different lexicons.

2.4.4 Emotion detection from text

Emotion detection from text has been a relatively new avenue of research in contrast to sentiment detection, mainly due to higher complexity and lack of resources such as emotion labelled corpora and emotion lexicons. In fact, sentiment detection is a simplified version of emotion detection which assess emotion only in one or two dimensions, while emotion detection generally denotes assessing emotions with increased granularity (more than two dimensions). Because of this close resemblance, from the technical aspect, emotion detection techniques follow the same approaches taken by the sentiment detection techniques.

Current state-of-the-art emotion detection techniques are extensions to sentiment detection techniques. For instance, in rules based emotion detection techniques were developed extending rules based sentiment detection techniques by simply replacing sentiment lexicon with an emotions lexicon. Similarly, machine learning models designed for sentiment detection were trained on an emotion labelled text corpora to detect emotion.

Emotion lexicons were constructed in different granularities which often adherer to different emotional models discussed in Section 2.4.2. Most of these emotion lexicons are based on Ekman (1992) emotion model of six basic emotions or Plutchik (1980a) emotion model of eight basic emotions. NRC Emotion Lexicon (Mohammad and Turney, 2013) is one of prominent emotion lexicon which has emotion terms related to eight Plutchik emotions joy, sadness, fear, anger, anticipation, trust, surprise, and disgust. It has close to 14,000 emotion terms generated using crowd sourcing techniques.

Similarly, emotion labelled dataset were created mainly based on the emotion models Ekman (1992) or Plutchik (1980a). Alm et al. (2005) annotated sentences from children's stories with eight emotions (Ekman (1992) six basic emotions and positively and

negatively surprised). Similarly, Strapparava and Mihalcea (2007) annotated news headlines based on Ekman (1992) six basic emotions. Brooks et al. (2013) annotated 27,344 chat messages based on 13 emotions (eight basic emotions and five secondary emotions from Plutchik (1980a)). Unlike sentiment annotation, these emotion annotations are multi-label where each text can be labelled as containing multiple emotions. These annotated corpora are used to learn machine learning models (by the original authors and others) to detect emotions as multi-class classification problems.

2.5 Self structuring techniques

Self structuring or self organising techniques automatically learn a structure from a given dataset and then maps each data-point to the most similar structural element. This structure can be used to develop a coherent grouping of the dataset. Such techniques are essential to learn the underlying structure of social data.

2.5.1 Biological inspiration

Self structuring techniques were mainly inspired by the self-structuring capabilities of cells (neurones) in human cortex. Those cells apart from passing information to upper layers, self-structure horizontally as well. This self-structuring happens through short-range excitatory interactions between the neighbouring cells and inhibitory interactions between distant cells (Ratliff, 1965). These inhibitory and excitatory actions lead to competition and correlative learning which is the basis for self-structuring. Topographically ordered maps are observed in many parts of the cortex (Kertesz, 1983) including visual cortex (Van Essen, 1985), auditory cortex (Reale and Imig, 1980) and somatotopic cortex (Kaas et al., 1979). In these topographically ordered maps, different sensory input receptions are mapped to different parts of the cortex. Although the primary structure of the cortex is determined before birth, these subsequent topographical orderings are due to self-structuring based on the sensory receptions.

This self-structuring learning mechanism in the cortex is studied and theorised by many research works. One of the most pioneering and prominent work is Hebbian theory (Hebb, 2005) which states that "When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic changes take place in one or both cells such that A's efficiency as one of the cells firing B, is increased". This theory explains that synaptic strength among neurone cells strengthen overtime if activation of one cell repeatedly and persistently leads to the activation of another cell. This can be mathematically stated as the synaptic weight between input cell and output cell is proportional to the correlation between input and associated output. Similar prominent theories are proposed elsewhere. Marr's theory of the cerebellar cortex (Marr and Thach, 1991) states that cerebellum is an associative memory which maps the state of the body learned from cutaneous and proprioceptive receptors into motor commands. Malsburg's theory on self-organisation in visual cortex (Von der Malsburg, 1973) states that retinotopic organization (a mapping from retina to visual cortex) is learned through

self-organisation process where neighbouring cells in visual cortex mapped to similar input cells in retina.

2.5.2 Self organising maps

The biological process of self-structuring in the brain has inspired the design of several unsupervised computational techniques such as Willshaw-Malsburg Neural Network Model (Willshaw and Von Der Malsburg, 1976) and Kohonen's Self Organising Map (SOM) (Kohonen, 1990, 1997). However, simplicity, computational efficiency and scalability of the Self Organising Map is unparalleled to any other such techniques which have lead SOM to be the only successful computational technique inspired by the biological self-organising process.

Kohonen's Self Organising Map (SOM) (Kohonen, 1990, 1997) follows the Hebbian learning rule which emulates the self-structuring process in brain, but is simplified to reduced the computational complexity by replacing pre- and post-synaptic layers with the computational efficient two (or one) dimensional map structure. As illustrated in Figure 2.4, this map self-learns a structure from a high dimensional input space to a low dimensional map space while preserving the topological relations exist in the high dimensional input space.

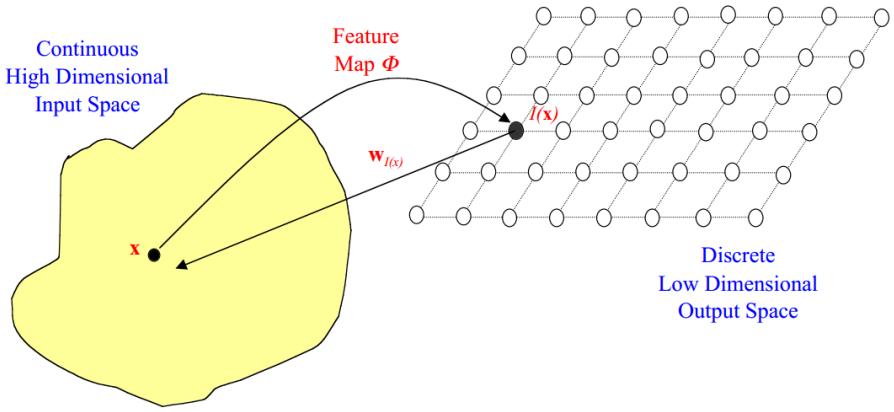


Figure 2.4: Self Organising Map(SOM) creates a low dimensional structure from a high dimensional input space while preserving the topological relations exist in high dimensional input space. Source: (Haykin, 1994)

The learned SOM structure consists of nodes where each is denoted by a vector which has the same dimensionality as the input space. As shown in Figure 2.4, once learned from the input data each node $I(x)$ in SOM structure maps to a space x in the input space and represent that space in the SOM map. Since SOM is learned while preserving the topological relations in the input space neighbouring nodes in the learned structure maps to adjacent spaces in input space.

SOM structure is initialised as a fixed 2-dimensional grid of nodes where each node is randomly initialised with a d dimensional vector (where d is the dimensionality of the input space). Once initialised SOM uses a combination of competition, cooperation, and

adaptation phases to self learn its topology preserving low-dimensional structure from the input space.

The competitive phase is used to find the *winning node* or the *best matching unit* for each input. This *winning node* N_w for each input is selected using a similarity function which assesses all nodes against a given input and selects the closest node to the given input:

$$sim(X, N_w) \geq sim(X, N), \forall N \in \{N\}$$

where $\{N\}$ includes all nodes in the SOM structure. Euclidean distance is often used as this similarity function and select the node with the lowest Euclidean distance to the input. This competitive process closely resembles the process in the brain where input neuron excites an output neurone.

Next is the cooperative phase in SOM which emulates the lateral interactions formulated among the neighbourhood of the neurones where a fired neurone excites its adjacent neurones more than the distance ones. In SOM, the topographic neighbourhood of a node is determined using a neighbourhood function $\mathcal{N}(d_{i,j}, t)$ that decays with the lateral distance $d_{i,j}$ between winning node i and neighbouring node j. Also, to reach the convergence, the neighbourhood function often decays with the number of iterations t, so that winning nodes excite a wider neighbourhood during the initial stages, and a smaller neighbourhood in later stages. A Gaussian function is often used as the neighbourhood function $\mathcal{N}$ as it peaks at the centre and decays exponentially with distance from centre.

Finally, the adaptive phase updates the weights of winning node and its neighbourhood based on the input vector. The weight vectors of winning node and its neighbourhood gets updated to reduce the distance to the given input. Let $w_1, \ldots, w_d$ be the weight vector (in a d dimensional space) of the SOM node j and $x_1, \ldots, x_d$ be the input vector, weight updates in each dimension k in iteration t is determined as follows:

$$w_{j,k}(t) = w_{j,k}(t-1) + \alpha(t) \times \mathcal{N}(d_{i,j}, t) \times (x_k - w_{j,k}(t-1))$$

where $\alpha(t)$ is the learning rate which determines the amount of learning captured into the winning node at each iteration. $\mathcal{N}(d_{i,j}, t)$ is the neighbourhood function where winning node is i and $\mathcal{N}(d_{i,i}, t) = 1$. In SOM, learning rate is a decaying function over the number of iterations (t), leading to substantial adjustments during the initial iterations and minor adjustments in later iterations. This weight adapting process emulates the Hebbian theory (Hebb, 2005) by adapting the weights of the winning neurone to reduce the distance between input and winning neurone so that the efficiency of exciting that neurone to same (or similar) input increases.

Observations from the input space are iteratively used in the SOM learning process. During initial iterations, with higher learning rate and bigger neighbourhood, rapid self-structuring of the SOM grid takes place. Over the iterations learning rate drops and neighbourhood size diminishes leading to a smooth convergence of the SOM grid to the learned topological structure. The number of iterations required for convergence depends on the complexity of the underlying structure of the input space. Convergence in SOM can

be assessed by monitoring the quantization error, which is the sum of absolute distances between input and the winning node $\sum^{d} |x_i - w_i|$. Quantization error rapidly drops during initial iterations and converges to a minimum value as SOM converges. The learning can be stopped when quantization does not show significant improvements over iterations.

Self organising map has been used in many application domains in both academia and industry over the last few decades (Oja et al., 2003; Kohonen, 2013). It has been frequently used as a tool for exploratory analysis of large datasets as it projects high dimensional input data into two dimensional grid which can be easily visualised to understand the underlying patterns exist in a dataset. In addition, it has been used in applications such as dimensionality reduction, clustering, classification, anomaly detection, and information organisation/retrieval.

2.5.3 Dynamically expanding self organising maps

A key limitation in self organising map is its fixed grid size which needs to be predetermined. SOM grid size determines the number of nodes trained which is indicative of the learning capacity of SOM. Using a smaller grid on a large complex dataset would result in over-generalisation of the learned structure where some nodes may represent more than one distinct patterns. On the other hand using an unnecessarily large grid is expensive in terms of run time and memory usage.

The key solution to overcome this issue is to use an expandable grid which dynamically expands during the training phase if more nodes are required to represent the underlying structure of the input dataset. There are several extended versions of self organising map developed to start with a small number of nodes and dynamically expand during training based on error metrics that indicates more nodes are required.

Growing Grid (Fritzke, 1995a) is a dynamically expanding version that starts with 4 nodes. In every iteration, it updates error of the winning node based on the distance to the input. Expansion of the grid happens after every $k \times m \times \lambda$ iterations (k and m are the dimensions of the grid), where it expands the grid based on the node q with the highest accumulated error. A new column or row is added to the grid between q and its furthermost connected node. This addition of an entire column or row may lead to adding too many nodes if trained on a large dataset (a large number of iterations).

Growing Neural Gas (GNG) (Fritzke, 1995b) creates a dynamically growing graph (instead of a grid). It also keeps track of the accumulated error and after λ iterations select the node with the highest error for further growth. Similar to Growing Grid, a new node is added in between node with the highest error and its furthest neighbour.

Growing Self Organising Map (GSOM) (Alahakoon et al., 2000) is another dynamically growing version of SOM. Similar to the above methods it keeps track of the error accumulated in winning nodes, but unlike above methods it triggers growth when the largest accumulated error exceeds a threshold. Growth happens from the node with the largest error and nodes added to all its remaining edges in the grid. Unlike Growing Grid, GSOM does not maintain a rectangle shape of the structure and nodes are only added to the impacted node.

The self structuring techniques were mainly designed for conventional datasets where it expects a real-valued dense feature matrix. Extended versions are required to handle social data which are often unstructured, sparse and high-dimensional.

2.6 Unsupervised incremental learning techniques

Incremental learning is the paradigm of learning where existing learned knowledge incrementally gets extended and updated as new data comes in. Incremental learning techniques are the machine learning techniques that are capable of learning knowledge incrementally from new data. Such techniques are mostly applied to scenarios where the entire dataset is not known priorly but available over time. A naive approach for such scenarios is to discard previously learned knowledge/model and re-learn from scratch using both old and new data. However, such an approach is highly resource consuming in terms of both processing (to process the entire dataset) and memory (model has to keep old data for retraining). Incremental learning techniques overcome these challenges by only learning from new data and updating the existing model to reflect knowledge in new data.

Incremental learning is a must for social data, as it is a constantly evolving data stream. There will always be new unseen patterns appear on social data streams. Also, previously appeared patterns may appear later, so learning techniques have to keep the previously learned knowledge intact without discarding.

Polikar et al. (2001) have specified four key characteristics of an incremental learning technique as follows:

1. It should learn additional information from new data.
2. It should not require access to the past data that it has already processed.
3. It should not suffer from catastrophic forgetting, thus should preserve the previously acquired knowledge
4. It should be able to accommodate new classes that may be introduced with new data

As mentioned in the first characteristic incremental learning techniques should capture any additional previously unseen information from the new data. Second characteristic delineates that learning from new data should not require access to previously processed data. This characteristic enables such techniques to be scalable to handle big data streams where retaining old data is not feasible. Although some techniques may keep samples of previous data that represent each class or group seen in the data. The third characteristic is related to a phenomenon called catastrophic forgetting (French, 1999) where models discard previously learned knowledge completely when new knowledge is presented. This ability to learn new knowledge while preserving the existing knowledge is known as the *stability-plasticity dilemma* (Carpenter and Grossberg, 1988). It is a dilemma because the incremental learning techniques have to sufficiently stable to be resilient to the noise in new data while flexible enough to learn new knowledge encapsulated in new data (Gama

et al., 2014). If it learns noise from new data, that could lead to dropping relevant previous knowledge causing catastrophic forgetting. On the other hand, if the technique is too conservative it won't learn important new knowledge from new data. The fourth characteristic is to accommodate previously unseen classes that may come with new data. This is primarily for supervised learning techniques. However, in the context of unsupervised learning, it refers to accommodating previously unseen patterns in data.

There are numerous supervised techniques that support *supervised* incremental learning, however, unsupervised incremental learning techniques were relatively less. Unsupervised incremental learning techniques maintain a dynamically evolving cluster structure reflecting the dynamics of the data stream, where new clusters may appear and others may disappear. Such incremental clustering techniques were mainly adaptations from the standard clustering techniques.

Leader (Späth, 1980) is one of the earliest partitioned based incremental clustering technique. It employs a user defined threshold to partition data into clusters. Every new input is assigned to the closest cluster if its distance to the cluster centroid is less than the given threshold, otherwise a new cluster centroid (leader) is added based on that input. This approach is simple but unstable as it depends on a user defined threshold for separating clusters. Another technique is the single pass k-means (Farnstrom et al., 2000) which is an incremental learning adaptation of k-means. It keeps a fixed sized buffer and when the buffer is full from incoming data, runs k-means to identify k cluster. After running k-means, only cluster centroids are retained in the buffer, and when it is full again runs a weighted k-means using new data and existing centroids. This technique incrementally update clusters as new data arrives, however, fixed k value limits capturing of new emerging clusters.

Aggarwal et al. (2003) introduced CluStream, a stream clustering framework which separates incremental clustering into two phases online micro-clustering and off-line macro-clustering. Micro clusters are groups of data points which keeps statistical information of data and time stamps. A new data point absorbed into an existing micro-cluster if its distance to the centroid falls within that micro-cluster boundary. Only a fixed number of micro-clusters are maintained, and when it exceeds, older clusters were merged into others. In off-line macro-clustering, k-means runs over the set of micro-cluster centroids formulating the final clusters.

Furao and Hasegawa (2006) introduced self-organizing incremental neural network (SOINN) which is an incremental clustering technique based on self-organising-maps (Kohonen et al., 2000). SOINN uses a two layer network where the first layer represents density distribution of inputs and second layer separates clusters by detecting low dense areas in the first layer. For each input presented, first and second winner is identified from the first layer, and if the distance to those exceeds a threshold, then the new input is added as a new node (between class insertion). Otherwise, standard SOM weight updates occur updating first winner and its neighbours (within class insertion).

Another recent technique is Incremental Knowledge Acquisition and Self-Learning (IKASL) (De Silva and Alahakoon, 2010; De Silva, 2010) which self learns a layered structure across time generalising the knowledge embodied in data. Each layer learns from a buffered batch of data using GSOM self-structuring technique (Alahakoon et al., 2000). IKASL preserves the acquired knowledge in a generalised form therefore, it does not require access to the past data that it has already processed. The generalised version of the acquired knowledge from each layer (n) is used as the basis for the knowledge acquisition from the subsequent layer (n+1), thus it avoids catastrophic forgetting of the past knowledge. Moreover, while using the past acquired knowledge as the base, it incrementally acquires new knowledge that is embodied in the upcoming data.

Similar to self structuring techniques discussed in the previous section, unsupervised incremental learning techniques were mainly designed for conventional datasets where it expects a real-valued dense feature matrix. Extended versions are required to handle social data which are often unstructured, sparse and high-dimensional. Also, existing techniques do expect all features to be previously known, which is not the case in social data as new features will appear and existing features may disappear over time.

2.7 The formidable challenges to social data analysis

Previous sections discuss the existing machine learning and natural language processing techniques that are used to gain insights from social data generated in online social media platforms. As discussed within the above sections gaining insights from social data is highly challenging. Such challenges come from several avenues: (i) scale of data generation, (ii) time sensitivity, (iv) diversity and (ii) the unstructured nature of data.

2.7.1 Scale of data

As discussed in the previous chapter, over the past decade, online social media platforms have been increasingly embraced by hundreds of millions of people worldwide. Moreover, users are more frequently active in major social media platforms mainly due to the availability of mobile devices.

These large number of active users in prominent social media platforms generate large volumes of new social data continuously resulting in a high velocity data stream. The analysis of data streams of this scale requires specialised computational approaches that are optimised to handle large volumes of data.

2.7.2 Unlabelled data

As discussed in previous sections, supervised machine learning techniques were mainly employed in classification tasks such as sentiment, topic and event classification. Such tasks were carried out using two types of techniques: (i) dictionary based approach where a term dictionary is engineered to represent the concepts of each class to be classified and (ii) model based approach, where a labelled training dataset is employed to learn a predictive model using supervised machine learning techniques. The first approach requires

a term dictionary or a lexicon and the second approach requires a labelled dataset with relevant classes.

Since social data is not structured with any taxonomy or explicitly labelled, it does not carry any labels that are needed for such supervised or semi-supervised learning tasks. There are a handful of human labelled social data samples, however, such samples are disproportionate and not representative enough when considering the scale and variation in social data.

There are two key approaches to handle this issue. The first is to use unsupervised learning techniques which do not require any form of supervision. Unsupervised techniques leverage underlying patterns in data to separate them into groups based on similarity. Current use of unsupervised techniques in text (mainly topic extraction) is covered in Section 2.2.1.

Another approach to overcome this issue is to use folksonomies or user generated tags as a proxy for labels. For example, (i) 20 Newsgroups (Joachims, 1996) which is used for topic classification is taken from a news discussion forum where the news category is used as the label, (ii) Movie Review Dataset (Pang and Lee, 2005; Pang et al., 2008) used for sentiment classification is a set of user created movie reviews on the movie review website Rotten Tomatoes [1] in which the labels were obtained based on the user provided star rating for the movie.

2.7.3 Time sensitivity

Social data can be considered as a near real-time data stream which gets updated by freshly generated content from its users. For instance, Facebook users frequently publish status updates about how they feel and Twitter users frequently tweet their opinion about their current issues of interest.

Moreover, as pointed out by Hu and Liu (2012) these social data streams are not uniformly distributed, but bursty in nature. These bursts are mostly due to current events of interest that have disrupted (Sakaki et al., 2010b; Zhou and Chen, 2014) or captured the attention (Becker et al., 2011) of a significant swath of human population. Hence, it is apparent that social data is tightly coupled with the time of its publication and that time sensitivity needs to be considered for any analytic task.

Most of the above discussed conventional machine learning algorithms require the dataset to be a priori which is not possible in social data streams as new patterns appear in the data streams as it progresses over time. Therefore, machine learning algorithms need to be extended to incrementally learn from the data stream as new patterns appear over time.

2.7.4 Diversity

Diversity in social data is an important feature that is often overlooked in other studies. Social data is diverse due to multitude of reasons. It contains traces of multiple social

[1] www.rottentomatoes.com

behaviours. Also, even the same behaviour could be actioned in different depths and intensities. Moreover, at the highest granularity, contrasting differences among different individuals may appear due to individuality resulting from differences in socio-demographics.

All these reasons induce diversity to social data. For instance, social data consists of diverse linguistic patterns, a large number of discussion topics and differences in expression of emotion. Eisenstein et al. (2014) found that there are distinct linguistic patterns exist in Twitter among groups of similar geographic proximity as well as similar socio-demographics. This diversity in social data act as noise on patterns that exist in social data. Hence, the machine learning techniques have to first reduce the noise due to diversity by separating diverse social data into coherent groups which can then be used to extract insights.

2.7.5 Unstructured nature of data

Social data mainly consists of unstructured text posted by users in online social media platforms. This text corpora of social data are substantially different from the professional discourse (found in news articles and other formal documents) in many aspects such as brevity, lack of syntactic structure and use of out-of-vocabulary terms.

Brevity

Short length of social data partially due to restrictions imposed by some social media platforms (e.g., tweet is limited to 140 characters). Even in platforms without such restrictions, users prefer brevity since it allows them to express more efficiently. This brevity is mostly achieved by using shortened forms of terms (e.g., *u* for *you*), and relaxation of grammaticality (Baldwin et al., 2013). This brief nature of social data result in highly sparse feature representations that is challenging for conventional natural language processing techniques.

Lack of syntactic structure

Discourse in documents often follows certain syntactic structure which is driven by a set of language rules that governs the formation of clauses, phrases and sentences. Such rules were often established through repeated documentation over time. However, the discourse in social data loosely follows such syntactic structures. This is mainly because although a written language social data is the outcome of social conversations where individuals prefer to make it more relaxed from syntactic structures.

The above discussed differences in social data to any other formal discourse are highly challenging most of the conventional natural language processing techniques that rely on such language rules to capture different constructs from the discourse such as noun, noun phrases, co-references etc. Since those elementary constructs are then feed into other techniques to capture complex insights such as sentiment, these errors due to lack of syntactic structures often impacts most of the conventional natural language processing techniques.

Out of vocabulary terms

Users tend to coin new terms during social conversations in online social media platforms. Such terms are mostly developed as social tagging or folksonomy highlighting important aspects i.e., topics or emotions in the conversation text. Hashtags used in Twitter is a good example of such terms, where users create *hashtags* to represent certain events or topics (e.g., *#WinterOlympics*). Another type of new terms is constructed by repeating certain characters of standard terms (e.g., *cooolll*) which are mostly used to emphasise the expressed emotion (Brody and Diakopoulos, 2011). Use of such out-of-vocabulary terms is a challenge as such terms are often not included in thesauruses that are often used by the natural language processing techniques to derive certain properties of each word (e.g., sentiment dictionaries). Filtering out such terms not an option as they are tightly coupled to the intended meaning of the post.

From the above examples, it is apparent that as is application of natural language techniques which are designed for standard discourse on social data would yield sub-optimal outcomes (Baldwin et al., 2013). Such techniques have to be significantly extended to capture different above discussed aspects of social data.

2.8 Chapter Summary

This chapter has conducted a comprehensive literature review on current state-of-the-art machine learning and natural language processing techniques employed to generate insights from social data. It focused around three types of insights often captured from social data which are topics, events and emotions. Topic capturing includes topic modelling to model topic distributions and text clustering to capture topically coherent clusters. Event detection is twofold where specified event detection is used to capture events with known prior information and unspecified event detection is used to capture events with unknown or partially known information. The emotion detection approaches use emotion models based on physiological and cognitive theories of emotion.

As highlighted in the review most of the techniques employed are adaptations and extensions of machine learning and natural language processing algorithms developed for conventional text datasets such as news articles. However, as discussed in Section 2.7 social data comes with formidable challenges that are less prevalent in conventional text datasets due to its scale of data generation, time sensitivity, unstructured nature and diversity. Hence this approach leads to suboptimal outcomes.

The next chapter proposes an alternative approach to this conventional approach of social media analytics. This proposed novel approach goes beyond considering social data as just another data source by considering social data as traces of online human social interactions. It enables social data to be represented by the underlying drives of human social interactions which can then be transformed to generate more meaningful insights.

Chapter 3

The Intellection of a Conceptual Framework

Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less.

Marie Curie

The previous chapter presented current work in machine learning and natural language processing techniques that are being used to process, analyse and learn insights from social data. As shown, the majority of such approaches for social media analytics were adaptations and extensions of machine learning and natural language processing methods developed for conventional text datasets. Although such approaches generate useful results, they tend to overlook the human behaviours, emotions and thought process which are embedded in social data. Therefore a wealth of rich information which can be used to infer potential causality of events and behaviours lie unused. This thesis proposes a novel approach that goes beyond conventional social media analytics. It considers the data generated in social media platforms a.k.a social data as the traces of human social interactions online and thus can be better represent based on theories on social behaviours, which can then be used by machine learning and natural language processing techniques to generate more meaningful insights. This chapter presents the theoretical foundations for this novel approach which has been materialised in subsequent chapters.

This chapter begins with a discussion on the importance of social data in advancing an understanding of social behaviours (Section 3.1). In Section 3.2, behavioural theories from the social sciences are used to propose a new multi-layered conceptual framework that addresses this need. Section 3.3 concludes the chapter delineating the materialisation of this conceptual framework based on the paradigms of self-structuring artificial intelligence, which also lays the foundation for the rest of the thesis.

3.1 On the importance of social data

Social data is generated by humans through their social actions on online social media platforms in cyberspace. In fact social data can be considered as archived traces of human social actions/interactions in cyberspace. For instance, *tweets* are the outcome of social conversations and the expression of opinion that happens in the Twitter online social media platform.

Social data and online social media platforms are relatively new paradigms which have only been in existence for the past two decades. However, online social actions that generate social data are in fact adoptions from human social actions in the physical world. For example, computer mediated communication techniques used in online social conversations are adaptations from the fact-to-face conversations in the physical world. Social actions were vital for the survival of human hunter gathers since the early ages of human civilisations. However, social actions and exchange of social information increased significantly with the use of language as it enables the individuals to convey complex expressions and have lengthy conversations.

Online social media platforms are also digital adaptations from the social media platforms in the physical worlds. The first form of a social media platform could simply be a *campfire* where the hunter-gatherer groups gathered around at night. The relaxing, warm and safe environment around a campfire leads to social chit-chat in-contrast to work related discussions during day-time (Wiessner, 2014), leading to increased social interactions among the group. Similar social media platforms were present in coffee-houses (mainly in Europe) and tea-houses (mainly in Asia) which serves as centres of social interactions (Standage, 2013) where people chat with other visitors while sipping tea or coffee for hours. These interactions although seems casual, were important to formulate ties with individuals for economic or political benefit.

These social actions and social media platforms in the physical world have been studied for decades and theorised by scientists across many disciplines that have developed an in-depth understanding of social actions and their underlying causalities. Such causalities have been abstracted into multiple layers of the human decision making process, where abstract social actions are represented as a handful of social behaviours driven by social needs perceived through human cognition. This understanding has developed incrementally over many decades and collectively by many theories which provide explanations to the social actions and their causalities based upon cognition, social needs and social behaviours. Considering online social behaviours are adaptations of social behaviours in the physical world and the causalities that drive such behaviours are the same, this research proposes to incorporate those theories to enhance the computational techniques (e.g., machine learning techniques, natural language processing techniques) that captures insights from social data in online social media platforms. For instance, sharing of different types of personal information (e.g., demographics, emotions) can be better understood in terms of the social behaviour called *self disclosure*, while providing advice in online support groups

represent the social behaviour *altruism*. Such incorporation would enable the computational techniques to effectively capture patterns from social data as well as meaningfully aggregate such patterns based on the underlying causalities.

Moreover, capturing insights into human behaviours and their causalities from social data, would enable social studies on human behaviour to use traces of human behaviours encapsulated in massive volumes of social data accumulated in online social media platforms. Conventionally, such studies on human behavioural causalities are conducted using controlled social experiments and natural experiments (observational studies) (Neuman, 2013). Controlled experiments are designed to study the cause and effect of a certain condition on individuals against a control group while in natural experiments the behaviour of a group of individuals are observed in their natural environment. In both types of studies the outcomes are acquired from the individuals either by recording their behaviour or using interviews/survey instruments.

Meshi et al. (2015) point out several key advantages of the use of social data to understand human behavioural causalities, in contrast to the use of conventional social experiments.

1. **External validity:** It is the generalisability of the casual inference (Drost et al., 2011) from the *experimental setting* to the *natural setting*. External validity is one of the key challenges of controlled social experiments, mainly because human behaviour is inherently complex and irregular. It is difficult to exclusively isolate cause and effect of a certain behaviour in a controlled experiment. Therefore, the casual inference obtained from experimental settings may differ in a natural setting. In contrast, archived social data is the retrospective accounts of social actions which occurred in a *natural setting*.Thus, causal inferences obtained using social data does not have the bias of any experimental settings.

2. **Less recall bias:** Is the error caused by the human recollection of past experiences. Self experiences are stored in the *episodic memory* as emotional and contextual details of the experience. However, *episodic memory* is gradually forgotten over the passage of time. Kahneman and Riis (2012); Kahneman et al. (1999) argue that recollection of an experience is biased towards the most intense aspects of that experience. Therefore, retrospective report of an experience is different than being reported shortly afterwards (Robinson and Clore, 2002). This recall bias affects the conventional studies since the data collection often happens periodically and the participants were asked to recollect their accounts of experiences that happen sometime back. In contrast, social data is often recorded near real-time, mainly due to the ease of access to online social media platforms using mobile devices. Therefore, social data is more immune to recall bias.

3. **Large sample size:** Conventional social studies were often conducted using a handful of individuals, mainly due to the associated cost. Since human behaviour is so diverse across individuals, the casual inference learned from a small sample could be over-fitting to that selected cohort. On the other hand, social data of millions

of individuals are already being archived in online social media platforms, who are from diverse socio-economic backgrounds. Hence, social data enables social studies using large and diverse cohorts.

In a nutshell, social data can be considered as traces of online social interactions which enable the use of theories on social behaviours to better understand social data. This understanding can be leveraged to develop alternative approaches that better transform social data into insights representative of underlying social behaviours. These two approaches to transform social data is depicted in Figure 3.1, which shows a. the conventional approach of applying machine learning techniques on social data to produce data-driven insights, and b. the alternative approach of combining machine learning techniques with social theories to transform social data into insights representative of underlying social behaviours. Furthermore, Figure 3.1.c shows that the insights representative of underlying social behaviours learned from the proposed approach can be leveraged by social sciences to better understand human social behaviours in general and especially in online social media platforms. However, this feedback loop to improve the understanding of human social behaviours is beyond the scope of this thesis which is focused on building technical capabilities for harnessing insights from social data.

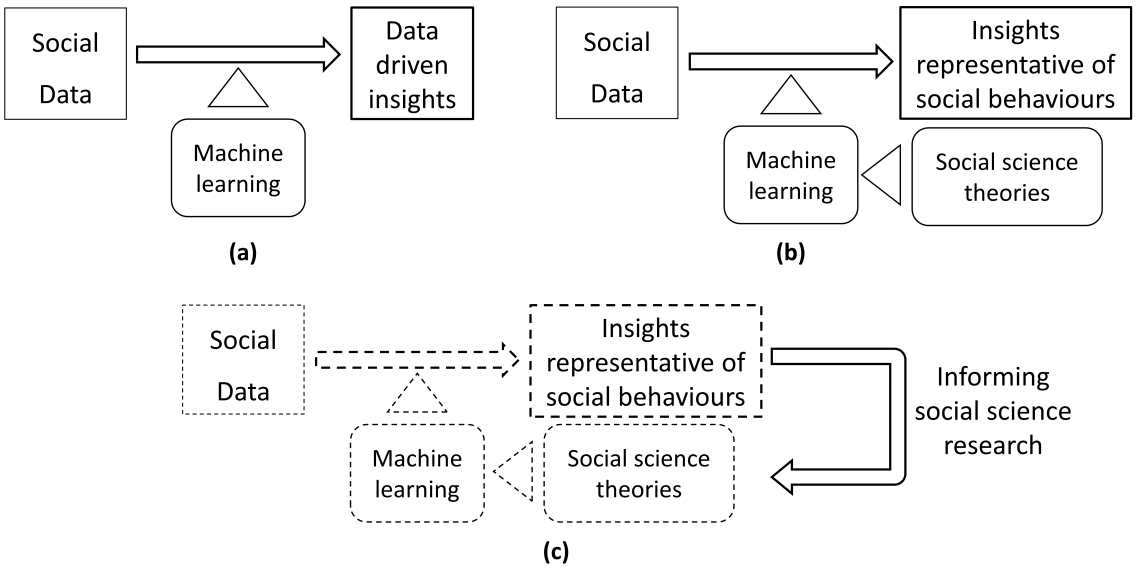


Figure 3.1: The approaches of transforming social data: a. the conventional approach of using machine learning techniques on social data which yields data-driven insights, b. combining machine learning techniques with social theories to transform social data into insights representative of underlying social behaviours, c. using insights representative of underlying social behaviours to better understand human social behaviours.

The new approach specified in Figure 3.1.b is further elaborated in Figure 3.2. Social science theories can be used to build latent representations from social data. Such representations can be based on either internal or psychological aspects such as emotions as well as external aspects such as topics of conversation or behaviours or events. Subsequently, machine learning and natural language processing techniques can be used to transform those latent representations into insights.

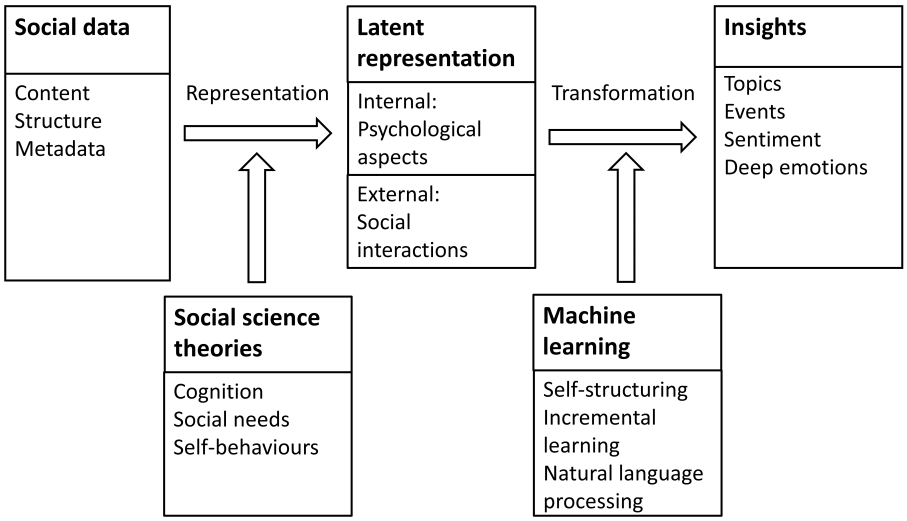


Figure 3.2: Incorporating social science theories and machine learning techniques to generate insights from social data.

The next section presents a hierarchical conceptual model that enables to represent social data as a layered latent representation which is derived from prominent social science theories.

3.2 The proposed conceptual framework

Figure 3.3 presents the proposed conceptual framework that can develop a latent representation of social data. The following subsections describe the layers of the conceptual framework and supporting theories from the social sciences.

3.2.1 Cognition

Human cognition is the mental process of acquiring or perceiving knowledge from thought, experience and senses. Cognition is thought to have physically based in the neural networks of the human brain within its over 100 billion neuron cells and their synaptic interconnections. In fact it is argued that cognition represents the collective functionality of all neural networks.

The cognitive process of acquiring or perceiving knowledge identifies behaviours that lead to positive or desired outcomes. This cognitive process dictates the execution of such desired behaviours by stimulating the motor neurons. These characteristics of cognitive process lead to the cognitive theory of human needs (Rosenfeld et al., 1992), which states that, from a cognitive perspective, a need is an abstraction of the desired outcome obtained using a set of similar behaviours. Therefore, needs originate during the cognitive process of perceiving knowledge.

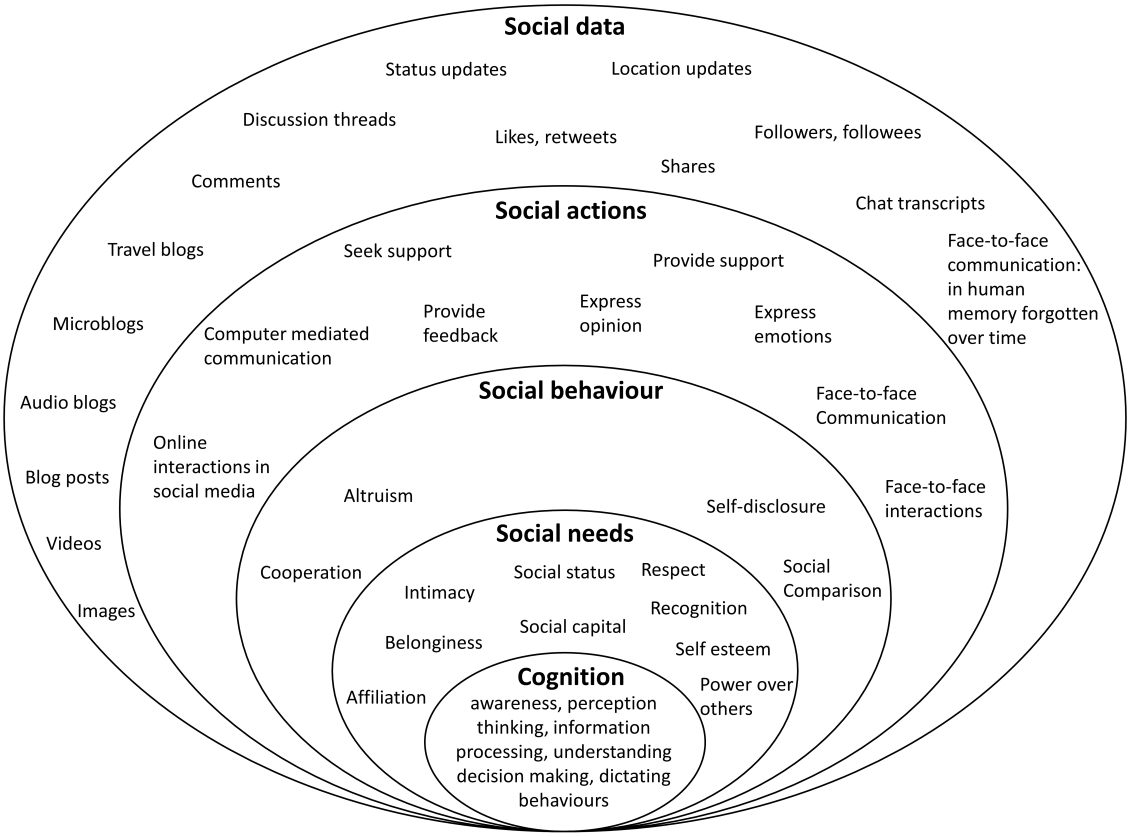


Figure 3.3: The conceptual framework for understanding social data using social theories of human social behaviours, social needs and cognition as the foundation.

3.2.2 Social needs

Based on this understanding of how cognition formulates human needs, this section discusses prominent human needs that involves social aspects i.e., the needs that have an emphasis on social interactions with other individuals or groups.

Evolutionary and biological perspective

It is argued in Darwinian evolution theory (Darwin, 1859) that social needs developed as a survival trait in human hunter-gather groups. Since, humans do not possess exceptional features such as teeth, strength or speed compared to other hominoid relatives such as gorillas or chimpanzees, their survival against the selective pressure of natural selection is probably the behavioural adaptation of belonging to a group. Although humans are now an ecologically dominant and no longer under threat of a predatory species, effective social interactions remains a key factor to survive or rather thrive among others in the modern complex society.

A similar idea is proposed by the social brain hypothesis, where Dunbar (1998) states that human intelligence initially evolves as a behaviour adaptation to survive as groups. The larger brains in humans is a result of increased computational demands of living in large social groups. In fact it is found that the optimum group size of primate social

groups correlates with the volume ratio of the neocortex (volume of neocortex relative to the brain) (Dunbar, 1998, 1993; Dunbar and Shultz, 2007).

The attachment theory (Bretherton, 1992; Bowlby, 1969) also describes that social needs of a human originate from the filial bonds developed between an infant and his close caregivers. It describes that an infant needs to develop a close and continuing relationship with at least one primary caregiver i.e., mother, in order for safety as well as cognitive, social and emotional development of the infant.

Psychological perspective

One of the earliest mentions of social needs appears in *pyramid of needs* presented in Maslow's theory of motivation (Maslow, 1943, 1962) which presents a hierarchy of human needs. The hierarchical formulation represents that humans look for needs in a particular layer only when the needs in respective lower layers are satisfied. In this pyramid of needs, socially relevant needs are placed third and fourth after physiological and safety needs. The third layer captures the human needs to feel a sense of belongingness among other individuals or social groups. Such relationships can range from close intimate relationships to work-related companionship. The fourth layer is self-esteem which is a sense of being respected and valued by others. It includes acquiring social status, fame or recognition from the groups that the individual is a member.

Alderfer (1969) formulated the Existence, Relatedness and Growth theory (ERG) by simplifying the Maslow's pyramid of needs. In ERD, relatedness is the desire to establish and positively maintain interpersonal relationships with other individuals. Such relationships could serve the need for belongingness as well as the need for self-esteem, thus relatedness is similar to third and fourth layers of *pyramid of needs*. However, in contrast ERD is not hierarchical and expect all three needs to exist simultaneously.

The Three Needs theory (McClelland, 1967, 1987), describes three types of social needs: the need for achievement, affiliation and power. The *achievement* need requires work on doable tasks and receive frequent feedback from others on the accomplishments. The *affiliation* is the need to establish and maintain social relationships; and feeling of belonging to social groups. The *power* need is the need to have power over others, social status and recognition

These different psychological theories of needs highlight the key social need as being belonging to social groups and maintain social relationships with other individuals, and within those relationships there exists need to receive feedback and recognition of achievement as well as the need to influence others.

In summary, social needs are a set of human needs that often need to be nurtured by others in society. Some needs such as love and belongingness often provided by those who are intimate to the individual (often family and friends of that individual). In contrast, social needs such as achievement, affiliation and power are related to receiving a form of recognition from society. Such needs are often satisfied by being part of different social groups.

3.2.3 Social capital

Social capital looks at social needs in the economic perspective, as a capital that each individual should possess to survive and thrive in modern society. Social capital describes the resources embedded in social groups as a form of capital i.e. an asset that can be economically useful and can be invested for future economic gain. Unlike other forms of *capitals*, that can be measurable, social capital is more conceptual and less measurable.

Over the past decade, many sociologists have provided slightly contrasting definitions on social capital and its utility (Woolcock and Narayan, 2000; Lin, 2002; Adler and Kwon, 2002; Putnam, 2001; Nahapiet and Ghoshal, 2000). Adler and Kwon (2002) define social capital as the *goodwill* (trust, reciprocity and sympathy) extended to an individual due to his social relations with individuals or groups, that will be available as information, influence, and solidarity.

Nahapiet and Ghoshal (2000) describe three types of social capital based on the form of existence as *structural, cognitive*, and *relational*. The structural dimension is based on the structure of the social groups which includes roles, rules and procedures of the social group that determines the approachability to the resources in the group. The relational dimension is based quality and nature of the relationships which determines the trust, obligations, expectations and sanctions (Putnam, 2001). The cognitive dimension is based on the shared reality (Bourdieu, 2011) which determines the shared values, beliefs and attitudes.

Another dimension of social capital is often defined based on the type of the relationship, which is initially defined as *bonding* or *internal* and *bridging* or *external* social capital (Putnam and Feldstein, 2009; Putnam, 2001; Adler and Kwon, 2002), base on the work *strong and weak ties* by Granovetter (1973). Bonding social capital arises from the relationships between people with similar characteristics i.e., homophily. Such individuals often have strong relationships with each other and peruse collective goals, thereby bonding capital facilitate support in terms of material, emotional and safety. In contrast bridging social capital arises from relationships between different social groups or individuals with different social identities. Bridging capital often operates on trust and reciprocity, and it provides access to important resources that are not available by bonding capital (Granovetter, 1973).

Later Szreter and Woolcock (2004); Woolcock et al. (2001) introduced *linking* social capital which is argued to be a special form of *bridging* social capital but the relationships and interactions occur across power or authority gradients of the society. The *linking* social capital mainly refers to the relationships with who have the decision making power over the individual, thus such capital provides access to services, jobs and resources (Szreter and Woolcock, 2004).

Social capital needs maintenance, where the individuals have to periodically renew the social links with others to make such links effective (Adler and Kwon, 2002). Moreover, leaving social groups reduces the social capital associated with that group (e.g., leaving an organisation) (Glaeser et al., 2002).

Initially, researchers speculated that the use of online social media platforms will reduce the time available for gaining and maintaining social capital. However, as new evidence emerge, now it is argued that use of online social media platforms not only provide an avenue to maintain social capital gained offline, it enables users to gain new social capital which were inaccessible offline due to barriers such as geographic proximity (Ellison et al., 2007). It is believed that online social media platforms mainly contributes to forming and maintaining *weak ties* with friends, colleagues and others with similar interests. Hence, most of the online social capital gains are in the form of bridging social capital (Resnick et al., 2001). However, some of the *weak ties* formed online may become stronger overtime leading to bonding social capital. Also, social media platforms help to maintain bonding capital when the individuals are geographically distanced (Burke et al., 2011). Ellison et al. (2007) also provide evidence that online social media platforms enables users with low self-esteem to achieve bonding social capital and improve their psychological well-being.

The Table 3.1 summaries the key theories discussed in relation to social needs.

Table 3.1: A summary of the key theories related to social needs discussed in this section.

Darwinian evolution theory (Darwin, 1859)	Social needs developed to support belonging to a group as a survival trait against the selective pressure of natural selection.
Social brain (Dunbar, 1998) (Dunbar and Shultz, 2007)	Human intelligence evolves as a behavioural adaptation to survive as groups and larger brains is a result of increased computational demands of living in large social groups.
Attachment theory (Bretherton, 1992) (Bowlby, 1969)	Human social needs originate from the filial bonds developed between an infant and close caregivers.
Theory of motivation (pyramid of needs) (Maslow, 1943) (Maslow, 1962)	Humans needs are hierarchical, reach higher order needs only when low order needs are satisfied.
Three Needs theory (McClelland, 1967) (McClelland, 1987)	Humans have three types of social needs: need for achievement, affiliation and power.
Social capital (Woolcock and Narayan, 2000) (Adler and Kwon, 2002) (Putnam, 2001)	Social capital describes resources embedded in social groups as a form of asset that can be economically useful and can be invested for future economic gain.

3.2.4 Social behaviour

Behaviour is defined as an intentional act to achieve a perceived goal (Ajzen, 1985). This goal oriented nature of the behaviour arises from having perceived needs which individuals plan to satisfy through their own behaviour. The theory of planned behaviour (Ajzen, 1985, 1991) states that execution of a behaviour is guided by the individual perceived

attitude towards that behaviour, perceived subjective norms of that behaviour and finally the perceived ability to control its execution.

Constructing on top of the general theory of behaviour, social behaviour can be defined as an intentional act that involves interactions between one or more individuals to achieve the perceived social needs of those who involved. Therefore, social behaviours are intentional acts of achieving social needs discussed in the previous section. Among social behaviours communication or interpersonal communication is the foundational social behaviour which is essential for any form of human social interaction. Other prominent social behaviours include self-disclosure, cooperation and social comparison. All these social behaviours are observed in the natural world as well as in online social media platforms. However, differences in the online social environments have led each social behaviour to have several adaptations from the natural environment to online social environments. The following subsections discuss the above mentioned social behaviours in terms of theoretical perspective, social needs they attempt to achieve and the different adaptations of them for online social media platforms.

Interpersonal communication

Interpersonal communication or simply communication is the act of transmitting information among individuals or groups. Any form of social interaction between humans involves passing information from an individual to others (intentionally or unintentionally). Therefore, interpersonal communication can be considered as a foundational social behaviour that itself is a social behaviour as well as required for all other social behaviours. The key dimensions of communication are sender, channel, message, and the receiver (Shannon, 1948; Schram, 1954). Sender encodes the desired information in a message and sends it to the receiver using an appropriate communication channel. The receiver decodes the message to receive the information. The encoding and decoding schemes have to be the same for the receiver to get the intended message, while differences in encoding and decoding may lead to unintended interpretations of the message received (Schramm, 1954).

Although communication happens among all living organisms, this discussion is exclusive to the behaviour of *interpersonal communication* which occurs between individuals or groups of humans. Conventionally, interpersonal communication happens face-to-face (FtF) where individuals visually see each other and exchange information using a combination of both verbal (using natural language) and non-verbal (e.g., body language, facial expression) communication. Verbal communication is often less ambiguous as language rules and the meaning of the words/phrases are well defined and common knowledge to all communicators. In contrast, the meaning of non-verbal communication often deviates based on different attributes (e.g., past experience) of the receiver.

Interpersonal communication as a social behaviour can be explained using the *uncertainty reduction theory* (Berger and Calabrese, 1975; Berger and Bradac, 1982). It states that during social interactions, individuals have a need to reduce the uncertainty about others, thus use interpersonal communication as a mean of acquiring more information

about them. Berger and Bradac (1982) further state that uncertainties exist as (i) *cognitive* which are the uncertainties on values, ideas and attitudes of each other; and (ii) *behavioural* which are the uncertainties on behaviours of each other in a given situation. Both verbal and non-verbal communication with each other leads to the exchange of information that leads to the reduction of such uncertainties and improvement of trust. Such reduction of uncertainties and improvement of trust is essential to initiate and establish social relationships with the society. Effective and frequent communication with closed ones (family and friends) is essential to maintain intimate relationships to achieve the social needs love and belongingness (bonding social capital). Similarly, communication with different social groups and broader society is the key to establish affiliations thereby improve bridging social capital. Also, interpersonal communication is used as a way of influencing others which yields influencing power. Therefore, interpersonal communication is a fundamental behaviour to achieve most of the social needs.

Computer mediated communication

Computer mediated communication (CMC) is the process of two or more individuals communicate using a computer platform. It is the communication process used in all online social media platforms. In early days it is primarily text based, however, recent CMC platforms allow a range of audio and video features to enhance the communication experience. Interaction in CMC can be either synchronous (e.g., Internet Relay Chat, Skype, Facebook Messenger) as well as asynchronous (e.g., email). Modern online social media platforms support both approaches. For instance, in Facebook platform, posts and status updates are mostly used for asynchronous communication while Facebook Messenger can be used for synchronous communication.

Most of the earlier research looked down upon CMC as an inferior communication channel in contrast to face-to-face communication (F2F). One of the prominent theory on that line of thought is the Media Richness Theory by Daft and Lengel (1986) and Trevino et al. (1987), which defines the richness of information as "the ability of information to change understanding within a time interval". The theory delineates that the *richness* of a communication medium is its capability to effectively transmit information to alter the understanding of an individual or a group (Daft et al., 1987; Daft and Lengel, 1986; Trevino et al., 1987). The theory further states that richness of medium is a function of the following capabilities of a medium: (i) transmit multiple communication cues, (ii) provide rapid responses, (iii) have a personal focus and (iv) support natural/conversational language for communication. These four capabilities of a rich medium is formulated considering F2F communication as the baseline standard and any other medium is evaluated based on how close it can resemble F2F communication.

One of the key criticism on media richness theory is that it focuses on the communication medium objectively and haven't accounted for different context/topic of the communication or the differences in the human actors involved in the communication process (Kock, 2005; Kahai and Cooper, 2003). In order to accommodate those features Carlson and Zmud (1999) proposed Channel Expansion Theory, which states that individuals perceptions on

the *richness* of a particular communication channel dependent on his relevant experience of using that channel and familiarity with the topic of discussion. Hence, frequent and prolonged use of a particular CMC channel in a particular context makes it *richer* for the associated individuals. Tidwell and Walther (2002) report that such improvements of the perception of richness of a CMC channel may result similar of greater richness than F2F communication.

Another key theory that discusses about the CMC medium is the Media Naturalness Theory which is proposed by Kock (2004, 2005) as a psychobiological model. It argues that humans (hominoids) have conducted F2F communication inside social groups as an act of survival. Hence, the evolutionary pressures over a long time period would have developed adaptations in both brain (e.g., sensory and motor neurons) and body (e.g., facial muscles, visual and auditory organs) to make F2F communication cognitively efficient, less ambiguous, and result in better physiological arousal. For example, human ears are sensitive to human voices than any other acoustic stimuli (Nass and Brave, 2005).

Kock (2004, 2005) further argues that due to the adaptive capability of the human brain, with frequent and prolonged use of a leaner CMC medium, human brain is capable of adapting to use such medium naturally with lesser cognitive effort, less ambiguity and increased physiological arousal. This phenomenon aligns with Channel Expansion Theory (Tidwell and Walther, 2002) which states similar outcomes with frequent and prolonged use of a CMC medium.

Several such channel expansion attempts have been clearly visible across the recent history of social media platforms such as emoticons and hashtags. Emoticons in short for *emotion icon* are a pictorial representation of facial expressions using a combination of characters. The most famous emoticons- smiley face :-) and frowning face :-(are introduced to online communication in 1982 as a suggestion to mark statements as jokes or serious statements. The use of emoticons has since spread across all text based communications channels such as emails, SMS and all social media platforms. A large set of emoticons are currently in use and also extended to multiple cultural variants such as Japanese *kaomoji*. Also, emojis are pictorial versions of emoticons that often provides a more lively feeling of facial expression. The main use of emotions/emojis is to reduced the ambiguity of discourse resulted by not having facial expressions and improve the social presence in text based CMC (Walther, 1992; Derks et al., 2007). Therefore, emotions/emojis expand CMC channel by providing a means to transmit facial expressions.

Another prominent channel expansion attempt is social tagging or folksonomy which is the practice of adding user generated tags to content published online. Such content published in social media platforms are mostly unstructured text and published in a less organised manner without adhering to any form of a taxonomy. Hence it is a daunting task to find content or discussions that are relevant to a given topic of interest. It also limits the formation of online social groups with similar interests. Social tagging is one of the solutions adopted by the online community to mitigate this limitation. Social tags start with # symbol and often comprises of a word or a concatenated phrase that the user wants to highlight. Use of social tags originated within the IRC (Internet Relay Chat)

community to signify different chat rooms. However, its adaptation into Twitter [1] in 2007 has lead to its widespread use across most of the prominent social media platforms. As pointed out in (Zappavigna, 2015), social tags serves different purposes such as: (i) highlighting the semantic domain of the post (e.g., #bigdata), (ii) linking the post to an existing social discussion (e.g., #auspol), and (iii) highlighting the affect of the post (#sad).

These examples highlight that although CMC is a leaner medium compared to F2F, humans have invented different channel expansion strategies to improve its richness and provide a better social presence.

Self-disclosure

Self-disclosure is another social behaviour, which is the act of disclosing self to an individual or a group (Jourard and Lasakow, 1958). Such disclosure can be intentional or unintentional, voluntary or by request, and verbal or non-verbal. The disclosure of information includes any type of information about the self such as socio-demographic information, emotions, opinions and needs (Cozby, 1973). Self-disclosure a key social behaviour that reportedly consumes around 30-40% of everyday human conversations (Dunbar et al., 1997). No other species including other primates engage such high level of self-disclosure, which suggests that humans may have intrinsic motivation to self-disclose personal information. In fact, recent studies on brain reveal that self-disclosure lights up mesolimbic dopamine system that relates to reward and pleasure indicating that self-disclosure is a rewarding experience for an individual (Tamir and Mitchell, 2012).

Self disclosure is often studied on long terms strategic relationships such as romantic relationships (Sprecher and Hendrick, 2004; Derlaga and Berg, 1987; Greene et al., 2006) and therapeutic (therapist-patient) (Derlaga and Berg, 1987; Knox et al., 1997; Hill and Knox, 2001). In addition, there are studies on spontaneous self disclosure to a complete stranger which is commonly known as the 'stranger on the train' phenomena (Rubin, 1975).

There are several theories that describe self-disclosure process and its objectives. Among them, social penetration theory (SPT) proposed by Altman and Taylor (1973) is the most widely regarded. The social penetration theory (SPT) (Altman and Taylor, 1973) was developed to explain how interpersonal communication develops from shallow to intimate through reciprocal exchange of self-disclosed information. SPT highlights two key dimensions of self-disclosure- *breadth* and *depth*. Breadth is the variety of topic in self-disclosure which could include topics related to profession, political affiliation etc. Depth is how deep an individual would reveal on a certain topic, where some disclosures are more of surface information while others disclose deeper information about the individual. SPT describes this multi-dimensional and multi-layered nature of personality using the 'onion metaphor'(Figure 3.4), where each layer of the onion represents different layers of personality. The outermost layer contains more visible and public information such as socio-demographics, while inner layers progressively contain more intimate information

[1] https://www.cmu.edu/homepage/computing/2014/summer/originstory.shtml

such as values, attitudes, emotions and goals. SPT states that self-disclosure process is similar to peeling the onion layer-by-layer, where it initiates by disclosing more visible information and then progressively disclose more intimate information as the interpersonal relationship grows.

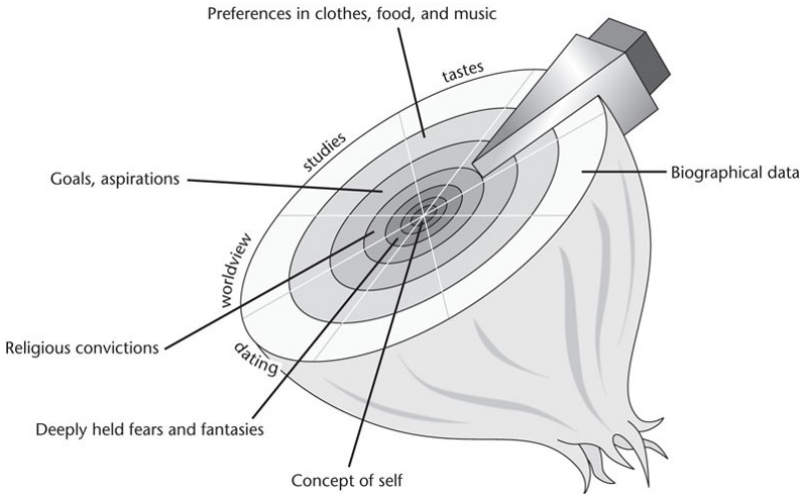


Figure 3.4: The onion like nature of personality (Altman and Taylor, 1973), where outer layers contain more visible information and inner layers contain more intimate information. *image source: Wikipedia*

Self-disclosure serves multiple social needs discussed in Section 3.2.2 which includes affiliation, intimacy, belongingness and recognition. Self-disclosure is a key behaviour required in formulating and maintaining any form of affiliation. Different topics are self-disclosed in different affiliations. For example, work related information such as skills and qualifications could be disclosed in professional affiliations, while sports team or player could be disclosed for affiliation related to fan clubs. The depth of self-disclosure is often shallow for affiliations. However, with reciprocity and over time self-disclosure goes deeper in some affiliations, making them more intimate relationships (become close friends or family) which often serves social needs of intimacy and belongingness. Also, self-disclosing information about achievements is essential to get recognition from individuals and social groups.

Self-disclosure is a prominently observed social behaviour in social media platforms as well. It is reported that sometimes self-disclosure amounts to around 80% of the content in online social media platforms (Naaman et al., 2010). This is known as the *online disinhibition effect* (Suler, 2004) where individuals self-disclose more frequently and intensely in online communication in contrast to face-to-face communication. Suler (2004) identified few factors that lead to this online disinhibition effect, which are dissociative anonymity, invisibility, asynchronicity, empathy deficit, dissociative imagination, and minimization of authority. Anonymity and invisibility are similar to the *stranger in the train* phenomenon where the individuals feel safe to self-disclose more in online social media platforms when their identity is not or loosely linked into the actual person. Similarly, asynchronous nature of the online social media platforms let the individuals self-disclose without worrying about an immediate response of others.

Suler (2004) further claims that the online disinhibition effect can be both benign and toxic. Benign online disinhibition is when the individuals use online social media platforms to express their feelings, issues, concerns etc. which they are reluctant to disclose in face-to-face conversations, potentially due to either embarrassing nature of the disclosure or introvert personality. It is found that such disclosures improve the mental well-being by reducing depression and anxiety among such individuals. Also, Davis (2012) found that some individuals (especially with introvert personality) can perform social interactions better in online social media platforms which enable them to achieve their belongingness and affiliation social needs as well as to improve their social capital.

Toxic online disinhibition happens when individuals use online social media platforms disclose hostility, hatred and threats targeted towards an individual or even a particular socio-ethnic group. Dissociative anonymity and invisibility in online social media platforms enable individuals to ignore the social norms that they often conform to in face-to-face conversations and disclose such toxic thoughts without any restrain in online social media platforms (Lapidot-Lefler and Barak, 2012).

Although self-disclosure is more frequent in online social media platforms, depth of self-disclosures varies across different platforms depending on a multitude of factors such as visibility (whether posts are public or only visible to members), strength of ties between dyads, homophily, and reciprocity. Bazarova and Choi (2014) have investigated self-disclosure in wall-posts and peer-to-peer messages in Facebook and shown that public wall-posts contain more peripheral information and less intimate information while peer-to-peer messages contain more intimate information and less peripheral information. This highlights that although individuals self-disclose more in social-platforms, they mostly control the depth of the self-disclosure for different audiences or different platforms.

Cooperation and altruism

Cooperation is the behaviour of working as a group to attain a goal that benefits everyone involved. Altruism is defined as an act which only benefits the recipient while incurring some cost to the actor (Hamilton, 1964). Both cooperation and altruism exist in the animal world in tasks such as hunting and bringing up offspring, however, their scale and scope in humans society are unmatched to any other species. Humans have the capacity to cooperate as large groups of genetically unrelated individuals (Boyd and Richerson, 2009) and the capability to cooperate in complex tasks.

In social brain hypothesis (Dunbar and Shultz, 2007) it is argued that the increased cooperation and altruism in humans is an evolutionary trait for survival as complex social groups. Another argument is that cultural and social norms (Fehr and Fischbacher, 2004), supporting prosocial emotion and the establishment of the legal system to punish those who violate such norms (Boyd et al., 2003, 2010) have contributed to increased cooperation and altruism.

Cooperation and altruism are often practised as a way to develop affiliations with social groups. It enables social interactions with the group members which often improve and establish interpersonal relationships with the other group members, thus improves the

bridging social capital of the individual. Such behaviours also lead to achieving recognition from the affiliated social groups. Also, cooperation and altruism with social groups enable the individual to take a leading role and influence others.

Behaviours related to cooperation and altruism are often observed in online social media platforms as well. Individuals form online social groups based on different interests and perform cooperative tasks. In contrast to social groups in physical world, online social groups are not bound by any geographic boundaries. Hence, online social media platforms enable individuals to perform cooperative tasks beyond geographical boundaries.

Large social media platforms such as Facebook and Twitter often becomes the birthplace of social movements (Golder and Macy, 2014) which originates as people discussing or sharing their views on certain socio-political issues. Others with similar views join those discussions. Some of those discussions transition into social movements either facilitating or demanding a corrective action for the issues. Tunisian and Egyptian revolutions that has overthrown the then governments are prominent examples for such social movements. Those social movements prominently used Twitter for organising protests/demonstrations (e.g., logistic coordination) and spreading news and updates to both local and global audience in near real-time (Lotan et al., 2011).

Another avenue for cooperative and altruistic behaviours are the online platforms for knowledge sharing where individuals from across the globe, contributes voluntarily by publishing user generated content. Such content is mostly based on their personal experience or professional experience. One of the well-known examples for such projects is Wikipedia which is a massive online encyclopaedia contributed by thousands of volunteers with different areas of expertise (Golder and Macy, 2014). In Wikipedia its community is delegated with multiple roles covering content generation, editing, validation and administration which drivers content generation as well as the integrity of the content.

Similarly, in other virtual communities such as self-help groups, cooperation and altruism displayed when individuals seek support by providing informational support and emotional support based on their personal experiences (Ziebland and Wyke, 2012). Unlike in Wikipedia most of the other virtual communities does not have set roles. However, each piece of information is either supported or opposed by other participants leading to community validated response.

Cooperative acts in online social media platforms primarily provide a sense of affiliation to the members of that online community. Especially when such contributions are commended by other members of the community. Moreover, online knowledge sharing communities have different approaches to provide social recognition to the contributors. For example, some platforms maintain contributor ratings either granted automatically based on the number of articles/posts or provided by the readers based on the usefulness of the contributions. In addition highest contributors receive privileged administration and moderation roles to assess the suitability of user content and user behaviour in the community. They have the power to remove inappropriate content as well as suspend users for inappropriate behaviour. Such roles provides a sense of power and influence over others in that online community.

Social comparison

Social comparison is another key behaviour which is an act of self-evaluation of an individual's attributes such as opinions, abilities, appearance, wealth, performance etc. against those of other individuals (Suls et al., 2002). Such comparisons enable the individual to understand self-worth as well as reduce any uncertainties about the compared attributes. (Festinger, 1954) proposed the initial social comparison theory which denotes social comparison as a behaviour that happens against individuals similar to oneself. People use social comparison for self-enhancement and to gain self-esteem.

Social comparisons can be broadly categorised as upward (compare against someone better) and downward (compare against someone worse) (Collins, 1996). In upward social comparisons, the comparison happens against individuals who are doing better in the compared aspects, such comparisons yield motivation and inspiration for self-improvement. However, it may lead to poor self-evaluation, and negativity as well. In contrast, downward comparisons happen against individuals who are doing worse in compared aspects. Such comparisons improve their self-worth by comparing to others who are worse off (Taylor and Lobel, 1989; Wills, 1981).

Traditionally, social comparison happens among family, friends and co-workers whom an individual often encounter. In contrast, there are significantly more opportunities for social comparison in online social media platforms. This is because most of those platforms allow and encourage users to publish a user profile where user can self-created their own persona. For instance, in Facebook, there are millions of user profiles which can be freely accessible to anyone which contains personal information as well as collections of photos/videos that provides a sneak peek to the lifestyle of each individual. Similarly, professional social media platforms such as LinkedIn contains comprehensive work related profiles. Therefore, social comparison is significantly high on online social media platforms (Vogel et al., 2014) where individuals compare their *offline* self to *online* self of others perceived based on the information presented on online social media platforms. However, most of the profiles in online social media platforms are built using selected content to attractively present the ideal self of the individuals. For example, people only publish their best looking photos and highlights of their life; and avoid publishing any negative looking content (Walther, 2007; Vogel et al., 2014). Because of this selective self-presentation in online social media platforms social comparison is not *like for like*, but rather a disadvantage for the *offline* self. This phenomenon leads to excessive upward comparisons where more realistic *offline* self gets compared to meticulously presented *online* self of others; and often results negative self-evaluations, diminished self-worth and even depressive symptoms (Vogel et al., 2014; Nesi and Prinstein, 2015). Chou and Edge (2012) report that the impact of such comparison is higher when compared to an *online* self of people never or less frequently met in person, and lower for people with frequent encounter.

Table 3.2 provides a summary of the key theories discussed in this section. In summary, the above discussed social behaviours are designed to achieve different social needs. Interpersonal communication is the foundational behaviour that is required to enable any

Table 3.2: A summary of the key theories related to social behaviour discussed in this section.

Theory of planned behaviour (Ajzen, 1985) (Ajzen, 1991)	A behaviour is an intentional act to achieve goal(s) perceived based on the needs of individuals.
Uncertainty reduction theory (Berger and Calabrese, 1975) (Berger and Bradac, 1982)	During social interactions, individuals have a need to reduce the uncertainty about others, thus use interpersonal communication as a mean of acquiring more information.
Media richness theory (Daft and Lengel, 1986) (Daft et al., 1987)	Richness of a communication medium is its capability to effectively transmit information to alter the understanding of an individual or a group. A rich medium can transmit multiple communication cues, provide rapid responses, have a personal focus and support natural language for communication.
Channel expansion theory (Carlson and Zmud, 1999) (Tidwell and Walther, 2002)	Perception on the richness of a communication channel dependent on the experience of using that channel. Frequent and prolonged use of a channel makes it richer for the associated individuals.
Social penetration theory (Altman and Taylor, 1973)	Self-disclosure is like peeling an onion layer-by-layer, where it initiates by disclosing more visible information and then progressively disclose deeper information as relationship grows.
Social comparison theory (Festinger, 1954)	Individuals compare themselves to others as a form of self-evaluation to reduce uncertainty about the opinions and abilities of themselves.

form of social interaction with others while the other discussed social behaviours utilise communication. Social behaviours are observed in both the physical world and the online environment (online social media platforms) however, some differences exist due to the differences in the environments. Table 3.3 summarises such differences of social behaviours in physical and online social environments.

As shown in the conceptual model in Figure 3.3, the above discussed social behaviours are mainly abstractions of human actions originated to achieve the perceived social needs, Those abstracted behaviours are executed in a multitude of ways as *social actions* depending on different functional and environmental aspects which will be discussed in the following section.

3.2.5 Social actions

Social action can be defined as an execution of a particular social behaviour detailed in the previous section. Social action for a particular behaviour is a specialised execution of that behaviour considering different functional aspects and environmental aspects. Functional

Table 3.3: Comparison of social behaviours in physical world and online social media platforms.

Social behaviour	Physical world	Online social media platforms
Interpersonal communication	Face to face social conversations Verbal and non-verbal cues Facial expressions, eye contact Participants in close proximity Immediate response required	Computer mediated social conversations Mostly text based, no access to non-verbal cues No access to facial expressions or eye contact Not bounded by geographic proximity No immediate response required
Self disclosure	Often disclose to the participants in close proximity More control over the recipients of the disclosure	Disclose to all the members or a selected group in social media platform Often self-disclose in a higher breadth and depth due to online disinhibition effect
Cooperation and altruism	Groups are bounded by geographic proximity Group have to work synchronously	Groups are often formed based on interest, not bounded by geographic proximity Groups can work synchronously and asynchronously Often form relatively large groups
Social comparison	Compare different attributes of oneself to others Compare against friends, peers, and colleagues	Compare different attributes of oneself to the online presented self of others Access to personal information of large number of individuals in social media platforms to do comparisons

aspects are determined by the cognition of the individual based on the perceived costs and benefits of that action. Functional aspects are often intensity, depth and duration, which determines the desired scale of the execution. For example, in self-disclosure, the depth of self-disclosure which determine the level of the self to expose, which can be a shallow disclosure of political views or a much deeper disclosure of emotional expression, and the intensity of the self-disclosure determines how much information to disclose.

Environmental aspects are the aspects that are external to the individual and the individual has no or minimal control over. Environmental aspects are mainly relevant to the medium of execution. Social actions need to be executed considering the characteristics of the medium of execution, where it needs be optimised to employ the favourable characteristics of a given medium to exercise the social behaviour effectively to yield the desired outcome. For instance, in a rich communication medium such as F2F communication, social actions can leverage both verbal and non verbal to exercise the social behaviour, while in a leaner medium such as Twitter, social actions have to use different strategies to achieve similar outcomes.

The next section describes social data, which are the archived accounts of social actions that are accumulated in online social media platforms.

3.2.6 Social data

Social data is the archived accounts of social actions. Before the era of digitisation, accounts of social actions only exist in the memory of the participants and forgotten over time, unless being recorded on paper or dramatised as folk stories, which only happens to a handful of important social actions.

The development of digitisation and digital storage techniques have enabled the archival of social actions in the physical world in digitised format. Initially, digitisation techniques were expensive, devices were bulky, and required technical expertise. Thus, such techniques and devices were mostly possessed by professionals and only used to record a few significant social actions. However, over the last few decades, (i) digital storage techniques have become cheaper and accessible, and (ii) digitisation devices has become cheaper, compact and easier to use. For instance, cloud storage has become cheaper and globally accessible through internet services. Also, smart phones packed with multiple digitisation techniques (e.g., text, voice recording, camera) has become affordable to the general public. The widespread availability of these techniques has enabled the general public to record and archive the social actions in the digitised form.

Accumulation of social data

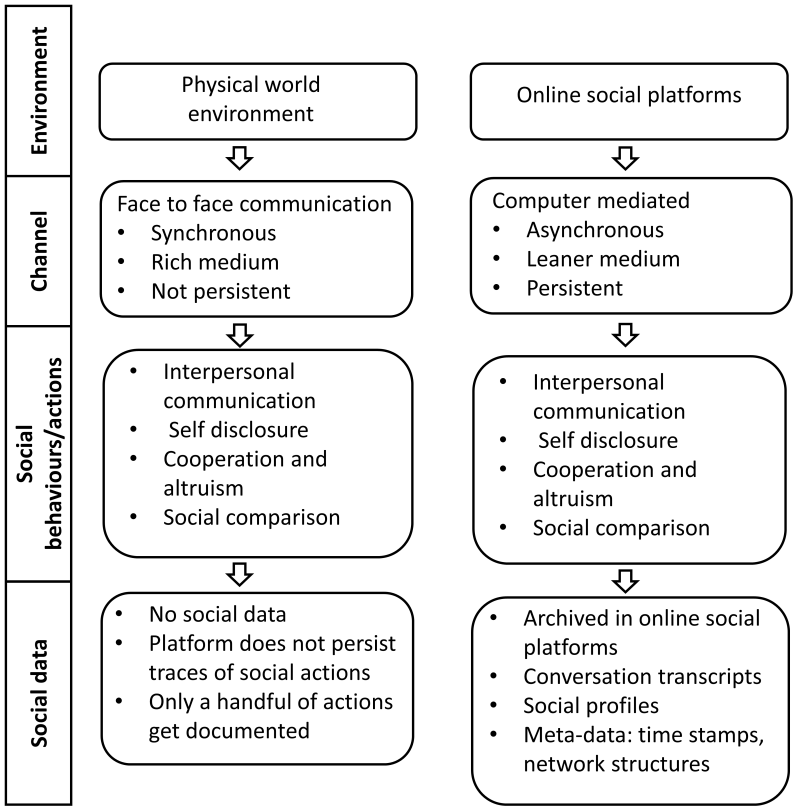


Figure 3.5: Accumulation of the traces of social behaviours in online social media platforms as social data.

Although the digitisation has enabled the archival of social actions digitally, such data are stored and distributed privately. Moreover, in most of the cases those are snapshots of selected moments of social actions which occur in the physical world.

The next paradigm shift comes with the rise of online social media platforms over the last decade. As discussed in previous sections, people start interacting with others on those online social media platforms and participate in social behaviours that are founded upon the social behaviours observed in the physical world. As shown in Figure 3.5 the CMC based channel in online social media platforms persist all interactions that happen through the platform, which include conversation transcripts, social profiles and metadata such as timestamps and social network structures. Those archived data can be considered as the traces of the corresponding social behaviours or social actions that took place on the online social media platforms.

3.3 An embodiment in self-structuring artificial intelligence

As discussed in Chapter 2 there have been attempts of qualitative methods as well as conventional machine learning approaches to transform social data into patterns that provide data-driven insights into social behaviours. However, such attempts yield suboptimal results due to a multitude of challenges present in social data that are discussed in Section 2.7. This thesis proposed a novel approach, the multi-layered conceptual model in Section 3.2. It is now pertinent to delineate the embodiment of this model, based on the paradigm of self-structuring artificial intelligence.

Figure 3.6 shows the key elements of the proposed framework which consists of (i) representing social data based on existing social sciences theories into useful latent representations such as emotions, topics and events, (ii) unsupervised self-structuring to automatically structure representations of social data into semantically coherent groups, (iii) incremental learning from social data streams to learn temporal changes, and (iv) natural language processing to extract semantic information about individuals and groups.

The unsupervised self-structuring machine learning techniques are employed to automatically generate representative structures from the feature transformed social data, based on the underlying similarities of the social behaviours present. Such structuring of social data highlights the common patterns of social behaviours among different social groups. These unsupervised algorithms are extended from the conventional machine learning algorithms to suit the challenges of the social data such as high velocity, diversity and unstructured nature.

Incremental learning techniques helps to capture temporal patterns which occur in social data streams at different granularities. For instance, at the individual level, changes are related to internal individual behavioural changes over time. Similarly at the group level, changes of group behaviour often are driven by external factors such as events.

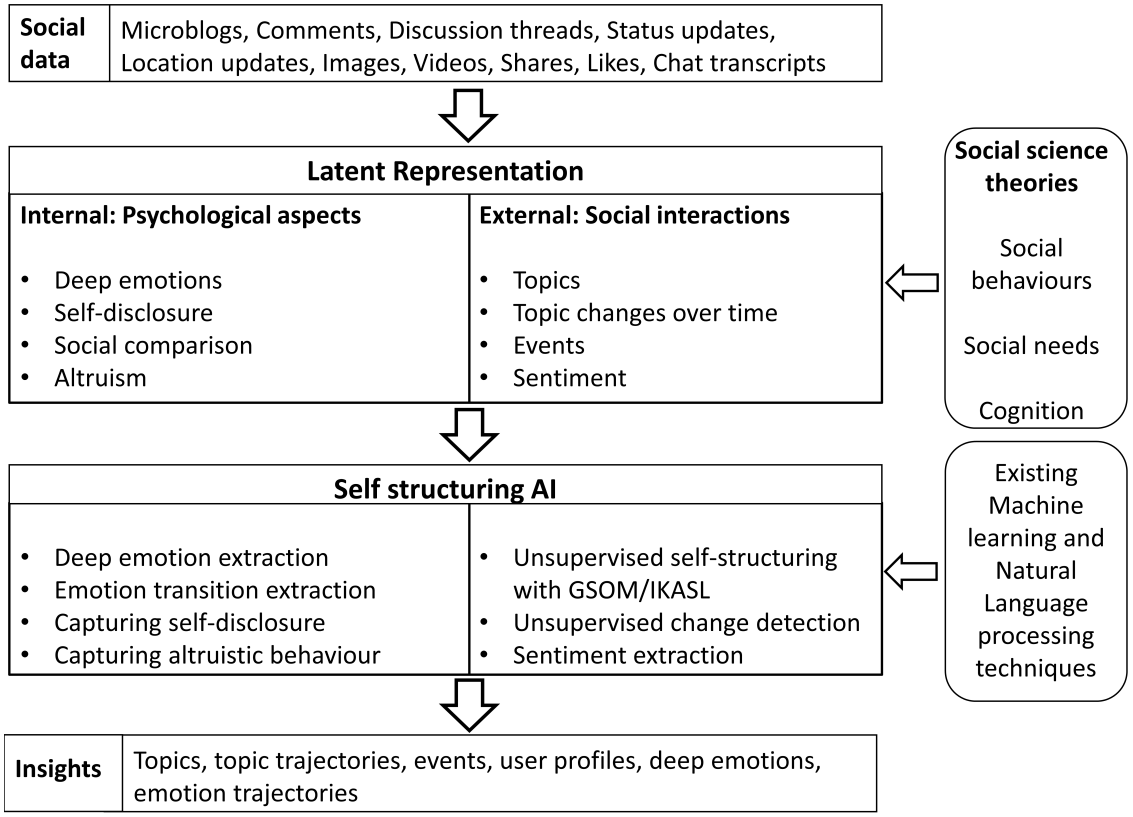


Figure 3.6: The proposed self-structuring artificial intelligence framework for generating insights from social data in online social media platforms.

The above framework is developed and implemented in the forthcoming chapters of this thesis, which describes the implementation of parts of the framework as well as its application on two social data sources with distinct characteristics.

Chapter four presents an implementation of the above framework for fast-paced social media platforms such as Twitter which mainly focus on the external aspects in Figure 3.6. Such social media platforms are mainly characterised by its fast-paced nature where discussions are used to rapidly disseminate information across the participants of the social media platform. The participants discuss topics related to recent events and express their opinion/sentiment. However discussion topics are changed frequently. An unsupervised self-structuring and incremental learning algorithm is developed that automatically structure the social data into coherent clusters based on the semantic similarity which exists in social data, and an event detection algorithm is developed to capture temporal changes of various behavioural change indicators.

Chapter five presents an implementation of the above framework for online support groups (OSG) which are slow-paced and show contrasting characteristics to Twitter. OSG discussions focus prominently on the internal or psychological aspects such as emotions specified in Figure 3.6. OSG discussions are tightly coupled to the main theme of OSG which is often an disease condition and participants engaged in longer discussions. Because of higher homophily and stronger ties people tend to engage in deeper self-disclosure such as expression of deeper emotions. Multi-granular emotional expression was structured

and incrementally captured overtime to generate individual emotional trajectories. Also, a suite of machine learning and natural language processing techniques are developed to capture different levels of self-disclosure.

Chapter six extends the algorithms developed in chapters five to gain insights from prostate cancer related OSG. Extensions include algorithms to capture the treatment decision making behaviour and decision factors.

3.4 Chapter Summary

This chapter developed a layered conceptual model for interpreting social data generated in online social media platforms and a self-structuring artificial intelligent framework to capture insights from social data.

The conceptual model is developed after an extensive analysis on existing theories of cognition, social needs and social behaviour. It depicts that social data is generated by social interactions happening in online social media platforms, and such interactions reflect different social behaviours driven by different social needs. It provides a deeper meaning to social data as archived traces of online social behaviours rather than just a corpus of unstructured text. In addition, four key social behaviours were conceptualised based on existing literature and a comprehensive comparison is provided comparing and contrasting exhibition of each behaviour in the physical world and online social media platforms.

The proposed self-structuring artificial intelligent framework is designed to overcome the challenges of social data highlighted in Chapter 2. It consists of four key elements (i) representing social data into useful latent representations based on above conceptual model , (ii) unsupervised self-structuring to automatically structure such representations into semantically coherent groups, (iii) incremental learning to learn temporal changes, and (iv) natural language processing to extract semantic information. The next chapter presents an algorithmic development of this self-structuring artificial intelligent framework to capture insights from fast-paced social media platforms.

Chapter 4

Ideation to creation: A new self-structuring artificial intelligence algorithm

No event can be judged outside of the era and the circumstances in which it took place.

Fidel Castro

This chapter presents the algorithmic development of the self-structuring artificial intelligence conceptual model proposed in Section 3.3. As discussed in previous chapters, social data accumulated on online social media platforms is a representation of social interactions and behaviours of individuals from diverse socio-demographic backgrounds.

This chapter is organised as follows. Section 4.1 describes the high-level design of the proposed algorithm. Section 4.2 focuses on incremental learning for topic separation and Section 4.3 presents IKASL: a recent unsupervised incremental learning algorithm that forms the foundation for the proposed algorithm. Section 4.4 presents the first technique of the proposed algorithm, for handling complexities of social text streams and Section 4.5 presents the second component, the event detection technique. Section 4.6 demonstrates the full algorithms using an illustrated example and Section 4.7 presents an empirical evaluation using two Twitter datasets.

4.1 The design of the algorithm

Given the diversity of social data generated by distinctive social media platforms, this development focuses on fast-paced online social media platforms such as Twitter which is a high volume/velocity text stream contributed by a diverse set of participants from all over the world. Also, social media messages are relatively short and highly unstructured.

Due to this fast-paced nature, such social media platforms are mainly employed for rapid dissemination of current and trending information through the social network to a large number of individuals (Kwak et al., 2010; Lotan et al., 2011; Lovejoy et al., 2012). Due to this rapid dissemination of information, any trending topic would capture the

attention of a large number of individuals quickly leading to a surge in discussions related to that topic. After a peak such trending topic often rapidly lose interest and discussions may switch to a different trending topic. As the primary intention is information dissemination rather than having continuing discussions, the ties between dyads are relatively weak and often lack homophily, reciprocity. Because of weak ties and lack of homophily, as discussed in Section 3.2.4, self-disclosures tend to be limited to revealing surface information about oneself. Also, the emotions expressed in *tweets* are often superficial and intense, with a clear direction towards positive or negative sentiment.

The developed algorithm consists of two techniques designed based on the above characteristics of the platforms. The first is a new unsupervised incremental machine learning technique that automatically captures discussions from a text stream of a fast-paced social media platform. This technique self-structures social text into coherent clusters which are indicative of distinct topics, and extend those clusters across time into pathways a.k.a *topic pathways* to capture changes of those discussions over time due to changes in trending topics and appearance new unseen topics.

The composition of a social text stream evolves continuously over time, which includes changes in the focus of the topics, the emergence of new topics and cease of certain topics. Considering a social text stream altogether makes it difficult to capture those interesting dynamics of the discussion topics. Therefore, it is important to separate the social text stream into dynamic topics that evolve over time i.e., topic pathways. This topic pathway separation reduces noise and separates social media messages in a meaningful way so that they are categorised into topic pathways based on their topical similarity. The proposed technique is designed to achieve this task by generating a self-adaptive structure automatically separate social text stream into topic pathways.

The second technique is a multi faceted event detection technique developed to monitor topic pathways for significant changes in social behaviours to detect significant changes over time and identify such changes as potential events of interest. Due to the above discussed characteristics changes in social behaviours are mainly on changes in self-disclosure as in changes of discussion topics and changes in emotional expression. Therefore, detection technique was developed to capture those changes.

These new techniques were designed to be capable to effectively handle the challenges related to the unstructured nature of social data outlined in Chapter 2.7. They are unsupervised and capable of incremental learning thus overcomes issues related to unlabelled data and time sensitivity.

This algorithm was trialled using two large datasets of tweets, one on a public figure and the other on an international organisation. The results highlighted the capability to automatically structure tweets semantically and temporally yielding coherent and meaningful topic pathways. Also, the automatically detected events from the detection algorithm were relevant and meaningful.

Figure 4.1 presents the high-level design of the proposed platform. A social text stream is denoted as a continuous stream of messages over time $< m_1, m_2, \ldots >$ where each message m consists of a *short text* and a *time-stamp* of its published time.

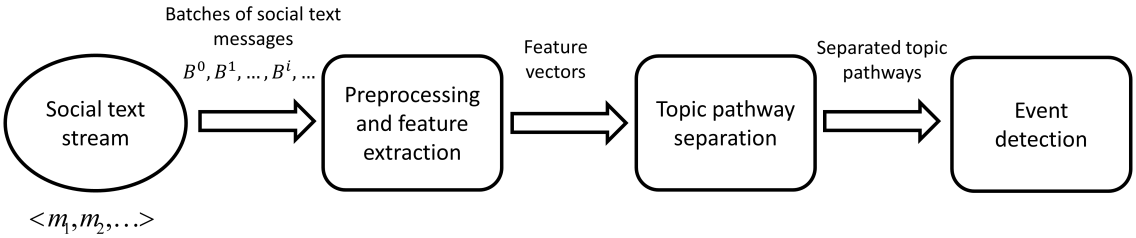


Figure 4.1: High-level design of the proposed platform for topic separation and event detection from a social text stream

Firstly, the social text stream is sampled using a fixed time-interval Δt based on the *time-stamp* of social media messages producing batches of social media messages $B^0, B^1, \ldots, B^i, \ldots$. Each batch is a set of social media messages collected for the time period Δt.

$$B^i = \{m\}_{t=i\times\Delta t}^{t=(i+1)\times\Delta t} \tag{4.1}$$

The granularity of the time period Δt (e.g., hourly, daily, weekly) can be adjusted to suit the velocity of the social text stream and also to match other application specific requirements.

Each batch is preprocessed to reduce noise due to unstructured nature of social text and feature vectors v are extracted from each social text message m. Note that technical details of preprocessing and feature extraction are discussed in Section 4.7 . The feature vectors of each batch are employed by the topic pathway separator to separate the social media messages into topics and topic pathways over time.

A *topic pathway* is a series of social media messages that discuss a distinct topic. It stretches over time across several batches of social media messages. A *topic segment* is a part of the topic pathway that belongs to a particular batch. Thus, each topic pathway is a chain of topic segments. A topic pathway D can be depicted as follows:

$$D = D^0 \rightarrow D^1 \rightarrow \ldots D^i \ldots \rightarrow D^n$$

where $D^i = \{m\}$ is a topic segment and $D^i \subset B^i$

Each topic segment is a set of social media messages that are semantically similar to each other. Hence, technically it can be considered as a clustering problem within each batch. A chain of such semantically similar clusters across batches forms a topic pathway. Therefore, this problem can be reduced to unsupervised incremental clustering algorithm that builds a set of semantically similar cluster chains across batches of social media messages.

As discussed in Section 2.3, and event can be broadly defined as an occurrence that can be bounded by space and time (Allan, Papka and Lavrenko, 1998). Events could happen anytime and anywhere which could impact a single person, a handful of individuals, or a large number of individuals (e.g., natural disaster, election). Since the interest here is to capture any event that has caught attention in online social discussions, any specifics of a particular event is not priorly known, thus *unspecified event detection* (see Section 2.3.1) technique was designed for this task. This unspecified event detection technique looks

for changes in online social behaviour of the impacted individuals for the duration of the event (e.g., change of discussion topics, changes of emotions). Such changes in human social behaviours related to an event appear as sudden bursts (Barabasi, 2005) in social text stream in between moderate or low level of activity. However, such bursts in a particular topic often diluted inside the diverse social text stream and can go unnoticed because the dynamics of other topics act as noise.

As a solution, we propose to detect events in the separated topic pathways because each topic pathway is a coherent stream of social media messages which continues a discussion about a certain topic. Hence, a burst of activity related to a particular topic would be easily detectable in relevant topic pathways as they are not (or minimally) affected by the noise of other topics. This proposed event detection algorithm monitors each topic pathway for topic specific events and detects such events using a multi-faceted event detector that looks for multiple event indicators.

The following sections describe the topic separation and event detection components of the proposed platform in detail.

4.2 Unsupervised incremental learning for topic separation

As discussed above, topic pathway separation is technically an unsupervised incremental learning task that builds a self-adaptive structure which consists a set of cluster chains over time across the batches of data.

As discussed in Section 2.6, unsupervised incremental learning is a challenging task. It has to learn previously unseen patterns from new incoming data while preserving the knowledge acquired previously (no catastrophic forgetting). Also, it should not require access to the past data that it has already processed.

Due to this challenging nature of the unsupervised incremental learning, there are only a handful of algorithms in literature that satisfies the above three characteristics (see Section 2.6 for details). Out of those, the recently published Incremental Knowledge Acquisition and Self-Learning (IKASL) (De Silva and Alahakoon, 2010; De Silva, 2010) is selected as the base for this work.

In IKASL, learning occurs as a structure of layers where each layer is learned from a new batch of data. In each layer, structure of the incoming data is self structured using the unsupervised self-learning mechanism introduced in GSOM algorithm (Alahakoon et al., 2000). GSOM employs the brain-inspired Hebbian learning rule (Kohonen, 1990) to learn the underlying structure from a dataset.

IKASL preserves the acquired knowledge in a generalised form therefore, it does not require access to past data that it has already processed. The generalised version of the acquired knowledge from each layer (n) is used as the basis for the knowledge acquisition from the subsequent layer (n+1), thus it avoids catastrophic forgetting of the past knowledge. Moreover, while using the past acquired knowledge as the base, it incrementally acquires new knowledge that is encapsulated in the upcoming data. The following section elaborates the functional details of the IKASL algorithm.

4.3 IKASL algorithm

This section describes the functional details of the IKASL algorithm. Table 4.1, states the symbolic notation that will be used in this section to detail the IKASL algorithm.

Table 4.1: Notations used for the functional details of IKASL algorithm

Symbol	Description
B^i	i^{th} batch of social media messages
d	number of dimensions in (i) feature vector of a social text message, and (ii) weight vector of a feature map node
LE^i	i^{th} learning phase
GE^i	i^{th} generalisation phase
n^j	j^{th} feature map node
TE^j	total quantisation error of node j
C^i	cluster representation vector learned during GE^i
$\mathcal{N}_{N_q}$	Set of N_q and neighbouring nodes of N_q

The IKASL algorithm functions in three key phases; initialisation, learning and generalisation which are delineated below.

4.3.1 Initialisation

IKASL is initialised in a similar manner to GSOM, using a map of four starting nodes $\{n\}^4$. Each node n is initialised with a weight vector w of size d. Initial values for the weight vector can be selected based on two different lines of approaches:

- **random initialisations:** this type of approaches include completely random initialisations, initialise using randomly selected data points and random sampling from a statistical distribution.
- **data distribution based initialisation:** these type of approaches include the use of K-means clustering to initialise weights to identified cluster centres and heuristic methods such as principal component analysis to assign optimal values to weights.

Each approach comes with pros and cons however, often such gains are marginals and only affect the rate of convergence. Interesting readers can refer to (Attik et al., 2005) for further details.

For the sake of simplicity, IKASL weight vectors are initialised by sampling from a uniform random distribution in the range of [0, 1].

4.3.2 Self structuring

This learning phase LE^i happens in each layer of IKASL in which it self learns a topology preserving two-dimensional feature map form the data batch B^i. This self structuring function employs the GSOM (Alahakoon et al., 2000) algorithm to learn the feature map. The GSOM algorithm is an enhanced version of Self Organising Map (SOM) (Kohonen, 1990) discussed in Section 2.5.

Similar to SOM, GSOM uses the concept of competitive learning, where a set of nodes (neurones) ins a feature map compete for the ownership of input data vectors. The winner or the *best matching unit* is the node that is most similar to the input.

Once the winner is selected, brain inspired Hebbian learning rule (Hebb, 1949) is used to update the weights of the winning node and its neighbourhood so that their distance to the particular input is reduced. This process iteratively learns the topology preserved feature map. One of the key limitations of SOM is that the number of nodes in the feature map need to be pre-determined, which is often done experimentally. GSOM overcomes this issue by initiating with a small number of nodes in the feature map and using a node expansion function to dynamically expand the feature map during the training phase. This self structuring phase executed in two stages (i) Self-Organisation with node expansion and (ii) Smoothing. The details are described below.

Self structuring and node expansion

This stage presents the inputs in a randomised order and learns the feature map using competitive learning and Hebbian learning rule (Hebb, 2005). Node map is dynamically expanded based on the quantisation error accumulated in individual nodes.

- A random input vector v is selected from the dataset and presented to the network of nodes in GSOM.
- The network of nodes compete for the ownership of the input v and the node n_q that is closest to the input vector v is selected as the winner:
  $sim(v, n_q) \geq sim(v, n)$, $\forall n \in \{n\}$.
  Note that similarity is measured using a designated distance function.
- The feature map updates the weight vectors of the winning node and its neighbouring nodes $\mathcal{N}(n_q, i)$ to reduce the distance between nodes and the input vector v The weight vector w_j of each selected node n_j is updated as follows:

$$w_j(k, i+1) = \begin{cases} w_j(k,i) + \alpha(i) \times (v(k) - w_j(k,i)), & \text{if } n_j \in \mathcal{N}(n_q, i) \\ w_j(k,i), & \text{otherwise} \end{cases} \tag{4.2}$$

  Note that $\alpha(i) \leq 1.0$ is the learning rate, i is the iteration and $k = 1, 2, \ldots, d$.
  The learning rate α is a tuning parameter which is determines the amount of weight correction in each selected node due to the input v. It starts with a higher values and gradually decreases as $i \to \infty$.
  $\mathcal{N}(n_q, i)$ is the neighbourhood of n_q that receives weight updates. This process simulates the neuro-biological phenomenon of *lateral excitation* which happens in the neocortex of brain where the activation of a neurone excites the neurones that are connected to it (Dalva and Katz, 1994). The size of the neighbourhood $\mathcal{N}(n_q, i)$ gradually decreases as $i \to \infty$ and eventually reduces to $\{n_q\}$.
- The winning node updates its quantisation error based on the distance between the node and the input v: $QE_{n_q} = QE_{n_q} + |v - w_q|$. This accumulation of quantisation

error is recorded for each node in the network. It is an indication of the level of generalisation of each node, where higher values indicate that the node is over-generalised and represents a large input feature space (Alahakoon et al., 2000).

- IKASL uses the node expansion method proposed in GSOM (Alahakoon et al., 2000) which uses the accumulated quantisation error QE_{n_q} to trigger expansion of the network from that node. When $QE_{n_q} > \tau_Q$, node expansions is triggered on n_q and new node are added to the vacant spaces of n_q in the grid formations as shown in Figure 4.2.

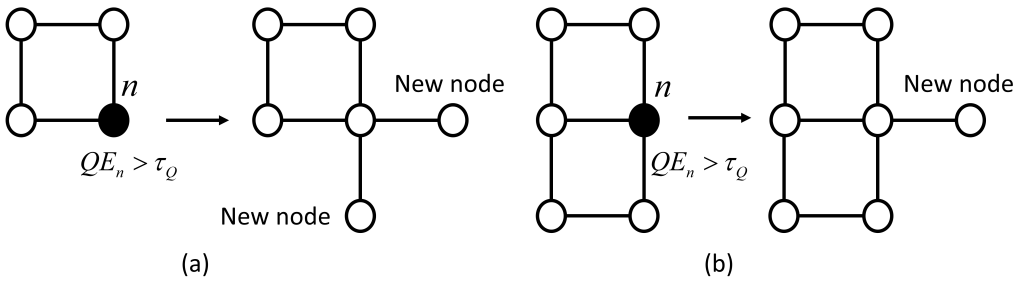


Figure 4.2: Two node expansions scenarios

Smoothing

This stage further fine tunes the weights of the node map by re-feeding the inputs. It uses the same competitive learning and Hebbian learning rule based approach to fine tune the node weights. However, in contrast to the previous stage, this stage uses a small learning rate α and a small neighbourhood in order to avoid substantial weight changes due to inputs. Moreover, this stage uses a fixed node map learned from the previous stage and do not perform node expansion.

4.3.3 Generalisation

This generalisation phase GE^i extracts a summarised representation as a set of cluster representation vectors $\{C\}$ from the topology preserved node map learned from the corresponding self structuring phase LE^i. Nodes that are able to claim a significant proportion of inputs are identified as *hit nodes*. The cluster representation vectors $\{C\}$ are learned by aggregating the weights of hit nodes and weights of its neighbourhood. This phase is described below:

- Input vectors are re-fed into the node network and let the nodes claim inputs based on similarity.
- Hit count $h(n)$ of each node n is the number of inputs claimed by each node.

$$h(n_j) = |\{\forall v \in B^i : n_j = arg \max_{n \in \{n\}} (sim(n, v))\}| \tag{4.3}$$

- Nodes that claimed more than τ_H proportion of input were marked as hit nodes, where hit threshold τ_H is a tunable parameter. Note that, hit threshold $\tau_H \in [0, 1]$

controls the granularity of the summarised representation where higher τ_H results few hit nodes, thus, more generalised representation.

- The *cluster representation vector* C_{h1} is formed by aggregating the weights of an identified hit node n_{h1} and its neighbourhood $\mathcal{N}_h(n_{h1})$.

$$C_{h1} = aggregation(n_{h1}, \mathcal{N}_h(n_{h1}))$$

Note that IKASL has employed a *fuzzy integral* as the aggregation function. However, any suitable aggregation function can be used for this task.

4.3.4 Learning from generalisation

After the first layer, learning phase LE^i is based on the set of cluster representation vectors $\{C\}^{i-1}$ learned during previous generalisation phase GE^{i-1}. Each *cluster representation vector* becomes the weights of the starting node of a separate feature map in LE^i. Therefore, $|\{C\}^{i-1}|$ number of feature maps are learned in LE^i.

Firstly, the feature vectors are assigned to their closest *cluster representation vector* based on an employed similarity function. Those assigned feature vectors are then used to learn the respective feature maps initialised from that *cluster representation vector.*

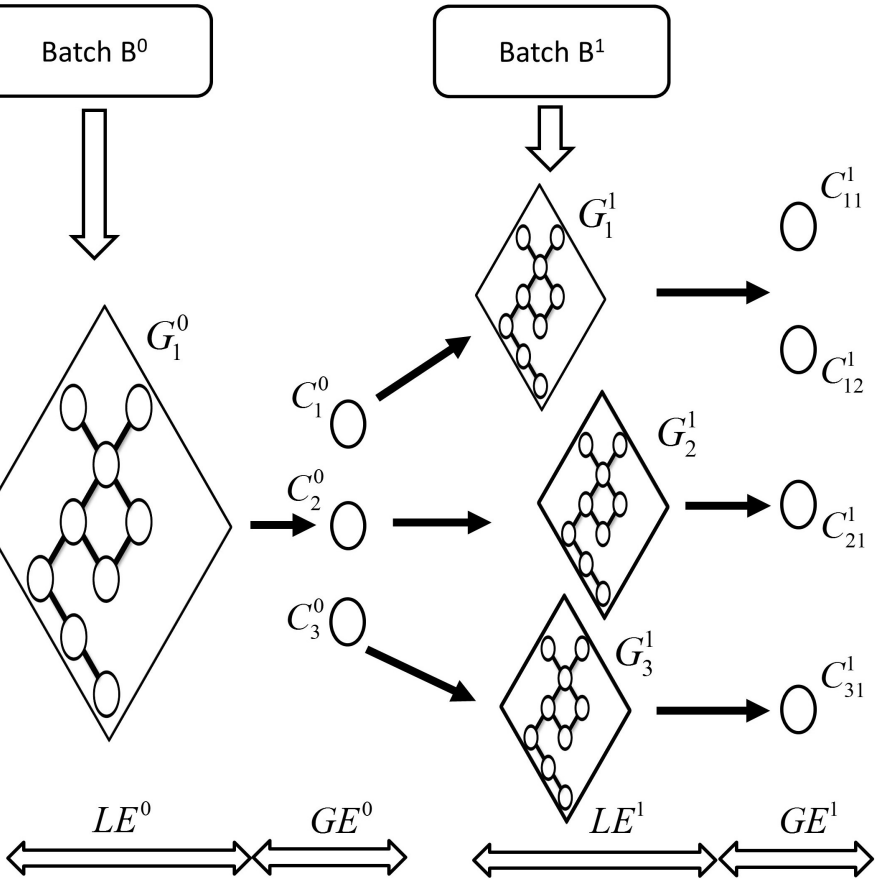


Figure 4.3: Layered structure of IKASL algorithm

Figure 4.3 highlights the self-learned layered structure of the IKASL algorithm. Learning phase LE^1 is initialised using the *cluster representation vectors* $\{C\}^0$ of layer 0. Hence, previously learned knowledge from B^0 is preserved and used as the basis for the acquisition of new knowledge. The new knowledge from batch B^1 of data is acquired by the learned feature maps in LE^1. Subsequently, generalisation layer GE^1 generates a summarised representation of the acquired knowledge as a set of *cluster representation vectors* $\{C\}^1$.

4.3.5 Topic separation with IKASL

The features vectors derived from batches of social media messages are used to self structure the adaptive structure that consists of chains of cluster representation vectors. Each batch of social media messages are assigned to the most similar *cluster representation vector* generated from the same batch, thus forming clusters of social media messages. Chains of *cluster representation vectors* form chains of social media messages clusters that become the topic pathways.

IKASL addresses the issue of time sensitivity (see Section 2.7) in social media messages by sampling them using their published time and incrementally learning topic pathways that capture the dynamic changes of each topic. The learning of IKASL has linear complexity with number of social media messages, hence it is scalable to handle large volumes of social text data. Moreover, the acquired knowledge is stored in a summarised form as *cluster representation vectors*, which makes it scalable in memory as well.

While IKASL addresses several issues related to topic pathway extraction from a social text stream, there are several identified issues that are not addressed. Hence, an extended IKASL algorithm (IKASL*text*) is proposed to address those identified issues. Those issues and the resolutions in IKASL*text* are presented in next section.

4.4 IKASL*text*: Extended IKASL for topic separation from a social text stream

This section describes three key extensions included in IKASL*text*, the extended IKASL algorithm for topic separation from social text stream, which is optimised to handle the challenges in social text.

Section 4.4.1 describes the developed technique to overcome the issue of dynamic vocabularies, Section 4.4.2 describes the technique the handle the brevity in social media messages and Section 4.4.3 presents the technique to detect new topic pathways and improve coherence in existing topic pathways.

4.4.1 Handling dynamic vocabularies

In online social media platforms users often coin new words that are not found in standard vocabulary (Eisenstein, 2013). As discussed in Section 2.7, some of these terms are tag words that are coined to represent certain events or topics (e.g., hashtags in Twitter). Such terms are important to aggregate the social media messages of that topic/event and

also to understand about the evolution of that topic/event. Often these new terms appear as bursts where its usage peaked rapidly, subsequently declines and seldom used after.

The conventional approach would be to use a static global vocabulary of significant terms in the dataset to generate feature vectors from social media messages. However, there are two key issues associated with this conventional approach.

1. The tag words that are used in social data streams are not known a priori (Mahendiran et al., 2014) thus it is not possible to create a global static vocabulary that includes all possible terms.

2. Consideration of large number of terms as features significantly increases the sparsity of the feature vectors.

In order to resolve these issues of a global vocabulary, a new method is developed to facilitate dynamic vocabularies which can incorporate new significant terms (e.g., hashtags) to the incremental learning task as they appear in new batches of data. A vocabulary V^i comprising a set of significant terms are learned from each batch B^i of social media messages. This vocabulary V^i is local to batch B^i and used for the self structuring stage of that batch. Therefore, the learnt cluster representation vectors $\{C\}^i$ from the batch B^i are based on features extracted using V^i.

The candidate terms for the vocabulary V^i is selected based on local term-frequency (Salton and Buckley, 1988b) i.e. frequency of the term in batch B^i, where the terms that have a local term-frequency higher than τ_V in B^i are included in V^i. τ_V is a tuning parameter in which setting a higher value can miss informative terms and setting a lower value can increase the sparsity of the feature representation of social media messages.

The key challenge of this approach is how to measure similarity between feature vectors based on two distinct vocabularies. This is because in IKASL, the cluster representation vectors $\{C\}^i$ learned from the batch B^i are employed as the basis for the self structuring stage for subsequent batch B^{i+1}, in which V^{i+1} might contain different terms that are not in V^i.

In order to solve this problem the intersection between the two vocabularies is considered in similarity calculations. For example, let C^i_m be any cluster representation vector from stage i and v^j be any input vector from data batch B^j (where $j > i$), the Cosine similarity between C^i and v^j is defined as follows:

$$sim(C^i, v^j) = \left[\frac{C^i \cdot v^j}{\|C^i\| \times \|v^j\|}\right]^{V^i \cap V^j} \tag{4.4}$$

where $V^i \cap V^j \neq \emptyset$. Note that Cosine similarity is a popular similarity measure for sparse feature vectors and often used in text *bag-of-word* representations.

Figure 4.4 shows the local *term frequency* plot of two terms in the social data stream. Term A maintains a local *term frequency* higher than τ_V throughout hence it is included in vocabulary of all batches shown. In contrast, term B appears as a burst in B^3, reached peak local *term frequency* in B^5 and gradually declined subsequently. Its local *term frequency* is greater than τ_V only in batches 3 to 7, hence included in vocabularies $V^3, \ldots, V^7$.

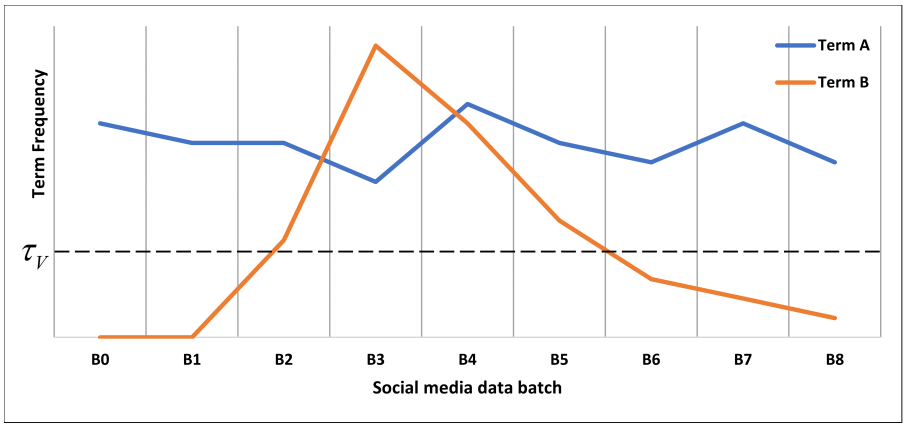


Figure 4.4: Local *term frequency* of terms A and B across social media message batches $B^0, \ldots, B^8$. τ_V is the local *term frequency* threshold that determines inclusion of exclusion in the respective vocabulary

This method overcomes the problem of new terms as it uses a dynamic vocabulary and incorporates new significant terms in the feature space as they appear in the new batches of social media messages. As in for *term B* shown in Figure 4.4, dynamic vocabularies include the new term as a feature when it is significant and relevant. Thereby, self structuring and generalisation stages on relevant stages consider the dynamics of the particular new term for topic pathway separation. Moreover, *term B* is automatically excluded from the feature space when its usage drops beyond the threshold τ_V thus less relevant. This subsequent exclusion reduces the sparsity of the feature vectors and thereby improves the clustering results.

4.4.2 Addressing the brevity issue

The short length of social media messages are imposed by some social media platforms (e.g., tweet is limited to 140 characters). Even in platforms without such restrictions, users prefer brevity since it allows them to express more efficiently. This results in a fewer number of words in social media messages. For example, tweets are on average 11 words long (O 'connor et al., 2010).

The feature vectors of the social media messages are formulated using the *bag of words* technique where each feature is a word in the text corpus of that batch of social media messages. The corresponding feature of each word present in the social media messages is assigned with the feature values of that word while other features are zero (not activated). Because of the brevity, these feature vectors of the social media messages have only a handful of (non-zero) features.

During the self structuring phase (Section 4.3.2) the activated features of a feature vector activate the corresponding weights of the winning node. Throughout the self structuring phase different sets of weights of the node get activated based on the feature vectors that the node claims as the winner. However, since the winner is claimed based on similarity which is determined using the activated features of the node, the already activated features often get further strengthened during the self structuring phase. These activated

weights of the node represent the acquired knowledge of that node. Because of the brevity issue, often few feature weights are activated in each node after the self structuring phase.

When the nodes have only a few activated weights and others are close to zero, the overlapping activated features of the hit node and its neighbourhood will be less. Hence, the existing summation based aggregation function used in the generalisation phase (see Section 4.3.3) would diminish the strength of the non-overlapping weights and the relevant acquired knowledge will be lost.

This issue is inherent in summation based aggregation methods which favours frequently activated features, while features activated in one or few considered nodes are discriminated (Murray and Perronnin, 2014). In order to minimise this issue, we adapted an aggregation method known as *max-pooling* which is popular in Bag of Visual Words (BoVW) based techniques in visual recognition tasks (Peng et al., 2015). In fact, (Boureau et al., 2010) have theoretically shown that *max-pooling* performs better than *average-pooling* (averaging based aggregation) on features with a low probability of activation.

Max-pooling takes the maximum feature value for each feature from hit node and its neighbourhood. In a d dimensional feature space, a *cluster representation vector* is calculated from the neighbourhood $\mathcal{N}(n_h)$ (neighbourhood includes the hit node) of a hit node n_h as follows:

$$\forall k \in 1, \ldots, d, C(k) = \max_{n \in \mathcal{N}(n_h)} (n(k)) \tag{4.5}$$

This approach does not average-out activated features that only exist in few nodes to the near-zero values in other nodes. Intuitively, in the context of *bag of words* feature extraction, *max-pooling* aggregates the significant words learned by each node considered and include all of them as significant words in the respective *cluster representation vector*.

4.4.3 Identifying new topic pathways

The impact of dynamic and time sensitive nature of social data streams extends beyond the granularity of new terms into a much broader concept of new topic pathways. New topic pathways emerge when there is a substantial public interest on a new discussion topic i.e., *social media buzz*. Such emerging topic pathways are often linked to real world events that have captured the attention of a significant portion of the general public. Hence, it is important to capture newly emerging topic pathways which are loosely related to existing pathways.

However, the original IKASL algorithm links clusters to previous layers, thus all new clusters are formed having a cluster representation vector learned from previous layers as the base. Hence, when a new topic emerges, the relevant social media messages would map into an already existing topic pathway and only seen as incremental changes to an existing topic pathway. Such an approach would hide the new topic as well as decrease the coherence of the existing topic pathways.

Moreover, text datasets often face the issue of feature sparsity and as discussed in Section 2.7 this sparsity more so in social data. Sparse feature spaces make nearest neighbour queries unstable (Beyer et al., 1999) and less meaningful (Aggarwal et al., 2001). Therefore, the process (described in Section 4.3.4) of assigning social data messages to most relevant topic pathway i.e., the closest *cluster representation vector* based on similarity would assign less similar feature vectors to respective topic pathways. Hence, the knowledge acquisition process from the new unseen messages can increase the variance within topic pathways decreasing the topic coherence of topic pathways over time.

In order to overcome these two issues, we employed a similarity threshold based method which is often used in high-dimensional clustering algorithms (Ester et al., 1996; Ertöz et al., 2003) to filter out *noise* feature vectors that are loosely related to any existing clusters. However, instead of filtering out such feature vector as *noise* our proposed technique identifies any potential new topics encapsulated in them. Therefore, in this proposed method, if the similarity with the closest *cluster representation vector* is greater than the similarity threshold τ_{topic}, then such features vectors are used to learn a new feature map G_{new} that is randomly initialised as in LE^0 as below. Thus, for a feature vector v, the feature map G is selected as bellow:

$$G = \begin{cases} G_j, & \text{if } C_j = arg \max\limits_{C \in \{C\}} (sim(C, v)) \text{ and } sim(C_j, v) \geq \tau_{topic} \\ G_{new}, & \text{otherwise} \end{cases} \tag{4.6}$$

τ_{topic} is the *topic pathway coherence threshold* that represents the expected coherence of the identified topics. Low τ_{topic} values result in few topic pathways with less coherent content in which some might be loosely related to the particular topic. τ_{topic} positively correlates with topic coherence in topic pathways and negatively correlates with number of topic pathways. High τ_{topic} values result in more topic pathways with increased coherence while low τ_{topic} values result in less topic pathways with low coherence.

This method improves the coherence of existing topic pathways as only the feature vectors that are similar to the respective cluster representation vector, are used in the self structuring stage to incrementally learn dynamic changes of that topic pathway. In contrast, the generation of a new feature map leads to the formulation of cluster representation vectors that are not influenced by the knowledge acquired in previous layers and thus, becomes the basis for new topic pathways.

It should be noted that the proposed technique separates topic pathways autonomously. The coherence of these topic pathways can be evaluated using topic coherence measures (Röder et al., 2015). The details of topic coherence are further discussed in the experiments section, alongside results from experiments.

This extended IKASL algorithm has the same time-complexity as the vanilla IKASL algorithm. It primarily depends on the number of inputs n and the number of nodes in the GSOM k. Each training step requires a nearest neighbour search between the input and GSOM nodes leading to a complexity of $\mathcal{O}(nk)$. However, the expansion of GSOM slows over training iterations and eventually saturates. Thus, when the input data set is sufficiently large ($n >> k$) it can be assumed that the time complexity of IKASL algorithm

is linear with the input size i.e., $\mathcal{O}(n)$. It is a very useful property of IKASL which enables its use on very large datasets.

4.5 Event detection

As mentioned in Section 4.1, an can be characterised as sudden burst of activity in a certain topic. In recent literature as discussed in Section 2.3.1, detecting such burst of activity in a topic is often simplified to detecting bursty key words or tag words (hashtags in Twitter) (Weng et al., 2011) with the hypothesis that such burst key words are indicative of an event related to the topic inferred by the meaning of those key words. In contrast, (Paltoglou, 2016) recently introduced a sentiment based event detection technique where significant changes of sentiment can be used as event indicators with comparable results.

As discussed in Section 4.1, it can be hypothesis that an *event* in social data stream creates a disruption to related topic pathways so that the characteristics of social behaviours in the affected topic pathways change substantially. Considering the nature of characteristics in fast-paced social media platforms Such changes are reflected in various aspects of the topic pathway such as change of volume or change of sentiment.

Hence, in the proposed event detection approach we employed a multi-faceted event detector. It looks for different event indications such as volume changes (I_v), positive sentiment changes (I_{PS}) and negative sentiment changes (I_{NS}). In addition to those generic event indicators, domain specific event indicators (I_{DS}) such as (i) volume of disaster related words e.g., shaking (Sakaki et al., 2010b), and (ii) volume of civil unrest related words e.g., protest (Ramakrishnan et al., 2014); can also be incorporated in to the event detection module.

This proposed technique linearly combines the event indicators to obtain the final event score $\mathcal{I}$. Let D_j^i be a topic segment of topic pathway D_j, the final event score of D_j^i is defined as follows:

$$\mathcal{I}(D_j^i) = r_V \times I_V(D_j^i) + r_{PS} \times I_{PS}(D_j^i) + r_{NS} \times I_{NS}(D_j^i) + \ldots + r_{DSE} \times I_{DSE}(D_j^i) \quad (4.7)$$

where $r_V, r_{PS}, r_{NS}, \ldots, r_{DSE} \in [0, 1]$ and $\sum r = 1$
r_k represents the sensitivity of the event indicator I_k to the final event score, where high or low value of r_k adapts the impact of I_k to the final event score accordingly. r values can be empirically set based on the importance of certain indicators to suit different applications. It is also possible to learn r values by training a classifier using pre-labelled records.

$\mathcal{I}$ can take values between 0 to ∞, and its value is proportionate to the significance of the event. The potential events of interest can be identified by applying an event threshold τ_e where $\mathcal{I}(D_j^i) \geq \tau_e$ are flagged as events.

The following subsections describes the volume based and sentiment based event indicators respectively.

4.5.1 Volume based event indicator

The volume based event indicator looks for significant increase in volume of a topic pathway. Such increase is an indication of increased interest in that topic pathway, which can be due to a potential event relevant to that topic pathway.

In order to capture significant increases in volume, we set I_V as the ratio between the proportion of messages in a topic segment and the moving average of proportion of messages in that topic pathway as:

$$I_V(D_j^i) = \frac{V_P(D_j^i) \times W}{\sum_{x=i-w-1}^{i-1} V_P(D_j^x)} \tag{4.8}$$

where W is the window size for the moving average.
The volume proportion of a topic segment $V_P(D_j^i)$ is determined as:

$$V_P(D_j^i) = \frac{|D_j^i|}{|B^i|} \tag{4.9}$$

The volume proportion is more robust than the absolute volume (number of social media messages) of a topic segment. This is because absolute volume of a topic segment is biased to the volume of social text message batch which varies significantly over time despite fixed sampling time Δt. These changes in volume of social text message batch are often due to seasonal changes which depends on hour of day (daytime vs night-time) and day of week (weekdays vs weekends). In contrast, volume proportion is more resilient to bias as it normalise the seasonal variations.

4.5.2 Sentiment based event indicator

Sentiment based event indicators capture the changes of public opinion (Thelwall et al., 2011). The concept that events are associated with changes of positive and negative sentiment strength has been successfully used for a recent event detection (Paltoglou, 2016) task from a social text stream. The hypothesis here is that an event can alter the existing level of public opinion of relevant topic pathways and such changes can be detected by measuring the sentiment changes of topic pathways over time.

As discussed in Section 2.4, sentiment analysis on social media messages is a challenging task mainly because of its inherent unstructured nature (see Section 2.7) where authors loosely follow grammatical rules of language (Baldwin et al., 2013) and heavily use emoticons to express sentiment. Moreover, word lengthening (e.g., *sorrrry*) is widely used to emphasis the sentiment of statements (Brody and Diakopoulos, 2011). Recent research reports that such variations have resulted in significant performance degradations in conventional sentiment analysis tools (Eisenstein, 2013).

Therefore, it is important to use sentiment tools that are designed to facilitate the above mentioned factors of social text. In Section 2.4.3, we have investigated the techniques of several state-of-the art sentiment analysis tools (Esuli and Sebastiani, 2006; Socher et al., 2013; Thelwall et al., 2012), and found that SentiStrength (Thelwall et al., 2012) is well

suited for our requirement; mainly because it has been specifically engineered to capture sentiment related features in social text. It employs a word list with sentiment strength and polarity to derive both positive and negative sentiment strengths. It considers emoticons, boosting words and negation words to strengthen or weaken the sentiment value. In addition, it considers repeated letters in a word (e.g., *sorrrry*) as an indication of adding more emphasis to the sentiment of that word.

SentiStrength provides positive sentiment values ranging from 1 to 4 and negative sentiment values ranging from -1 to -4 respectively. Two sentiment based event indicators are designed utilising both positive and negative sentiment values to capture sentiment variations in both positive and negative sentiment. The sentiment value of a topic segment is determined by aggregating sentiment values of social media messages belong to that topic segment and taking the average. The variations of sentiment in a topic segment D_j^i is captured by comparing the sentiment to the moving average of sentiment in that topic pathway D^j. The positive sentiment event indicator I_{PS} and the negative sentiment event indicator I_{NS} is defined as follows:

$$I_{PS}(D_j^i) = \frac{avgPositiveSentiment(D_j^i) \times W}{\sum_{x=i-w-1}^{i-1} avgPositiveSentiment(D_j^x)} \tag{4.10}$$

$$I_{NS}(D_j^i) = \frac{avgNegativeSentiment(D_j^i) \times W}{\sum_{x=i-w-1}^{i-1} avgNegativeSentiment(D_j^x)} \tag{4.11}$$

where W is the window size for the moving average.

4.6 Illustrative Example

This section demonstrates the functionality of the proposed platform for separating topic pathways and event detection using an illustrative example shown in Figure 4.5.

As shown in Figure 4.5 social media messages are sampled into batches and extracted feature vectors are fed to the topic separation technique. Once the cluster representation vectors $\{C\}^i$ are learned, social media messages in B^0 are mapped to their closest cluster representation vector $C_j^i \in \{C\}^i$. The topic segment D_j^i is formed by the set of social media messages that are mapped to the cluster representation vector C_j^i. As shown in Figure 4.5, For example, C_j^0 is the closest for $m_3^0, m_4^0 \in B^0$, thus $D_j^0 = \{m_3, m_4\}$.

C_j^1 is learned based on C_j^0 (the feature map that used to derive C_j^1 is initialised using C_j^0), and is timely evolved acquiring relevant new knowledge from B^1. Hence C_j^0 shares similarity with C_j^1, C_j^1 shares similarity with C_j^2, and C_j^2 shares similarity with C_j^3. Because of these relationships among the cluster representation vectors, the topic segments formed based on these also contain semantically similar social media messages. Therefore, such topic segments are linked to form a topic pathway $D_j = D_j^0 \rightarrow D_j^1 \rightarrow D_j^2 \rightarrow D_j^3$ that contains semantically similar social media messages over time.

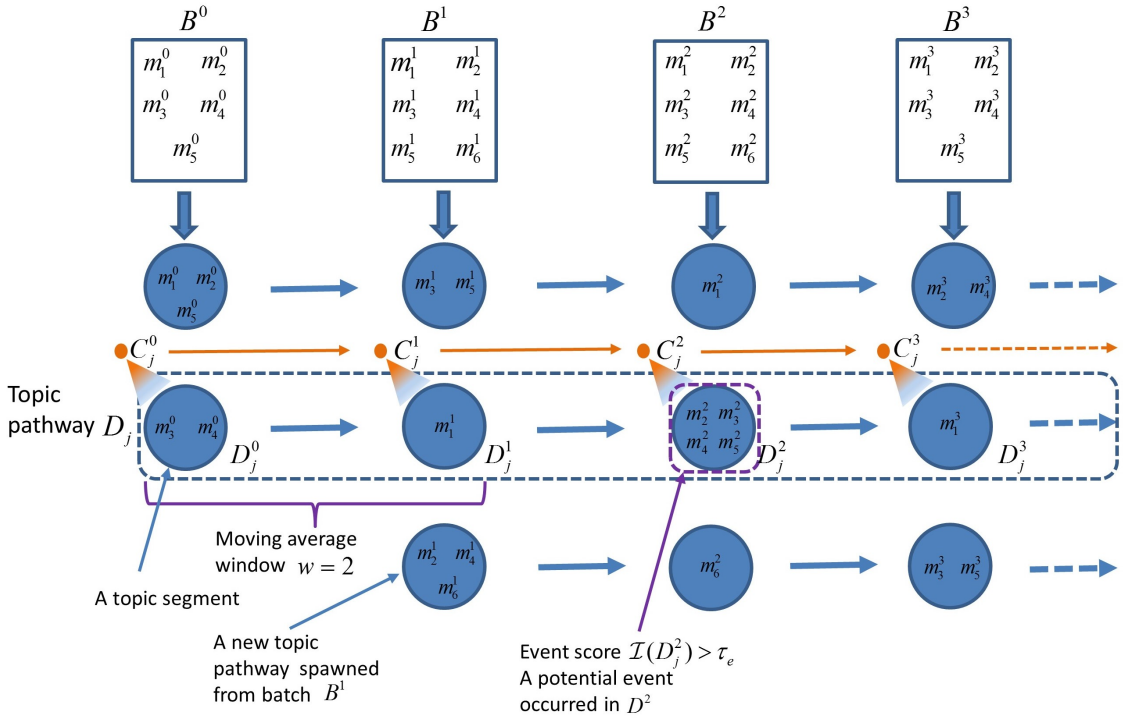


Figure 4.5: An illustrative example showing topic pathway separation and event detection

Moreover, the proposed IKASL*text* facilitates new topic pathways hence as shown in Figure 4.5 such new topic pathways can be spawned from any batch.

Those identified topic pathways are continuously evaluated by the proposed event detection algorithm. It collects information about event indicators and aggregates them into an event score $\mathcal{I}$. As shown, event indicators of D_j^2 are determined based on changes in volume proportion, positive sentiment and negative sentiment compared to the previous topic segments D_j^1 and D_j^0 (moving average window of 2). If $\mathcal{I}(D_j^2) > \tau_e$ then D_j^2 contains information about a potential event that is related to the topic pathway D_j.

The next section presents the results of an empirical evaluation of the proposed topic separation and event detection techniques using two Twitter datasets.

4.7 Experiments

This section demonstrates the core capabilities of the above proposed platform for topic separation and event detection. Twitter was used as the experimental online social media platform due to its popularity and rich API. The first subsection describes the datasets and new two subsections describe preprocessing and feature extraction steps. Subsequently, comprehensive accounts of the core capabilities, topic pathway separation, evolution of topic pathways, emergence of new topic pathways and automatic event detection are presented in separate subsections.

4.7.1 Datasets

This section describes the two Twitter datasets collected for the experiments.

#Obama Dataset: President Obama maintains a diverse range of public relations, both locally and internationally and he is the most followed political figure on Twitter with approximately 77.8 million followers. Thereby, analysing tweets about him is a challenging task. Separating these into topic pathways would showcase the capabilities of the proposed technique and its value over existing techniques. #Obama Dataset was collected using the keywords *#obama, @obama, obama, #potus* for the time duration 01/12/2014 to 19/04/2015 (20 weeks). 4,230,985 tweets were collected which contain 173,903 out of which 57,998 are *hashtags*.

#Microsoft Dataset: Microsoft Corporation has a diverse online social presence due to a large portfolio of technology-centric products and services that are consumed by a variety of end-users. This diversity is a fitting testbed to further investigate the separation of a tweet stream into different topic pathways on products, services and competitors. #Microsoft Dataset was collected using the keywords *#microsoft, @microsoft, microsoft* for the time duration 08/12/2014 to 28/06/2015 (30 weeks). 1,953,243 tweets were collected which contain 84,758 out of which 29,655 are *hashtags*.

The characteristics of the two datasets are summarised in Table 4.2. #Obama Dataset has significantly more unique tokens compared to #Microsoft Dataset. Interestingly a third of unique tokens are hashtags in both datasets (33% in #Obama Dataset and 35% in #Microsoft Dataset) indicating its widespread use in tweets. Out of these unique tokens only around 5% has a frequency of 5 or more in both datasets.

Table 4.2: Characteristics of the twitter datasets.

	#Obama Dataset	#Microsoft Dataset
Number of tweets	4,230,985	1,953,243
Date range	01/12/2014 to 19/04/2015	08/12/2014 to 28/06/2015
Terms used as filters	#obama, @obama, obama, #potus	#microsoft, @microsoft, microsoft
Number of unique tokens	173,903	84,758
Number of unique hashtags	57,998	29,655
Tokens with term frequency ¿5	8,112	4,151

4.7.2 Pre-processing

Pre-processing is carried out on text data to reduce noise. This step is important as it helps to reduce the sparseness of text data. Note that, informal language is widely used in tweets, therefore they are often noisier than standard text documents, and hence a series of pre-processing steps are necessary.

Duplicate removal: tweets that are duplicated within the same batch are often spam or advertisements generated by automatic bots (Wang, 2010). Duplicated tweets affect the topic pathway separation process because they tend to form separate pathways. Such pathways are not informative for understanding public opinion as they are published for advertising purposes and can confuse attempts at event understanding and detection.

Stopword removal: stopwords are often less informative words exists in text (e.g., 'a', 'the'). It is a common text processing practice to remove those stopwords before extracting features. In addition to those, tweets also have Twitter specific stopwords such as 'rt' (a tag word used for retweets). Web links in tweets were also removed, as they do not directly relate to the social expression. In addition, a set of domain specific stopwords were identified and incorporated into the standard stopword list. For example, in dataset Obama, words such as 'obama', 'barack obama', 'president obama' and 'POTUS' were filtered out.

Twitter specific tag removal: Tweets also contain user mentions i.e. '@username' and hashtags '#keyword' which are two unique features of tweets used to mentions a specific user and tag a keyword respectively. We removed user mentions from the tweet text as they are often user names. As discussed in Section 3.2.4, hashtags are unique keywords coined by public for social tagging or folksonomy. Such tags are often represents the content of the tweet and therefore indicative of the the topic. Therefore, hashtags were preserved as they supports the separation of topic pathways.

4.7.3 Feature extraction

In this step, representative features were extracted from the pre-processed tweets. We employed the *bag of words* model, in which terms (word or phrase) are considered as features of the tweets and features scores are determined based on the frequency of that word. We use noun phrases from tweets as the features, which are extracted using JATE toolkit (Zhang et al., 2008). Noun phrases that exist in less than a 0.5% of tweets in each batch (vocabulary threshold τ_V discussed in Section 4.4.1) were omitted to reduce the sparsity of the feature space.

We employed inverse document frequency (idf) as the representative feature value of a term. tdf_w of term w is given as follows:

$$idf_w = \log(\frac{N}{1 + df_w})$$

Where N is the number of tweets in that batch and df_w (document frequency) is the number of times the term w appeared in tweets of that batch.

4.7.4 Topic separation

This subsection demonstrates the first core capability; topic pathway separation based incremental learning outcomes.

Pathway separation occurs as the algorithm incrementally learns from the Twitter stream. Some pathways remain prominent, continuing to receive a significant number of tweets across each topic segment, while others lose popularity and number of tweets decreased over time.

#Obama Dataset

Six prominent (high volume) topic pathways were identified in #Obama Dataset (labelled TP^x_{Obama}, where x is 1-6). Each topic pathway consists of a group of frequent terms that signify the topic of that pathway. These words are more frequent among the tweets of that pathway compared to the tweets of other pathways.

Table 4.3 presents the top 10 frequent terms of these pathways. Column 3 of this table summarises the frequent terms into a focus area.

Table 4.3: Frequent terms in the prominent topic pathways of #Obama dataset

Topic pathway	Frequent terms	Key focus of the topic pathway
TP^1_{Obama}	putin, russia, ukraine, economy, iran, respect, west, merkel, war, head	Relations with Russia and Ukraine
TP^2_{Obama}	republican, party, question, tcot, congress, world, politics, cromnibu, gop, american people	Internal Politics relevant to Republican party
TP^3_{Obama}	iran, israel, sanction, nuke, netanyahu, congress, tcot, irandeal, war, nuclear weapon, nuclear deal	Relations with Iran and Israel
TP^4_{Obama}	nation, liberal, racism, reason, voter, democrat, consequence, tcot, gruber, pelosi	Internal Politics relevant to Democratic party
TP^5_{Obama}	obama news, cbs news, obama video, obamacare, sore throat, washington dc, gop, castro, cnn obama, summit	News about Obama
TP^6_{Obama}	war, afghanistan, isis, power, congress, protection siyra, iranian dissident, iraq	Terrorism, mainly ISIS

It should be noted that some of the pathways focus on internal political issues (TP^2_{Obama}, TP^4_{Obama}) while others are on international political issues (TP^1_{Obama}, TP^3_{Obama}, TP^6_{Obama}). TP^5_{Obama} is discussions about new reports on Obama.

In order to distinguish the salient terms that define a pathway, it is necessary to investigate their presence and prominence in other topic pathways. Table 4.4 tabulates this investigation. It presents the distribution of top five frequent terms of each prominent topic pathway, across the six prominent topic pathways. The rows represent the top five frequent terms of each prominent topic pathway and the columns represent the six topic pathways.

This table highlights that the top five terms of the six prominent pathways have a significant presence in the relevant pathway compared to other pathways. For example, in pathway TP^1_{Obama}, the top three words putin, russia, and ukraine have more than 80% within that pathway. The term iran has a low percentage in TP^1_{Obama} because it has a larger presence in TP^3_{Obama}.

Figure 4.6 shows the tweet volume of these topic pathways. It shows that different topic pathways peaked at different time periods. For example, in the latter part, the topic

Table 4.4: Distribution (%) of top five frequent terms of six prominent topic pathways in #Obama dataset

Topic pathway	Frequent terms	Distribution (%) of terms					
		TP^1_{Obama}	TP^2_{Obama}	TP^3_{Obama}	TP^4_{Obama}	TP^5_{Obama}	TP^6_{Obama}
TP^1_{Obama}	putin	97.9	0.3	0.5	0.0	0.3	0.0
	russia	85.2	0.7	9.3	0.2	0.2	1.7
	ukraine	84.6	1.6	1.3	0.3	3.1	3.5
	economy	54.7	3.4	1.5	1.5	5.9	0.5
	iran	2.2	2.0	69.8	0.4	1.7	3.8
TP^2_{Obama}	republican	1.0	81.4	1.5	2.0	2.8	0.5
	party	1.7	68.5	1.1	1.7	1.7	4.4
	question	3.1	67.3	4.3	0.0	0.6	7.4
	tcot	4.1	16.8	14.3	7.6	2.3	9.2
	congress	1.5	11.1	14.7	1.2	9.0	9.4
TP^3_{Obama}	iran	2.2	2.0	69.8	0.4	1.7	3.8
	israel	1.8	3.6	53.9	1.8	1.7	2.7
	sanction	5.8	2.6	57.6	0.6	1.2	1.2
	nuke	2.8	1.2	70.6	0.0	0.4	0.4
	netanyahu	1.6	3.2	16.9	0.7	3.0	1.9
TP^4_{Obama}	nation	0.3	1.6	0.6	90	0.9	0.9
	liberal	0.0	0.0	0.8	98.4	0.0	0.4
	racism	0.0	0.5	0.0	97.5	0.0	0.0
	reason	0.6	2.2	2.8	75	0.0	3.3
	voter	0.0	0.9	0.0	96.4	0.0	0.0
TP^5_{Obama}	obama news	0.5	0.2	0.6	1	90.4	0.4
	cbs news	0.0	0.0	0.8	0.0	90.5	0.8
	obama video	2.7	6.7	0.9	1.3	49.3	1.8
	obamacare	2.5	5.7	1.3	3.2	69	1.3
	washington dc	0.0	0.0	0.9	11.4	86	0.9
TP^6_{Obama}	isis	1.4	1.7	1.6	0.8	2	81.4
	war	7.6	2.8	10.7	2.2	1.6	56.4
	iraq	2.3	1.9	9.1	1.1	0.4	74.9
	siyra	7	19.9	21.6	0.0	0.6	37.4
	power	5.8	7.7	7.1	8.3	1.3	43.6

pathway TP^3_{Obama} becomes popular to reflect the incidents related to the nuclear deal between USA and Iran.

#Microsoft Dataset

Seven prominent topic pathways were identified in #Microsoft Dataset (Table 4.5), which were labelled using a similar approach to the above dataset.

Out of these seven, four topic pathways focused on products and services offered by Microsoft Corporation such as Windows, Windows Phone, Microsoft Office and Xbox. Remaining three focused on relations with competitors such as with Google, Facebook and Apple. Figure 4.7 shows the tweet volume of these topic pathways.

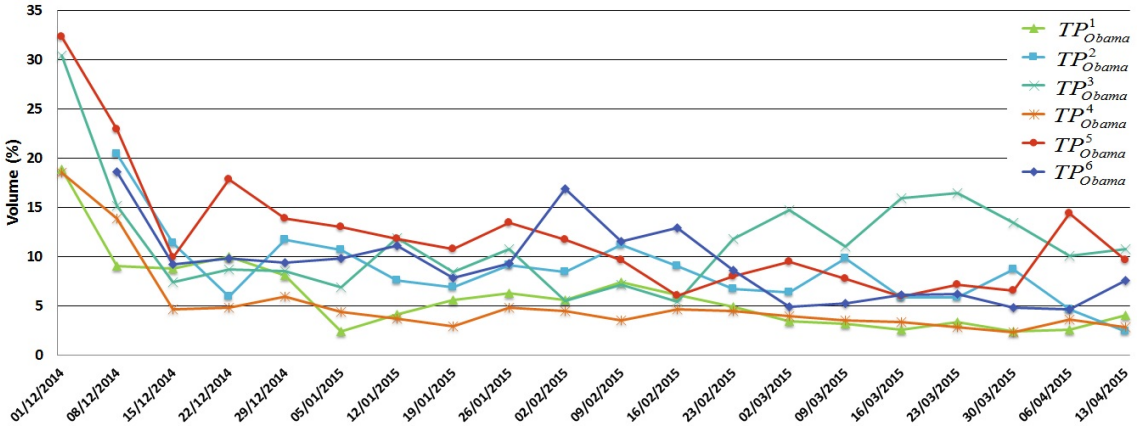


Figure 4.6: Volume (%) of tweets in six prominent topic pathways of #Obama Dataset

Table 4.5: Frequent terms in the prominent topic pathways of #Microsoft dataset

Topic pathway	Frequent terms	Key focus of the topic pathway
$TP^1_{Microsoft}$	Google, android, amazon, apple, cyanogen, ibm, work, sony, app, window	Relations with Google
$TP^2_{Microsoft}$	Xbox, ps4, update, windows 10, xboxone, game, Cortana, sony, bitcoin, microsoft xbox	Microsoft Xbox related products and services
$TP^3_{Microsoft}$	microsoft office, ipad, android, mac, android tablet gmail, windowsphone, google, onenote, microsoft tech, office	Microsoft Office and its compatibility in different devices
$TP^4_{Microsoft}$	Windows, version, windows 10, os news, future, microsoft windows, upgrade, microsoft windows10 Week, update	Microsoft Windows
$TP^5_{Microsoft}$	Facebook, bing, google, microsoft bing, zdnet Apple, favour, monumental deal, search result, virtual reality	Relations with Facebook
$TP^6_{Microsoft}$	windows phone, app, office, android, bitcoin, onedrive, update, bitcoin payment, tablet, ipad	Windows Phone
$TP^7_{Microsoft}$	Apple, google, cloud, amazon, Samsung, microsoft store, bigdata, fight, device, patent	Relations with Apple

The topic pathway that focuses on Windows has the highest volume in most of the weeks as Windows operating system ($TP^4_{Microsoft}$) is the most used product of Microsoft. Similarly, the topic pathway focuses on relations with Google ($TP^1_{Microsoft}$) has the highest number of tweets among the pathways that focus on relations with competitors since both Microsoft and Google are direct competitors of several products including operations system, mobile device, search engine and virtual reality technology.

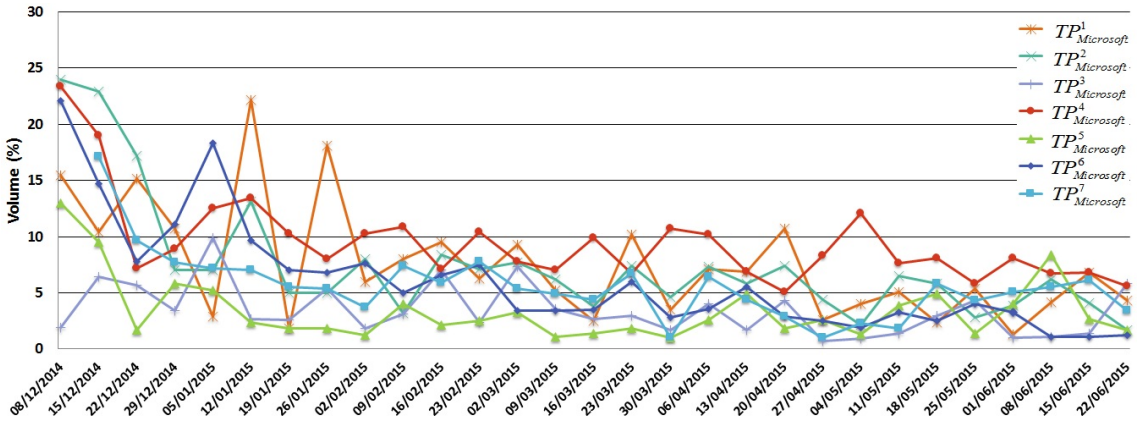


Figure 4.7: Volume (%) of tweets in six prominent topic pathways of #Microsoft Dataset

Topic Coherence

It is important to evaluate the quality of the automatically identified and learned topics based on their interpretability to a human analyst. There have been recent developments in quantitative metrics to evaluate the coherence of learned topics (Röder et al., 2015).

Mimno et al. (2011) have demonstrated such a metric based on the principle that significant terms belonging to a topic are likely to co-occur in same documents. This topic coherence metric delineates as follows:

$$C(T; V^T) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(v_m^T, v_l^T) + 1}{D(v_l^T)}$$

where $V^T = (v_1^T, \ldots, v_M^T)$ are the list of M terms that are most probable (frequent) in topic T. $D(v_l^T)$ is the document frequency of the term v_l^T. $D(v_m^T, v_l^T)$ is the document frequency of the co-occurrence of terms v_l^T and v_m^T.

This metric aggregates the term co-occurrence scores of a topic T for the M frequent terms of that topic. They have empirically verified that this metric correlates with the judgement of human analysts.

We have adopted this metric to measure and evaluate topic coherence within individual pathways. Our experiments measure the topic coherence scores of each topic for $M \in [2, 100]$ and the results are presented in Figure 4.8 for the two datasets; (a) #Obama and (b) #Microsoft. Note that, topic coherence score calculated for the entire dataset (without pathways separation) is plotted as the baseline.

The above results highlight that topic pathways have significant improvements in topic coherence compared to that of the entire datasets emphasising the semantic coherence of topic pathways learned by the proposed algorithm.

4.7.5 Evolution of topic pathways

This subsection demonstrates the second core capability of the proposed algorithm; capturing evolution of topic pathways and temporal changes in topic segments within a topic pathway. In order to demonstrate this capability, we selected the topic pathway,

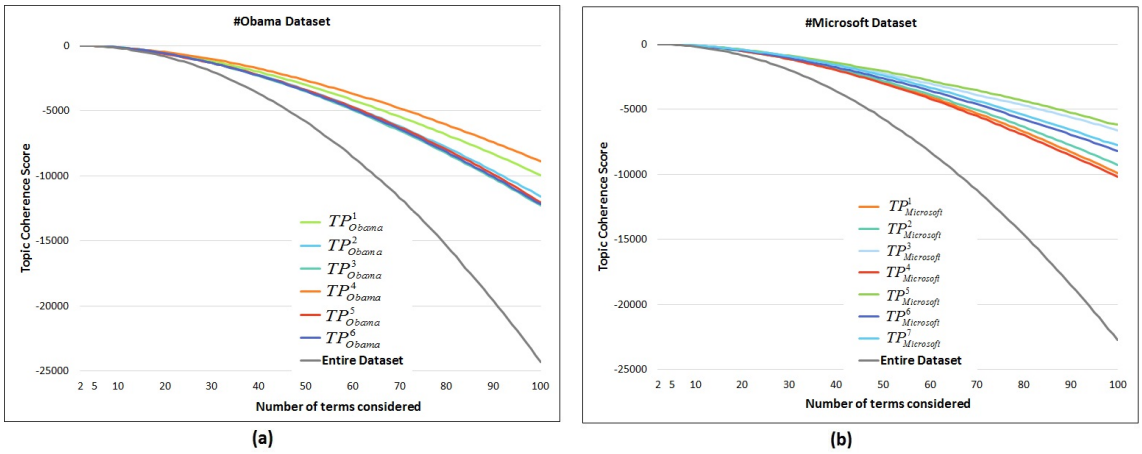


Figure 4.8: Topic coherence scores for prominent topic pathways and whole dataset (for baseline) in (a) #Obama Dataset and (b) #Microsoft Dataset.

$TP^1_{Microsoft}$, which captures the relationship between Microsoft and Google. It is common knowledge that both these organisations are dynamic and innovative in the technology domain, thereby twitter activity associated with both entities is equally dynamic with new subtopics (i.e., short-term topics within the particular topic pathway) being discussed. We demonstrate that the selected topic pathway $TP^1_{Microsoft}$, is capable of identifying each new subtopic as it emerges without a loss of focus on the overarching theme of Microsoft-Google relationship. Thereby, the entire topic pathway maintains focus on Microsoft-Google relationship, while each segment in the pathway focuses on different subtopics associated with this relationship as they emerge over time. The proposed algorithms autonomous capability to capture the evolution of a topic pathway in this manner provides an analyst unique insights into short-term changes within a long-term trend. In this instance, the long-term trend is the relationship with Google and the short-term changes are the subtopics associated with both entities.

We selected nine topic segments from pathway $TP^1_{Microsoft}$ for this demonstration. Figure 4.9 presents word clouds for these nine topic segments. Table 4.6 further examines content of each of the nine topic segments. Word clouds were selected for the intuitive value in visualising fluctuations in the usage and frequency of terms. The primary observation here is that each word cloud has a unique set of frequent terms which are indicative of different subtopics.

In Table 4.6, each row represents a topic segment. The first column is a label to identify each segment (i.e. first row in Table 4.6 *Segment 08/12/2014* corresponds to the first word cloud in Figure 4.9, also labelled *Segment 08/12/2014*). The second column denotes representative frequent terms for the corresponding segment and the third column denotes the subtopic for each segment. It should be highlighted that side by side, second and third column specify the relationship between top frequent terms and the subtopic. Although the subtopics change per segment, the main theme (or focus) of the pathway remains the same. The fourth column presents a robust evaluation of the subtopic of each segment, in the form of title of corresponding news articles published in mainstream online social media platforms during the same period of time.

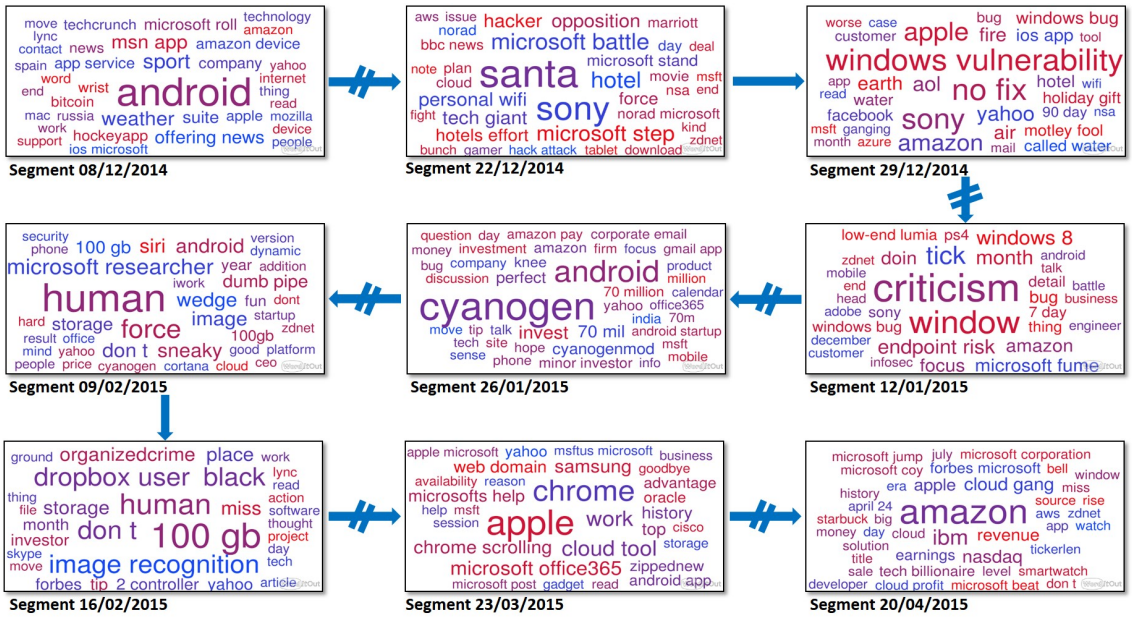


Figure 4.9: Word clouds generated for nine topic segments in topic pathway $TP^1_{Microsoft}$, which represents Microsoft-Google relationship. Note that most frequent terms *Microsoft* and *Google* were removed from word clouds for improved clarity.

For example, in *Segment 08/12/2014*, the discussion is about Microsoft providing MSN apps for android and iOS platforms. The second column presents representative frequent terms from this segment, the third column presents the subtopic and the fourth column validates both terms and the subtopic with the title of the corresponding news article. Some representative frequent terms are common across consecutive segments which indicates that some subtopics are discussed over several weeks. For example, frequent terms in week 22/12/2014 such as 'hotel' and 'wifi' appear in the word cloud of week 29/12/2014, but with a relatively lower frequency. Similarly, 'human' and '100 gb' appears in consecutive segments *Segment 09/12/2015* and *Segment 16/02/2015*.

Representative frequent terms within a segment are related to each other based on the subtopic. Representative frequent terms across segments are related to each other based on the topic pathway evolution. As noted, this topic pathway focuses on Microsoft-Google relationship, and each segment represents a different subtopic (or phase) of this relationship. Top terms across segments are able to capture this evolving topic pathway, despite differences in subtopics in each segment.

Figure 4.10 presents a contrasting view. It illustrates outcomes from conventional analysis of a social media platform, which analyse all tweets corresponding to the topic pathway as a single dataset. It only provides a concise view of what is discussed. Unlike Figure 4.9 and Table 4.6, it is impossible to capture evolution of a topic pathway or short-term changes in topic segments. The most frequent term (apart from 'Google') is 'android' as it is one of the main competitive product for windows operating system. Other frequent terms include 'amazon', 'ibm' and 'apple' which are competitors in the same space as Microsoft and Google. Insights gained by an analyst will be severely limited if Twitter activity is explored in this manner.

Table 4.6: Further examination of topic segments illustrated as word clouds in Figure 4.9

Segment	Representative frequent terms	Subtopic(s) of the segment	Relevant News Article(s)
Segment 08/12/ 2014	android, ios, msn app, sport, weather, offering news	Microsoft rolls out a set of MSN applications such as Weather, News and Sports for iOS and Android devices	MSN consumer apps arrive on Android, iOS, Amazon-CNET (11/12/2014)
Segment 22/12/ 2014	(i) microsoft battle, personal wifi, marriott, hotel (ii) santa	(i)Google and Microsoft step in to oppose the attempts of Marriott Hotels to block personal Wi-Fi inside the hotel (ii) Google and Microsoft provides interactive features to track Santas journey during Christmas	(i)Microsoft, Google join opposition to hotels' wi-fi blocking-ZNET(23/12/2014) (ii) Tracking Santa with help from Microsoft, Google-CNET (22/12/2014)
Segment 29/12/ 2014	windows vulnerability, no fix, bug, windows bug	Google has openly published a Windows 8.1 vulnerability that allows users to get administrator privileges	Google discloses unpatched Windows vulnerability-ZDNET (31/12/2014)
Segment 12/01/ 2015	criticism, windows, windows bug, microsoft fume, endpoint risk	Microsoft criticised Google for releasing information about security vulnerabilities to public	Microsoft slams Google for spilling the beans on Windows 8.1 security flaw-ZDNET(12/01/2015)
Segment 26/01/ 2015	android, 70 mil, cyanogen, android startup, 70 million, invest	Microsoft is taking part in a $70 million investment round in a startup that is developing a competitive operating system to Android	Microsoft to Invest in Cyanogen-Mod: What Could It Mean For Google?-Tech Times (29/01/2015)
Segment 09/02/ 2015	human, microsoft researcher, image	Microsoft and Google are working on deep learning systems that can beat human at image recognition	Microsoft researchers say their newest deep learning system beats humans and Google-VENTUREBEAT (09/02/105)
Segment 16/02/ 2015	(i) 100 gb, dropbox user, storage, move (ii) human, image recognition	(i) Microsoft offers 100GB cloud storage free as a promotion to get more cloud users. (ii) Microsoft and Google are working on deep learning systems that can beat human at image recognition	(i) Microsoft offers free 100GB OneDrive space to Dropbox users worldwide-ZDNET (20/02/2015) (ii) Microsoft researchers say their newest deep learning system beats humans and Google-VENTUREBEAT (09/02/2015)
Segment 23/03/ 2015	chrome, chrome scrolling, microsofts help	Google is going to fix a scrolling issue of the Chrome browser using a Microsoft API	Google will fix Chrome's scrolling with Microsoft's help-ENGADGET(26/03/2015)
Segment 20/04/ 2015	amazon, cloud gang, nasdaq, revenue, cloud profit	NASDAQ spiked to its highest in 15 years, propelled by the revenues from cloud services of Amazon, Microsoft and Google	Techs surge on earnings, boosts Nasdaq, S&P 500 to record close-CNBC (24/04/2015)

Other topic pathways of both datasets show similar characteristics. Due to space restrictions, details of frequent terms for four prominent topic pathways (two each from two datasets) are provided in Table 4.7 and Table 4.8. Note that, the identifying terms in each pathway ('xbox' in $TP^2_{Microsoft}$; 'facebook' in $TP^5_{Microsoft}$; 'russia' and 'ukraine' in TP^1_{Obama}; and 'obama news' in TP^5_{Obama}) have been excluded to highlight subsequent frequent terms.

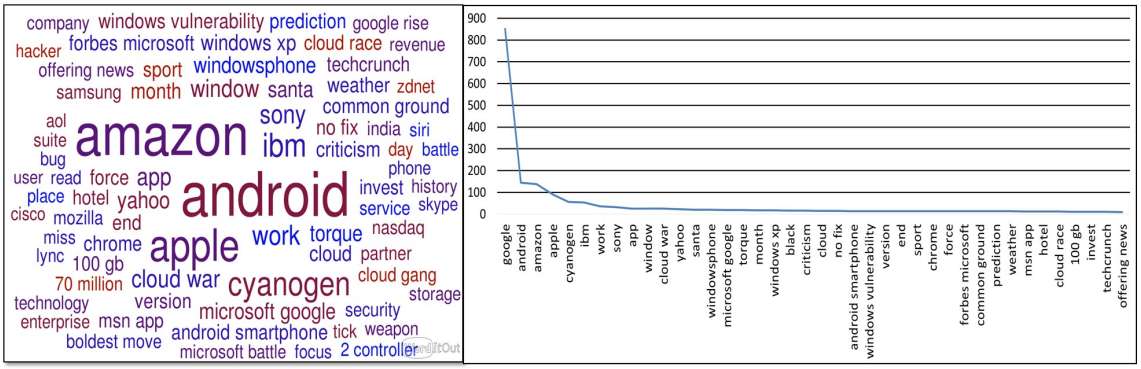


Figure 4.10: A word cloud and a word frequency graph generated for the topic pathway $TP^1_{Microsoft}$. Note that most frequent terms *Microsoft* and *Google* were removed from word clouds for improved clarity.

Table 4.7: Two topic pathways from #Microsoft Dataset represented using high frequent terms of the topic pathway and high frequent terms of top five most voluminous topic segments

Topic pathway	Frequent terms of topic pathway	Frequent terms of top five most voluminous topic segments				
$TP^2_{Microsoft}$	update, xboxone, game, windows 10, sony, bitcoin, plan, lumia, ps4, kinect, windows, playstation	*08/12/2014* bitcoin, windows10, november, sony, playstation	15/12/2014 game, pandora, apple tv, app, chromecast, xboxone	*22/12/2014* sony, christmas, ddos attack, playstation, lizardsquad	*16/02/2015* plan, week, update, ps4, xboxone, lumia	*11/05/2015* xboxone, ps4 , 10 april, playstation 4, sony
$TP^5_{Microsoft}$	bing, google, zdnet, apple, favour, monumental deal, search result, ganging, yahoo, sense, future, search	*08/12/2014* bing, search tool, safe preference, monday, graph search, 240 million, built-in search	*15/12/2014* bing , google, facebooks split, zdnet key event, amber alert, motley fool	*29/12/2014* google, apple, ganging, bing, worse, microsoft service, hotmail, msn, yahoo search	*18/05/2015* google, future, lead, bing, voice communication, sweden, eurovision, winner	*08/06/2015* monumental deal, virtual reality, alibaba, sense, google, oculus rift, investor, gamescom

$TP^2_{Microsoft}$ globally contains frequent words that are indicative of different products related to 'xbox' like 'xboxone', 'kinect' and 'windows 10'. Also, the word 'sony' denotes competitive manufacturer while 'ps4' and 'playstation' denotes competitive products. Notable subtopics include (i) *Segment 08/12/2014*: Microsoft accepts 'bitcoin' as a valid payment method to pay for Xbox games; (ii) *Segment 22/12/2014*: Xbox and Sony ps4 services are down due to a 'ddos attack' by a hacking group called 'lizardsqsuad' on Christmas day.

$TP^5_{Microsoft}$ has frequent terms 'bing', 'znet' and 'monumental deal' that are frequent terms of different segments. In addition, there are other organisation names such as 'apple',

'google' and 'yahoo'. Notable subtopics include (i) *Segment 08/12/2014*: Facebook drops Microsoft Bing in favour of its own search tool; (ii) *Segment 08/06/2015*: Facebook and Microsoft jointly working to make the Oculus Rift virtual reality headset works with Windows 10 and Xbox.

Table 4.8: Two topic pathways from #Obama Dataset represented using high frequent terms of the topic pathway and high frequent terms of top five most voluminous topic segments

Topic pathway	Frequent terms of topic pathway	Frequent terms of top five most voluminous topic segments				
TP^1_{Obama}	putin, economy, iran, respect, war, cuba, west, india, fight, china, weapon, merkel, pressure, europe, sanction, leader	*15/12/2014* putin, cuba economy, sanction, west, work, crimea, problem, foxnews	*22/12/2015* economy, dow, gas, foxnews, putin, cuba, head, golf, hawaii	*02/02/2015* putin, wwiii, merkel, kiev, arms, fight, Moscow, weapon, china, Kerry, respect	*09/02/2015* putin, weapon, war, merkel, poroshenko, fighting, hollande, minsksummit	*16/02/2015* putin, merkel, hollande, respect, allies, weapon, sanction, shot, war, west
TP^5_{Obama}	cbs news, obama video, obamacare, sore throat, gop, cuba, castro, acid reflux, summit, congress, prayer breakfast, iran, hospital, seattle, cnn, abc news	*22/12/2014* troop, 6 year, hawaii, apron, hacking row, cbs news, obamacare, afghanistan	*05/01/2015* community college, france, seattle, malia obama, union tour, congress, housing move	1*2/01/2015* obamacare, cameron, methane emission, iran, middle class, obama video	*02/02/2015* slavery, prayer breakfast, isis, crusade, obamas, comparison christianity	*06/04/2015* castro, cuba, summit, stage, saturday, panama, abc news, historic meeting

TP^1_{Obama} contains opinions about Obama handling Russia and Ukraine issues. Frequent terms of this pathway contains the names such as 'putin' (President of Russia), and 'merkel' (Chancellor of Germany). Also, the word 'sanction' is frequent showing that people talk about sanctions on Russia. Subtopics include (i) *Segment 15/12/2014*: Obamas support for new U.S. sanctions to Russia; (ii) *Segment 09/02/2015*: ahead of Minsk summit, Obama threatened the Russian President with serious consequences for involvement in the Ukrainian conflict.

TP^5_{Obama} contains news tweets about Obama often tagged as '#obama #news' . Tweets posted by news agencies have tags used by them such as 'cbs news' , 'cnn' and 'abc news' . Topic segments in this pathway is often a mixture of subtopics about different news articles. Notable subtopics: (i) *Segment 05/01/2015*: Obama proposes free two years at community colleges, Malia Obama supports anti-cop rap group, Obama to tout Arizona housing initiative; (ii) *Segment 12/01/2015*: Obama to call for tax increase on

rich to help middle class, proposes reduction in methane emissions from oil and gas, and Obama and Cameron talk counter terrorism and economy.

4.7.6 Emergence of new topic pathways

This subsection demonstrates the third core capability of the proposed algorithm; the detection of emerging topic pathways in the twitter stream. These new topic pathways are loosely related to previously identified pathways. Such new topics can be internal (e.g., new product or service launch) or external (e.g., interaction with a new organisation) to the business entity. Subsequently, some of them fade away as people lose interest, while some topic pathways prevail due to continuing interest.

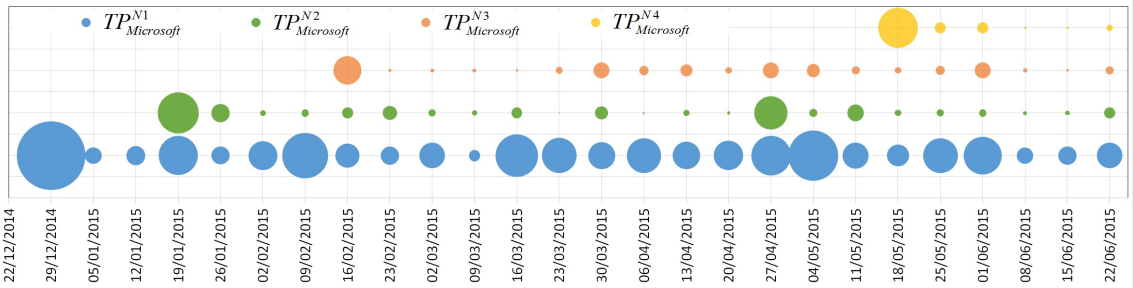


Figure 4.11: Four new topic pathways emerged in #Microsoft Dataset. Each topic pathway is represented using a chain of bubbles where each bubble is a topic segment. Diameter of a bubble is set to number of tweets in the topic segment.

$\boldsymbol{TP^{N1}_{Microsoft}}$: (frequent terms- windows 10, internet explorer, version, pirate, upgrade) Initiation of this new topic pathway coincides with the week where Microsoft announces that there will be a new web browser in Windows 10 operating system. Since then it has a significant amount of tweets that mainly discuss about new features in Windows 10. It is interesting to note that this pathway was separated as a new topic despite having a topic pathway focused on Windows) in general. We hypothesis that it is due to differentiating terms (e.g., windows 10, release, future, upgrade). Unlike other new topic pathways, this pathway retains its popularity throughout the time span of the dataset.

$\boldsymbol{TP^{N2}_{Microsoft}}$: (frequent terms- microsoft hololen, world, hologram, future, google glass) This pathway initiated when Microsoft introduced its virtual reality glasses 'Hololens' during 'Microsoft Consumer Preview' on 21/01/2015. After the initial peak it declined to few tweets a week except for the week 27/04/2015 in which the capabilities of Hololens was demonstrated at Microsoft Build Developer Conference.

$\boldsymbol{TP^{N3}_{Microsoft}}$: (frequent terms- linux, service, windows, python, bigdata) This pathway is initiated when Microsoft announces the incorporation of Linux operating system in its big data services provided in Microsoft Azure cloud platform. This move was widely discussed as Linux is a long term direct competitor to Windows and this is the first time that Microsoft incorporate Linux as part of its product/service suit.

$\boldsymbol{TP^{N4}_{Microsoft}}$: (frequent terms- salesforce, talk, msft, price, 55b) This pathway is focused on Microsoft's attempted acquisition of the company 'Salesforce' . It created hype on twitter as the news emerged and progressed for a couple of weeks. However, later declined

in popularity when it failed due to the hefty price tag of *55 billion USD* ('55b') demanded by Salesforce management.

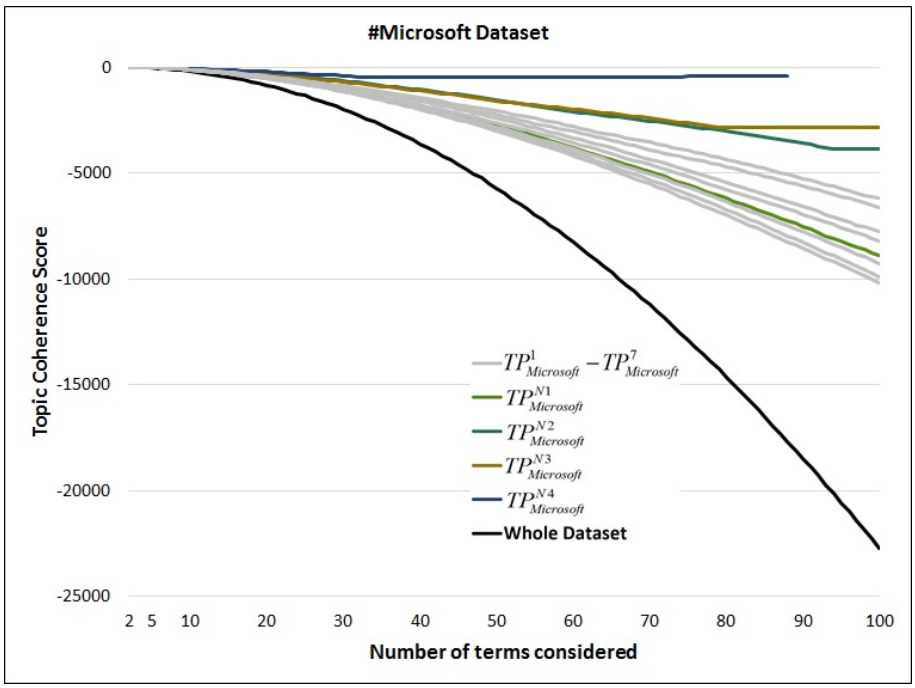


Figure 4.12: Topic coherence of newly emerged topic pathways.

Topic Coherence

Figure 4.12 presents the topic coherence metric for the above mentioned topic pathways plotted alongside the seven original topic pathways of the #Microsoft dataset. It shows that new topic pathways (except $TP^{N1}_{Microsoft}$) have increased topic coherence scores than the other prominent pathways. This is because they are focused on topics which emerged as a result of particular events and thus contain more closely related terms that are uniquely used to discuss them. $TP^{N1}_{Microsoft}$ has frequent terms shared with the topic pathway focused on Windows in general ($TP^{4}_{Microsoft}$) which results in relatively low coherence scores.

Detection of these new topic pathways highlights the capability of our proposed algorithm to identify previously unseen topics at any stage of the social text stream. This capability is due to the third extension of the extended IKASL*text* algorithm described in Section 4.4.3, which facilitates learning a new randomly initialised feature map from social media messages that are less similar to the existing topic pathways and spawning new topic pathways when there is a burst of social media messages on a new topic.

4.7.7 Automatic event detection

The aim of this section is to demonstrate the fourth capability; automatic detection of events from topic pathways using event indicators.

As previously shown in Equation 4.7, event score is determined by combining three event indicators for each topic segment in topic pathways.

The r parameters in Equation 4.7 decides the impact of each event indicator on the final event score. However, we have observed that the frequency and intensity changes of volume is higher compared to the changes of positive and negative sentiment. In order to compensate for this behaviour r_V is set to a lower value than others as follows $r_V = 0.1$, $r_{PS} = 0.45$ and $R_{NS} = 0.45$. We also employed $W = 2$ for the calculations of I_V, r_{PS} and r_{NS}. τ_e is set to 1.0 for this experiment and we filtered-out events that belong to topic segments containing less than 1% of the tweets in that batch.

Table 4.9 and Table 4.10 present the top 10 events identified based on event score $\mathcal{I}$ in the two data sets. The frequent terms that are related to an event are derived by capturing the frequent terms that appear in the topic segment containing the event but does not appear in the frequent terms of the W previous topic segments of the same pathway. We analysed online news articles from major news agencies to verify the significance of events detected by the algorithm.

The above tabulated results demonstrate and verify the authenticity of events detected by the proposed algorithm. It is interesting to note that these events were captured exclusively based on social expressions on Twitter. It is also interesting to observe that different events were triggered by high values from different indicators. Volume based indicators are high in the first event of Table 4.9, which is about an event where Microsoft demonstrated the capabilities of its virtual reality product Hololens. As shown in Figure 4.11 a new topic pathway is spawned when Hololens was first introduced by Microsoft on 21/01/2015.

Another example where I_V is high is the third event in Table 4.9, which is due to Microsofts statement that Windows 10 would be a free upgrade for all users (including those with pirated). This statement attracts attention on online social media platforms as it was a surprise move by Microsoft (quite often aggressive about software piracy). This peak of volume can also be seen in Figure 4.11 for $TP^{N1}_{Microsoft}$.

There are several events in both tables that are significantly high in negative sentiment event indicator score. Event 2 in Table 4.9 is the highest triggered by the event where Xbox Live online services were offline due to a hacker attack on Christmas day. This event has caused issues for online gaming community who took it into Twitter criticising Microsoft for lack of security and sluggish response.

Event 2 in Table 4.10 is highly negative, which was due to firstly, people condemning the release of nine suspects held in Guantanamo Bay and secondly the Charlie Hebdo Shooting in the same week. Social commentary combined both incidents stating that releasing of terror suspects can lead to more attacks. This incident is identified from topic pathway focused on Iran/Israel relations, which is loosely related to the event since most of the released suspects are from Middle East.

Events with high positive sentiment are comparatively rare. This is largely because online social media platform Twitter is more often used to express negative sentiment more than positive. It also supports the finding that popular events are often associated with negative sentiment (Thelwall et al., 2011).

This subsection aptly demonstrates the final capability of the proposed platform, automated event detection from social text streams. The incorporation of several event

Table 4.9: Top 10 events detected based on event score $\mathcal{I}$ for #Microsoft Dataset

Segment	Topic pathway	$\mathcal{I}$	Event indicator scores 0.1 $\times I_V$	0.45 $\times I_{PS}$	0.45 $\times I_{NS}$	Frequent terms related to the event	Verification evidence based on news articles
27/04/2015	$TP^{N2}_{Microsoft}$	1.62	0.69	0.53	0.4	microsoft hololen, google glass, build2015, augmented reality, future	Microsoft demonstrated the capabilities of Hololens in Microsoft Build Developer Conference
22/12/2014	$TP^{2}_{Microsoft}$	1.37	0.1	0.52	0.75	psn, ddos attack, christmas, lizardsquad, credit, lizard squad	Xbox Live online services were down on Christmas day after a DDoS attack by a group called 'LizardSquad'
16/03/2015	$TP^{N1}_{Microsoft}$	1.31	0.52	0.36	0.43	windows 10, pirated version, summer, pirate, password	Microsoftannounced that it will give free upgrades to Microsoft 10, even for pirate copies
30/03/2015	-	1.26	0.48	0.42	0.36	Iphone, rogue window, android phone, office lens, document, powerful scanner	Microsoft launches Office Lens scanner app to iPhone and Android for the first time
12/01/2015	$TP^{1}_{Microsoft}$	1.22	0.18	0.54	0.5	criticism, window, tick, endpoint risk, windows 8	Google released malicious code which can be used to exploit Microsoft Windows
11/05/2015	-	1.16	0.2	0.4	0.56	microsoft hyperlapse , video, desktop, hyperlapse, shaky video, launch	Microsoft launches Hyperlapse, a video smoothing software for Android and Windows phones
30/03/2015	$TP^{4}_{Microsoft}$	1.15	0.15	0.52	0.48	open source, windows sensor, open-source, microsoft exec, sale	Microsoft executive says that it is possible that Windows become open-source in future
09/02/2015	$TP^{7}_{Microsoft}$	1.14	0.24	0.4	0.5	google app, note, iwork, windows user, microsoft office, 700 billion, watch, google threaten, work	Google just threatened to expose Apple and Microsoft security flaws. Apple opens up iWork to Windows users for free
09/03/2015	-	1.13	0.3	0.45	0.38	microsofts cortana, report, standalone app, ios device, digital assistant, siri	Microsoft announced its intelligent personal assistant Cortana will be released for iPhone and Android
09/02/2015	$TP^{N1}_{Microsoft}$	1.13	0.2	0.48	0.45	public preview, lumia , first preview, microsoft windows10forphone, lumia 635	Microsoft delivers first public preview of Windows 10 for phones

indicators allow us to capture events across diverse domains without having to develop tailored event capture mechanisms for individual applications.

4.8 Chapter Summary

This chapter presented an algorithmic development of the conceptual model proposed in Chapter 3. It consists of two novel algorithms designed to capture insights from fast-paced social data streams. The first is a new unsupervised incremental machine learning algorithm developed extending GSOM self-structuring algorithm and IKASL incremental learning algorithm. It automatically structures a social data stream into topics and extends those across time into topic pathways. It also captures changes in topics over time as well as distinct new topics as new topic pathways at different points of time. This algorithm was designed to overcome the challenges present in social data with respect to its brevity, unstructuredness, and diversity. The second algorithm is a multi faceted event detection algorithm developed to monitor topic pathways for significant changes in online social behaviours over time, and identify such changes as potential events of interest. The changes in social behaviours were identified using automatic event indicators such as changes in volume, positive sentiment and negative sentiment.

These techniques were trialled using two large Twitter datasets containing 6 months of tweets on two entities a politician and an organisation. As shown in the experiment results, the topic pathway separation algorithm successfully captured the key topic pathways representing ongoing discussions related to the respective entities. Also, shifts in the discussions represented by new key terms were successfully learned and associated with the relevant topic pathway. Moreover, new distinct topics were automatically captured as new topic pathways. The event detection algorithm monitors those topic pathways and automatically captured significant changes in human behaviour using changes in volume and sentiment. Those captured events were aligned with contemporary news articles that discussed those events.

The next chapter extend the above developed algorithms into a technology platform to capture insights from online support group discussions.

Table 4.10: Two topic pathways from #Obama Dataset represented using high frequent terms of the topic pathway and high frequent terms of top five most voluminous topic segments

Segment	Topic pathway	$\mathcal{I}$	Event indicator scores			Frequent terms related to the event	Verification evidence based on news articles
			0.1 $\times I_V$	0.45 $\times I_{PS}$	0.45 $\times I_{NS}$		
19/01/2015	-	1.22	0.33	0.52	0.37	tuesday, netanyahu, india visit, proof, nice, world, obamainindia	Obama visits India
05/01/2015	TP^3_{Obama}	1.22	0.1	0.42	0.7	gitmo, france, freedom, paris, press, terror, war	Obama released 9 suspects held in GTMO prison. Obama condemns Charlie Hebdo attack saying terrorists fear freedom
26/01/2015	TP^6_{Obama}	1.16	0.08	0.43	0.65	white house, taliban, muslim province, white-house, japanese journalist, islamicstate, beheading	ISIS threatens to behead Obama and transform America into a Muslim province. ISIS killed second Japanese Hostage
30/03/2015	TP^3_{Obama}	1.16	0.23	0.49	0.44	irandeal, lying , obama kerry, detail, framework, trust, irantalk, good deal, nuke agreement	
02/02/2015	TP^6_{Obama}	1.15	0.12	0.41	0.62	death, january, crusade, unemployment rate, campaign, christianity, muslim leader, teaparty, immigration	Obama, at National Prayer Breakfast, compares ISIS to violence of Christian Crusades. USA unemployment rate rises to 5.7% in January
19/01/2015	-	1.13	0.07	0.57	0.49	battle, marijuana legalization, tntweeters obama, crisis, obama cybercare	Obama on Thursday said he expects more states to experiment with marijuana legalization
16/02/2015	-	1.13	0.14	0.6	0.39	fabric, founding, woven, love, obama islam, giuliani, obamas love, gov walker, mayor giuliani	Obama created controversy by saying Islam Has Been Woven Into the Fabric of Our Country Since Its Founding. New York Mayor Rudy Giuliani said Obama doesn't love America
22/12/2014	TP^5_{Obama}	1.13	0.15	0.5	0.48	6 year, apron, hacking row, novice, obama motorcade, afghanistan, trend, malaysian leader, afghan milestone	Obama hails end of combat operations in Afghanistan
23/03/2015	TP^5_{Obama}	1.12	0.1	0.49	0.53	obamacare, price, double funding, antibiotic resistance, national plan, senate, superbug	Obama announced a 5-year action plan to prevent lives lost from antibiotic resistance
19/01/2015	TP^1_{Obama}	1.12	0.19	0.57	0.36	respect, russia, ukraine, saudi arabia, pressure, saudi king, putin , india trip, russian aggression	Obama cuts short presidential trip to India, and travelled to Saudi Arabia to pay respects to King Abdullah. Obama pledges more pressure on Russia as Ukraine clashes broaden

Chapter 5

Expatiation from algorithm to technology platform

I alone cannot change the world, but I can cast a stone across the waters to create many ripples -Mother Teresa

This chapter presents the advancement of the conceptual model for social behaviour understanding proposed in Chapter 3, and the new incremental machine learning algorithms, based on the principles of self-structuring artificial intelligence, proposed in Chapter 4, into a functional technology platform, that delivers on the ambition of common good and contribution to human society. More specifically, the focus of social good is on patient-centred healthcare and the role of online support groups (OSG).

OSG are social forums that provide peer support, mainly for physical and mental health related issues. Unlike fast-paced mainstream social media platforms, OSG are relatively low volume social data streams where the bigger OSG generate few hundreds of social media messages daily compared to millions in Twitter. The participants of these platforms engage in longer discussions, and such discussion mainly focused around the main theme of OSG. Engaging in longer discussions lead to the formation of stronger ties between the OSG members where they support each other to cope with the hardships of the disease conditions. Because of these stronger ties participants tend to frequently do deeper self disclosures which include information about their disease conditions and deeper emotions related to their wellbeing.

The proposed technology platform operates in multiple stages to better facilitate the different information needs of its stakeholder groups: consumers, researchers and health professionals. It extracts information encapsulated in free-text discussions of OSG using a suite of machine learning, natural language processing and emotion analysis techniques and automatically generate individual profiles for each OSG user which enriches the information retrieval process.

The chapter is organised as follows: Section 5.1 describes an OSG, its purpose and functions, Section 5.2 defines the stakeholder groups of OSG. Section 5.3 delineates limitations of OSG and Section 5.4 presents the proposed technology platform that addresses

such limitations and advances the role of OSG in understanding social behaviours. Finally, Section 5.7 presents a statistical evaluation of the technology platform, followed by Section 5.8 which demonstrates its capabilities on two high-volume OSG.

5.1 Online Support Groups (OSG)

Health related information seeking has become one of the major use of internet, where surveys reported that in 2010, 80% internet users (59% of all adults) in United States have looked online for health related information such as specific disease and treatment options (Fox, 2011b). Moreover, 34% internet users have read online resources that contain consumer generated content about health issues and 18% internet users have searched internet to find others with similar diseases or health concerns (Fox, 2011b, 2013).

One of the growing sources of consumer generated online health information is Online Support Groups (OSG) which are also known as online health communities and health forums. OSG are virtual communities that are established to discuss or exchange knowledge about health related topics. OSG are formed in many different forms of online social media platforms such as discussion forums, blogs, wikis, groups in mainstream social media platforms (e.g., Facebook groups), and Q/A websites (e.g., Yahoo! Answers [1]). While there are few professional-led OSG, most OSG are peer-to-peer and moderated by few self-appointed senior users of the OSG. It is reported that seeking such peer-to-peer support is prominent among patients living with chronic illnesses such as high blood pressure, diabetes, heart or lung problems and cancer (Fox, 2011a).

OCSG discussions are organised as discussion threads where each thread starting with a question, comment or an experience about corresponding individual‘s health concerns. Other OSG members (patients, partners of patients or other caregivers) respond to such concerns and thereby create discussion threads.

Figure 5.1.a shows the home page of a popular OSG, *patient.info*[2]. It contains a *free-text* search and a list of high level topics (e.g., Mental health and Women's health). Figure 5.1.b shows a sample OSG thread where the initiating author starts with the title ‘stomach pain’ and a description about his experience, which includes demographics of the author as well as the symptoms that he encountered. Figure 5.1.b also shows four subsequent posts (out of the six replies in the thread), in which first, second and fourth seems to be advice and the third is further information by the initiating author.

5.2 Users of Online Support Groups

OSG are contributed by consumers (e.g., patients or caregivers), health professionals and health researchers who are been denoted as the three main user groups of Medicine 2.0 applications (Eysenbach, 2008). Figure 5.2 highlights the activities and interactions of these user groups.

[1] http://answers.yahoo.com
[2] http://patient.info/forums

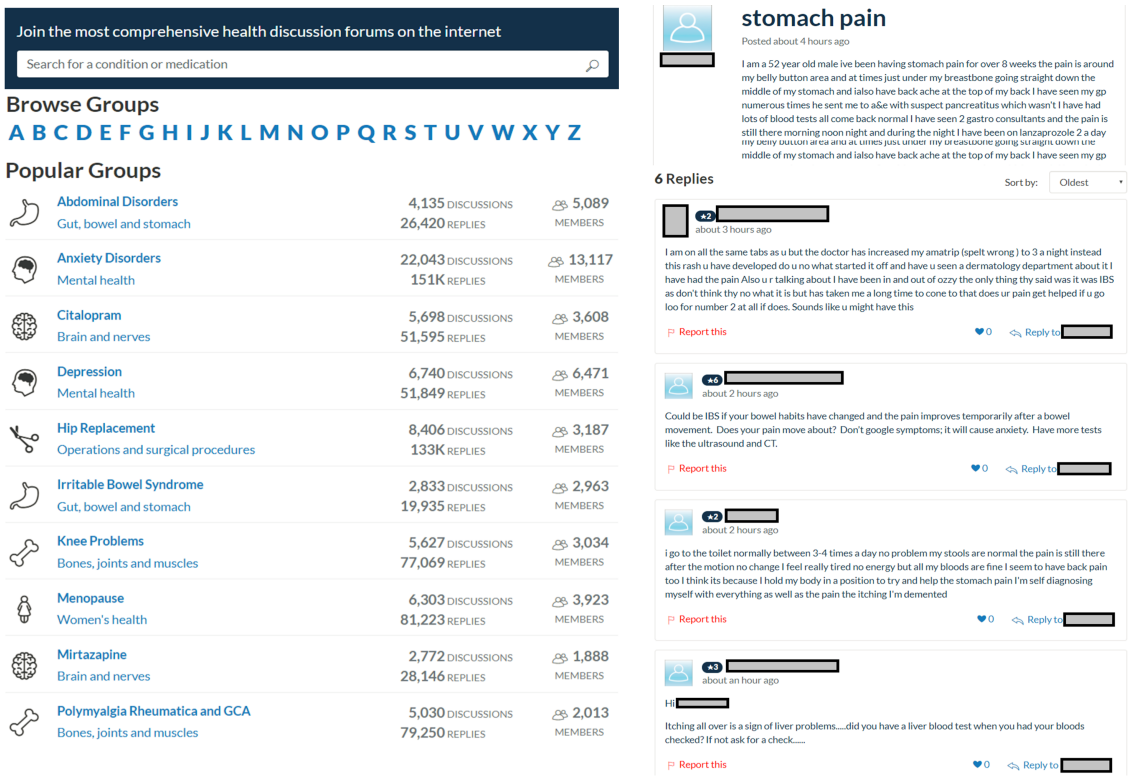


Figure 5.1: a. Home page of patient.info OSG, which shows the popular discussion groups and the simple search interface, b. A sample OSG discussion thread.

1. Consumers: OSG are mostly contributed by health consumers who are often patients or caregivers of patients (e.g., partner, children, family). They seek and receive support from other OSG users (peers) in OSG as well as provide support to others.

2. Health professionals: professional involvement in OSG varies, where some OSG have designated health professionals to monitors and mediate the health information exchanged, while in most OSG professionals involve voluntarily to provide their professional opinion to support seekers.

3. Researchers: researchers often scrutinise the OSG discussions thematically to understands discussions topics as well as to find suitable participants for research studies.

5.2.1 Motivations to participate in OSG

OSG are mainly used by patients and caregivers to fulfil some of their social needs discussed in Section 3.2.2 such as belongingness/affiliation, achievement and power over others. These were achieved using social behaviours such as self-disclosure and social comparison.

Functionally, motivations to participate can be categorised into three: (i) seek informational support, (ii) seek emotional support and (iii) provide support.

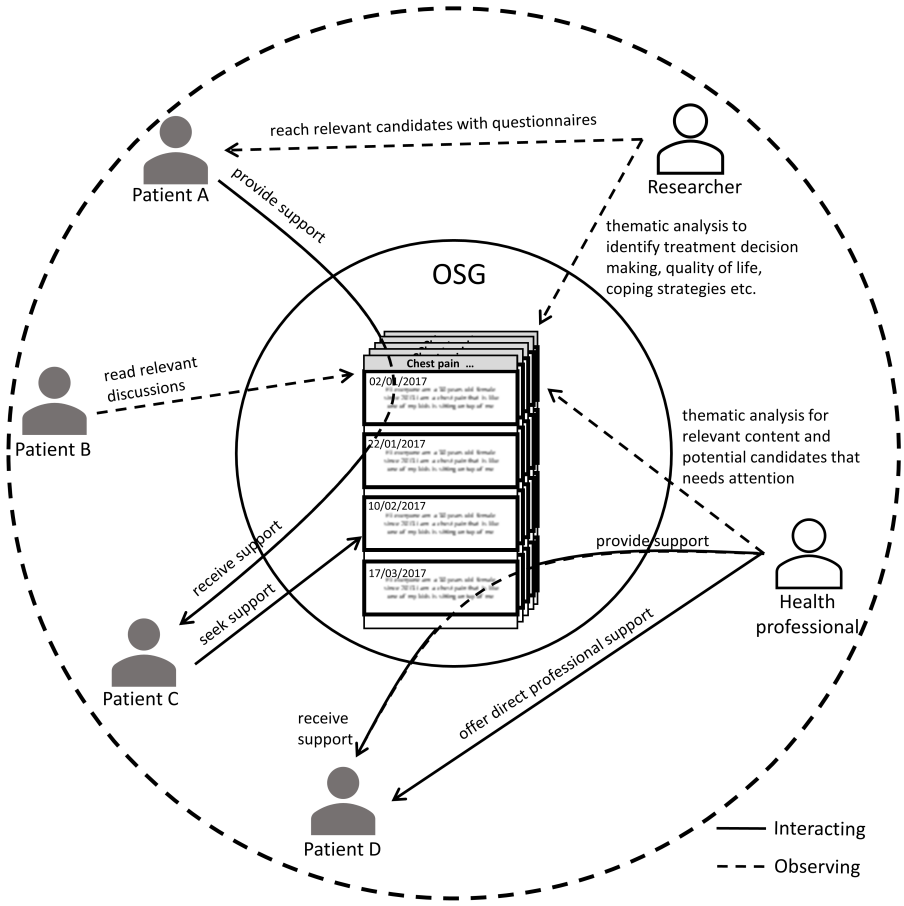


Figure 5.2: Stakeholders of Online Support Groups (OSG).

Seeking informational support

Information support is sought for information related to a particular health condition. Individuals first self-disclose their information related to the interested health condition and either ask specific questions or similar experiences. The objective of seeking information support is two-fold. First is to gain knowledge about the health condition. The second objective is to compare (social comparison) their situation to similar situations of others. Such comparisons enable individuals to get an understanding of their situation compared to others, and also as discussed in Section 3.2.4 downward social comparisons reduce distress related to the health conditions by knowing that others have similar or even worse conditions/sufferings.

Information support seeking happens at three different stages of the healthcare process such as initial diagnosis, treatment decision making, and post-treatment quality of life (Hu et al., 2012; Ziebland and Wyke, 2012).

Information seeking in initial diagnosis is mainly to re-affirm the diagnosis made by the clinician. It is reported that comparison of own symptoms and disease conditions against others with similar diagnosis helps to boost the confidence about the diagnosis as well as the clinician (Ziebland and Wyke, 2012; Mazanderani et al., 2012).

Treatment decision making process is another key stage where people seek informational support (Huber et al., 2011; Zhang, Bantum, Owen, Bakken and Elhadad, 2017;

Berry et al., 2003), especially when there are multiple treatment options with contrasting pros and cons. The accounts of people who have undergone different treatment options are scrutinised to capture the decision making criteria as well as their contemplation about those decisions afterword.

Another most sort after information is the post treatment quality of life from the people with chronic conditions, where people look for coping strategies and lifestyle issues (Ziebland and Wyke, 2012). Such information seeking is more prominent when the issues are in intimate and delicate in nature (Blank et al., 2010; Huber et al., 2011).

Seeking emotional support

Emotional support is sought for emotional issues related to disease conditions. Emotional support attempts to satisfy belonging social need by facilitating an understanding, supportive and empathetic audience where users can express their emotional issues related to the disease conditions. OSG members provide supportive, encouraging and positive response to such issues and thereby reduce the emotional burden of the disease conditions (Tanis, 2008; Bar-Lev, 2008).

Another key problem specifically faced by people with chronic illnesses is the disconnect or isolation from the community (work or living) where they were once part of (Williams, 1984; Bury, 1982). Some individuals may even feel embarrassed and stigmatised by their illness and afraid to face their living community. For such individuals, contributing to OSG provide a sense of belonging to a community that consists of others with similar circumstances, which helps to reduce the feeling anxiety and stress resulted due to the isolation of community (Bar-Lev, 2008; Harvey et al., 2007).

Moreover, it is reported that individuals with chronic health issues reduce their anxiety by comparing their situation to others with similar circumstances (Frost and Massagli, 2008). Bender (2011) explains this phenomenon using the theory of social comparison processes (Festinger, 1954), which states that during uncertain situations individuals tend to seek others with similar circumstances to compare their behaviours and abilities.

Providing support

One of the vital element of the OSG ecosystem is the voluntary support provided by peer OSG users. Chiu et al. (2006) argue that willingness to share knowledge to support peers is key to the fostering of virtual communities such as OSG. Providing support satisfy needs such as a sense of belongingness to the community as well as a sense of achievement for supporting others. Also, it may give a sense of power over others for influencing. The following are the ways that OSG users provide peer-support.

- **Answer direct questions:** Most of the direct questions are on information support discussed in Section 5.2.1. Users answer such direct information requests, based on their experience in similar health concerns. A study on users in a Diabetes OSG (Yan Zhang, 2016) has found that users often answer those question because they are confident about their understanding of the illness. Also, answering questions made

the users felt proud or empowered as a contributor and a mentor to the OSG (Yan Zhang, 2016; Kummervold et al., 2002).

- **Self-disclosure:** Users often self-disclose their story of coping with health concerns as support for other users facing similar situations. Such disclosures help to find important clues to cope with the situation and also it serves as emotional support for the receiving users as they tend to feel belonging to a social group with similar health concerns (Høybye et al., 2005; Ziebland and Wyke, 2012).

5.2.2 Role of professionals in OSG

Support groups (both face-to-face and online) often have different levels of professionals support which varies from professional-led to patient-led (Shepherd et al., 1999). Professional-led support groups are often small closed groups (by invitation only) which facilitates support for certain patient groups of a particular healthcare organisation. Such groups are mediated and quality controlled by a board of health professionals. While such quality control reduces the risk of spreading misnomers, researchers have found that excessive professional involvement tends to reduce patients supporting each other (Kurtz, 1990). In addition to mediation, professional therapist often involved in OSG to provide emotional support to patients and caregivers.

Most of the current popular OSG are patient-led, with no designated professional body to mediate or quality control. However, it is observed that health professionals voluntarily join the OSG as members and provide their opinions as informational and emotional support to the other OSG members. In addition, some health professionals refer to relevant OSG discussions to understand the patient concerns related to disease conditions.

5.2.3 Opportunities for researchers in OSG

OSG contain unsolicited accounts of first person experiences, which is a rich source of information for the medical researchers to study different aspects of patient experience (Robinson, 2001). There are two main avenues that OSG are used by researchers.

In the first approach, Medical researches use OSG to recruit potential candidates to administrate interviews or questionnaires for various research studies. OSG provide access to user groups aggregated with similar health concerns which are used by the researchers to send invitations to participate in various studies. One of the key advantage of OSG is it provide access to concentrated groups that are otherwise difficult reach with conventional methods, which can be due to either rareness of the inclusion criteria or social stigmatised nature of the illness (Wright, 2006). Moreover, it allows the researchers to reach out to a large group of potential candidates within a short period of time. This approach uses OSG only as a recruitment platform for conventional survey based methods, without paying attention to its peer-to-peer support ecosystem or the content of the discussions.

In the second approach, medical researchers retrospectively analyse OSG discussions of the selected groups with specific selection criteria. This approach is less intervening

compared to the previous, as it does not reach the users directly, but more exhaustively analyse the content of the OSG messages. The self-reported expressions of an individual are scattered across the OSG as parts of different discussion threads. However, once aggregated and ordered the OSG posts of an individual provides significant insights about the patient journey over time which can be exploited to analyse their demographics as well as the clinical outcomes and the associated emotions across the patient journey. Such analysis is more beneficial for a chronic illness as it provides the opportunity for the researchers to evaluate the self-reported quality of life of each individual against the clinical outcomes at different stages of the patient journey.

In the current research most of these studies are conducted manually or semi automatically, often as thematic analysis of the OSG discussions of chronic illnesses such as breast cancer (Høybye et al., 2005; Huber et al., 2017; Winzelberg et al., 2003), prostate cancer (Gooden and Winefield, 2007; Huber et al., 2011), HIV (Bar-Lev, 2008), and Autism (Zhang, Kang, Qiu, Zhang, Yu and Elhadad, 2017). The limitation of this approach is that due to the exhaustiveness of the process, such analysis is limited to few OSG participants and often limited to analysing only the clinical outcomes without correlating them with the associated emotions expressed. This mainly due to the challenging and nature of extracting such information from noisy and unstructured free-text in patient-authored OSG posts.

5.3 Limitations in finding relevant information

As discussed in Section 5.2.1 OSG plays an important role as an online platform that enables users to engage with virtual peers who have (or had) similar health concerns. One of the key elements of these engagements is finding users with similar health concerns. Based on a survey of diabetes OSG users, Yan Zhang (2016) points out that the users look for similar or related users mainly based on similar illness conditions and comparable demographics. In the existing OSG infrastructure users achieve this task by manually scrutinising the posts in OSG threads for clues about the illness conditions and demographics of the other users.

Moreover, as shown in Section 5.2.3 researchers look for OSG users with particular selection criteria (e.g., high risk prostate cancer patient who has gone through robotic assisted laparoscopic radical prostatectomy for prostate removal), where such user were either contacted to administrate questionnaires or their OSG discussions were analysed thematically depending on the research methodology. Such selection process is also carried out by manually going through the posts of OSG users while evaluating such users eligibility based on the self-stated information in OSG posts.

However, with the widespread use of internet and popularity in online health information seeking across individuals from all walks of life, use of OSG participation is growing rapidly. Table 5.1 presents several usages statistics of three large OSG. It shows that those OSG are consists of over 100,000 registered users (*healthboards* has over 1,000,000 registered users). Moreover, these OSG consists of over 100,000 discussion threads and more than 1,000,000 posts. Therefore, it is impossible to manually scrutinising these colossal

text croups or go though large number of users to find the information that either users or researchers are looking.

Table 5.1: Population and volume statistics of three large OSG. Note that this information are as of 02/04/2018.

OSG	registered users	threads	posts
healthboards	1,151,152	909,486	4,975,203
healingwell	161,571	323,288	2,946,335
patient.info	161,734	225,220	1,882,821

Based on the existing capabilities of these large OSG, consumers (users or researchers) narrow down their information retrieval process using two main approaches: (i) by navigating through the hierarchy of topic often exist in OSG platforms to find the topics relevant to theirs health concerns, and (ii) using the free-text search function that is available in some OSG to lookup threads based on relevant keywords.

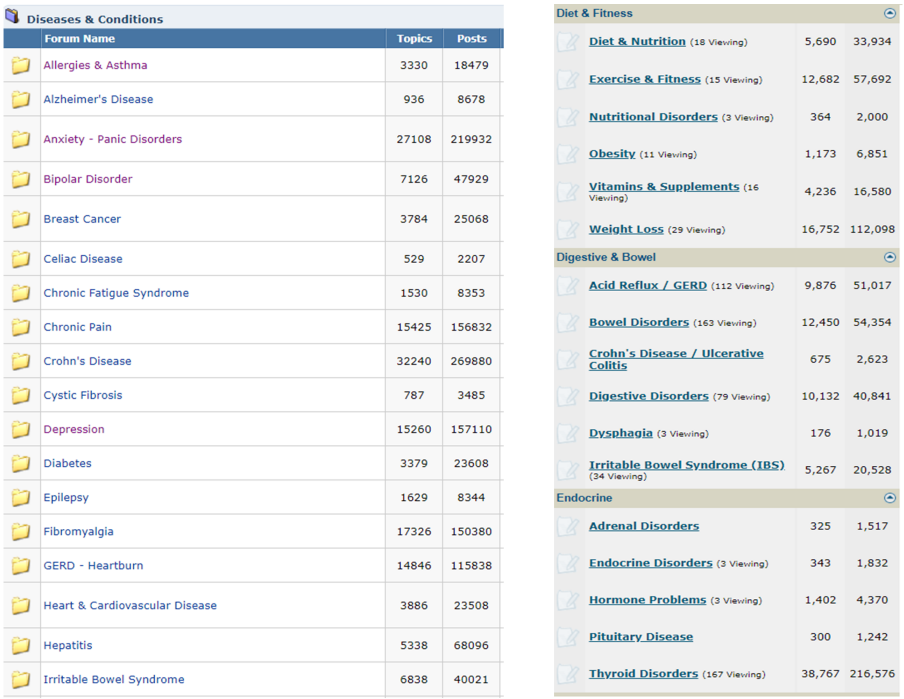


Figure 5.3: Topic hierarchies in (a) healingwell and (b) healthboards OSG.

Figures 5.1.a, 5.3.a, and 5.3.b show parts of the topic hierarchies in OSG *patient.info, healingwell, and healthboards* respectively. As shown, the topic hierarchies in OSG are fairly shallow, often limited to two levels. Hence, the narrowing down by topic can still leave large number of posts for manual analysis. Moreover, these topic structure of the OSG are user generated, often by the moderators as per popular requests from users. Therefore, the topic hierarchy is developed incrementally, and thus can be overlapping and sometimes confusing. For example, in healthboards (Figure 5.3.b), the second level topics under 'Digestive & Bowel'can be overlapping as 'IBS'can fall under both 'Digestive Disorders' and 'Bowel Disorders'.

Free-text search capabilities of those OSG are often basic and limited to a keyword search across the entire text corpus of the OSG, while some OSG allow keyword or phrase search across a selected topic. These search capabilities are facilitated by indexing the terms found in each post (except stop-words) so that they are searchable. Free-text search is demonstrated below using an example query "Im a 40 year old woman taking Nexium for heartburn".

1. Tokenize the query into words (I, m, a, 40, year, old, woman, taking, Nexium, for, heartburn).
2. Remove stop-words and stem-words to their root form (40, year, old, woman, take, Nexium, heartburn).
3. Search the OSG database for the posts that have any of these words.
4. Determine a relevance score for each identified post by aggregating tf-idf scores of each matching word (tf-idf is a statistical measure of how important a word is to a document in a collection of documents).
5. Present the top results with the highest relevance score.

The top results consist of posts that contain several matching words in the end-user query. However, it does not recognise that heartburn is a symptom, Nexium is a medication and the end-user is interested in posts that mention both of these words. In addition, it does not recognise 40 is an age mention and woman is a gender mention.

These limitations of topic based navigation and free-text search results the precision and recall of the current information retrieval capabilities of the OSG.

Low precision: in an information retrieval task, precision is the fraction of the retrieved documents that are relevant (Manning et al., 2008). Topic based navigation ends up with a substantial amount of OSG post that is of little relevance to the health concerns of the users. Also, as demonstrated in the above example, free-text search pulls out OSG posts that have one or few relevant terms, which may not relevant. For example, there may be OSG posts with the terms 'old' and 'woman' but with different symptom mentions of disease conditions, thus, not relevant to health concerns in the query.

Low reliability in recall: in an information retrieval task, recall is the fraction of relevant documents that are retrieved (Manning et al., 2008). Topic based navigation approach loose recall when users post relevant content under different topics, which can happen due to confusing topic structures or users misunderstanding the scope of a topic. Free-text search loose recall mainly because it is not capable of identifying the different types of information present in the query or OSG posts. For example, OSG posts from a user of age 41 with similar health concerns are a very close match for the above given query, but will not be given a higher relevance in the free-text search. Moreover, users only mention a portion of information in a single OSG post. For example, a user might have mention relevant health concerns in a one OSG post and relevant demographics in another, but the free-text cannot automatically aggregate them and show it as a relevant result.

The above issues are mainly due to insensitivity towards different types of information (e.g., demographic, clinical). These types of information are crucial to understand the context of OSG posts as well as the search criteria. Moreover, the information encapsulated in OSG posts needs to be aggregated by the author to get a more comprehensive picture of each author. The next section proposes such information structuring model which can overcome the limitations of the existing information retrieval capabilities in OSG.

5.4 Proposed technology platform

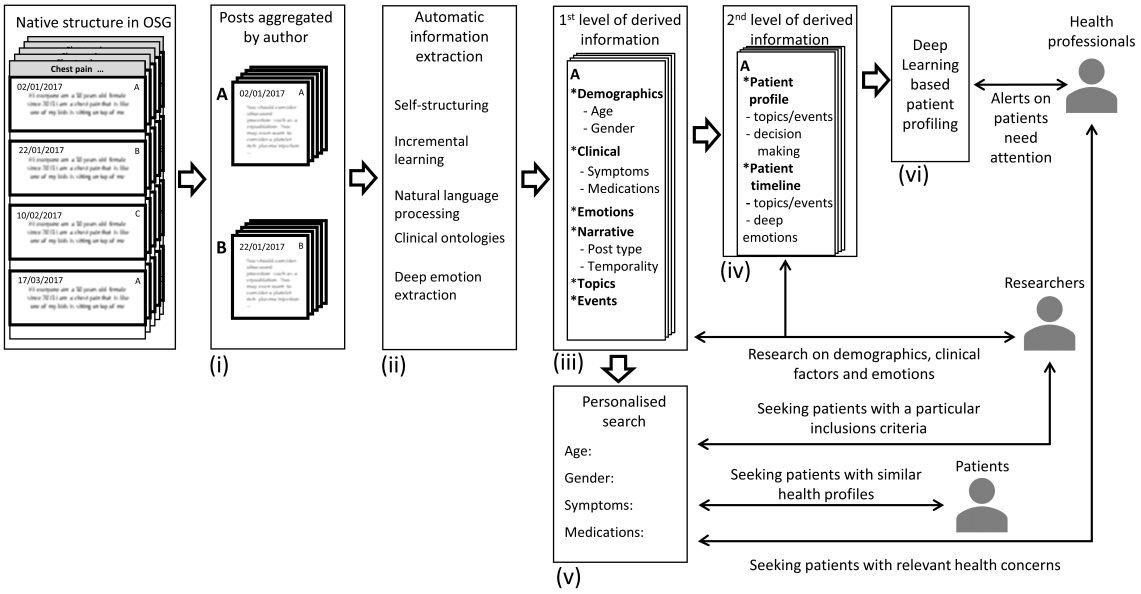


Figure 5.4: The proposed multi-stage self-structuring AI platform and its benefits for the three stakeholder groups in OSG.

Figure 5.4 presents the proposed multi-stage self-structuring AI platform which is an implementation of the proposed self-structuring artificial intelligence platform in Section 3.3. It employs machine learning and natural language based information extraction methods to transform the free-text corpus of discussion threads into a structured multi-dimensional information database pivoted around each user to better facilitate the information needs of three groups of stakeholders in OSG. Stages of this framework is as follows:

i The OSG posts are aggregated by the author ID of each OSG user. As discussed, OSG users include bits and pieces of information in different posts, posted across multiple discussion threads. Therefore, aggregating the posts by each author enables more comprehensive information extraction from each user.

ii Natural language processing and machine learning based techniques are employed; with the support of clinical ontologies and emotional thesaurus; to extract and infer information of each OSG user based on the free-text content of their aggregated OSG posts.

iii The first level of information consists of demographic, clinical, emotional state and narrative type information about each user extracted from the previous stage. This information is stored in a multi dimensional database pivoted by the user ID of each user.

iv The second level of information consists of complex information structures such as patient event timeline, which is constructed based on the clinical and emotional events derived in the previous stage.

v Structured search retrieves OSG members with a given set of clinical factors and range of demographics.

vi Deep learning techniques are employed to learn the personal traits (e.g., depression) of the OSG users based on their emotions and linguistic patterns in discussion threads.

The information structured in stages slowromancapiii@ and slowromancapiv@ opens up new opportunities for researchers to conduct cohort studies on the OSG population with particular health concerns. Emotional information can be employed in a similar manner as survey instruments to quantify the quality of life of patients. Aggregated quality of life across different dimensions can be used to compare and contrast different treatment types and side-effects. The patient profiles and timelines extracted in stage slowromancapiv@ can be employed for in depth research about patient decision making behaviour and decision factors as well as emotional trajectory over time of patients suffering from chronic illnesses.

The structured search in stage slowromancapv@ enables the patients to search up other patients with similar health concerns and demographics. Such functionality also benefits users who are keen to provide peer-support as they can easily look up the users with similar health concerns to provide support. Researchers can also use this structured search to apply their inclusions criteria of the research cohort to look up potential candidates for research. Health professionals can use it to lookup patients who have health concerns relevant to their field of expertise.

The stage slowromancapvi@ use deep learning techniques to model personal traits of the patients, based on their discussion topics, language pattern and emotions so that it can identify the behaviours of patients that needs the attention of a health professional. For example, a patient posting depressive content can be brought into the attention of a health professional on mental health.

The implementation, evaluation and validation of stages slowromancapiii@ to slowromancapvi@ of this framework is presented in the rest of this chapter. A use case of this platform on prostate cancer related OSG are presented in Chapter 6.

The next section presents the implementation details of stages slowromancapii@ and slowromancapiii@ and their evaluations can be found in Section 5.7. Section 5.6 presents the implementation of stage slowromancapiv@, and Section 5.8.3 provides implementation details of stage slowromancapv@ as part of the demonstration.

5.5 Information retrieval techniques

This section presents the information retrieval techniques developed for stages slowromancapii@ and slowromancapiii@ of the proposed information structuring platform discussed in the previous section (see Figure 5.4). This section is organised as follows: Section 5.5.1 presents the narrative type extraction, Section 5.5.2 presents age extraction, Section 5.5.3 presents gender extraction, Section 5.5.4 presents medical concept extraction and Section 5.5.5 presents the emotion extraction.

5.5.1 Narrative type extraction

Online social platforms such as OSG are used to seek, provide and exchange information (Zhang et al., 2015; Oh, 2012). As discussed in Section 5.2.1, individuals share similar experiences and also provide advice to their peers in the OSG.

Differentiating the posts as experience sharing or advice is an important task. Experience sharing posts are important as they contain self-disclosed information of patient factors (Sadovykh et al., 2015), which includes demographic, clinical and emotional information during various stages of the patient timeline. This information can be employed for patient profiling. On the other hand, advice can be used to understand the nature of peer-advice in OSG.

OSG community is consists of not only patients but also a significant proportion of caregivers who posts on behalf of the patients that look after. Hence, the expressions of experience in OSG can be further categorised as an expression of own experience or an expression of experience of someone else (often family member, close relative or a friend). Thereby, in this work the posts are grouped into three narrative type categories (i) experience: first-person, (ii) experience: second-person, and (iii) advice.

Detecting experience expressing posts is being studied in various application domains. Park et al. (2010) developed a classifier based on a set of linguistic features to detect sentences that discuss human activities from Weblogs. Liu et al. (2015) developed a similar classifier to extract sentences that express experience in online health forums (OSG) using linguistic features such as pronouns, verb tense and modality of the verb.

(Nguyen, 2012) has employed both bag-of-words (BoW) and linguistic features to detect experience expressing posts and identified that the most discriminative BoW features are 'I' and 'my 'and most discriminative linguistic feature is *first person pronouns* ('I' and 'my '). This is because users frequently use *first person pronouns* in expression of experience.

In this work, the narrative type is determined based on the main subject of each OSG post. The key intuition for this approach is that different types of nouns and pronouns are used as the subject in different narrative types. For example, the first column of Table 5.2 shows three sample posts taken from a OSG. First post is a patient expressing his experience, and the main subject is a first-person pronoun. In the second post, a son is expressing about his mother, so the main subjects are mother and third person pronoun

she (refers to mother). The last post is a piece of advice given to someone else, so the main subject is the second-person pronoun you.

Based on the above observations, a set of rules are designed to categorise OSG posts into the three narrative types based on the main subject of each OSG post as follows:

1. **experience first-person:** if the main subject is a first-person pronoun (e.g., I, my)
2. **experience second-person:** if the main subject represents a relationship (e.g., mother, friend, son) to the narrator of the OSG post.
3. **advice:** if the main subject is a second-person pronoun (e.g., you).

The key challenge of this approach is that OSG posts sometimes contain several nouns and pronouns as the subject of different sentences, for example, patients expressing their experience might mention their family history using words such as mother and third-person pronouns like her. Moreover, third-person pronouns are often used to mentions other people as well (e.g., I went to see GP and he prescribed ...).

Table 5.2: Examples for narrative type resolution from OSG posts

Post	Pronoun resolved post	Human mention nouns	Narrative type
My doctor ... yesterday. He ... daily. I took it ... and ... I felt ... Is this normal ...? I was ... because ... I haven't ... and I was ... with my fiance. I got paranoid about her ... even though I knew ... she ...	My doctor ... yesterday. <doctor> ... daily. I took it ... and ... I felt ... Is this normal ... ? I was ... because ... I haven't ... and I was ... my fiance. I got paranoid about <fiance> ... even though I knew ... <fiance> ...	I: 7 doctor: 2 fiance: 2 prominent noun: 'I'	experience: first-person
My mother is diagnosed ... She is suffering with ... I know that ... her heart. An endocrinologist suggested her to ... Also, he recommended her to ... But now she had ... She is suffering ...	My mother is diagnosed ... <mother> is suffering with ... I know ... <mother> heart. An endocrinologist suggested <mother> to .. . Also , <endocrinologist> recommended <mother> to ... But now <mother> had ... <mother> is suffering ...	mother: 5 endocrinologist: 2 I: 1 prominent noun: 'mother'	experience: second-person
You should consider ... such ... . You may even want to ... . Try to ..., but if you've ... your other options you may ...	You should consider ... such ... . You may even want to ... . Try to ..., but if you've ... your other options you may ...	you: 5 prominent noun: 'you'	advice

In order to overcome these ambiguities, initially, the gender specific third person pronouns in OSG posts are replaced with relevant nouns (e.g., she →mother). For this task, we employed a pronominal anaphora resolution (Mitkov, 2014) algorithm which predicts the relevant antecedents of third person pronouns.

Pronominal anaphora resolution is initially attempted using JavaRAP tool (Qiu et al., 2004) that implements an algorithm presented by Lappin and Leass (1994). However, this approach relies on parsing text to identify its structural elements and grammatical roles,

which is found to be less effective for OSG posts because as discussed in Section 2.7 they are not written conforming to the formal rules of grammar. Hence, a relatively simple rule based pronoun resolution algorithm CogNIAC (Baldwin, 1997) is selected. CogNIAC uses a set of rules to assign the most probable antecedent to each pronoun.

For this task, a partial implementation of CogNIAC is used to resolve pronouns to their antecedent. Note that the pronoun 'it' is ignored as it is irrelevant to the given task. The second column of Table 5.2 shows the pronoun resolved output where each pronoun is replaced with the relevant noun.

Once the pronouns are resolved it is easier to determine the key narrative type of the OSG post. For example, pronoun resolution in the second OSG post specified in Table 5.2 has aggregated 'mother'and 'she', which makes the subject 'mother'more evident, and that the narrative of that post is *experience: second-person*. Column three of Table 5.2 shows a list of human subjects identified from the sample posts and column four shows the predicted narrative type.

Note that the pronoun resolved output of the OSG posts are further used in Section 5.5.3 for determining the gender of the patient.

5.5.2 Age extraction

Age is an important feature for patient profiling because medical information such as symptoms are interpreted differently for different age groups. Patients often mention their age within the text of the forum post. However, such mentions are highly diverse, unstructured and often expressed using shorten forms of English terms. Some sample age mentions are shown in column 1 of Table 5.3. Also, they need to be differentiated from other number mentions such as duration, dose, weight etc.

Table 5.3: Positive and negative samples of age related phrases found in forum posts

Age Phrases (positive examples)	Non Age Phrases (negative examples)
when i was in my 20s	My SVT started about 10 years ago
In my 41 years of life	I had a 48 hr monitor
I am 42 yrs old	I am using atenolol 50 mg
my 25 year old son	doc prescribes 10 tablets 1 twice a day
operated on at 10 days old	Currently I weigh 16 3/4 stone

Previous literature on age extraction from forums is limited mainly due to its challenging nature. Liu et al. (2014) focused on standard age mentions such as 'I am 35 years old' and identify them as age mentions. Kim et al. (2013) employed regular expressions such as age+number to extract age. These approaches achieve higher precision but as shown in Table 5.3, age mentions are diverse thus these approaches would overlook a significant amount of age mentions resulting in a low recall.

Zhu et al. (2012) look for clue words such as years, old, aged etc. appear within two-word distance of each two-digit number mentions and such mentions are extracted as the age of the patient. In addition, they look for clue words such as teenager, toddler, child to approximate the age of the patient. This approach would result in high recall however,

can lead to a significant number of misclassifications (e.g., about 10 years ago can be misclassified as an age mention).

A common issue of these approaches is that they do not resolve whether the age mentions is about the patient. For example, a patient might mention the age of relatives when talking about their family history. Also, there can be age mentions in past incidents of the patient (e.g., I have this issue since 11 years old). Moreover, there can be multiple age mentions in a forum post. Thereby, it is necessary to develop an extraction method to resolve age from multiple age mentions.

In order to overcome the above mentioned issues and also to improve both precision and recall we employed a multi-step process shown in Figure 5.5 to resolve the age of each user in OSG. Each phase of this process is elaborated below.

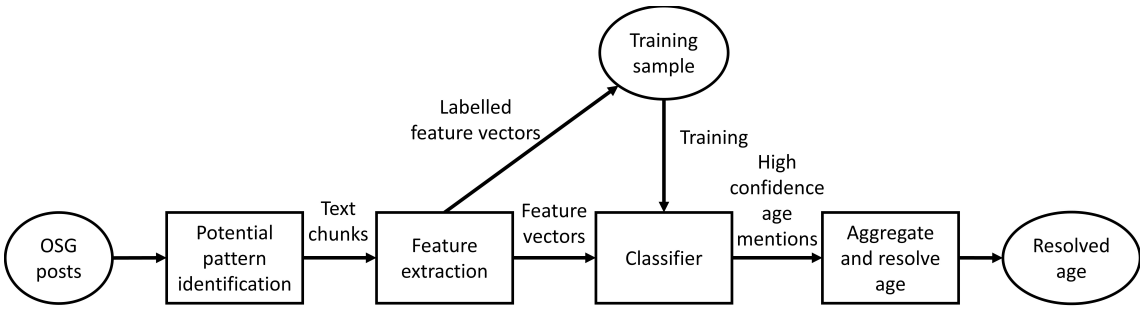


Figure 5.5: Age resolution process

Potential pattern identification

In this initial step, a regular expression based pattern in used to identify text chunks that are potential age mentions.

Firstly, each post is break into sentences using the sentence splitter in OpenNLP tool (Baldridge, 2005). Subsequently, the *regular expression*: `(\d{1,2}[a-zA-Z]{0,5})` is used to capture two-digit number mentioned in the text. Note that, the above *regular expression* is designed to accommodate numbers that co-exist with several letters such as 20s and 34yrs.

A text chunk of five words before the detected two-digit number and five words after are extracted as a possible age mention. Note that start and end of sentences are padded with `<S>` symbols to accommodate age mentions in edges of the sentence.

Let a set of OSG posts of a user A be $\{p\}^A$. For each post p, a set of text chunks $\{t\}$ with possible age mentions were identified $\{p\}^A \leftarrow \{t\}^A$.

Feature extraction

In this step the text chunks extracted from the previous step are transformed in to feature vectors ($\{t\}^A \leftarrow \{f\}^A$). These features are extracted to differentiate age mentioned from others (e.g., drug dose, clinical factors).

As shown in Figure 5.6, the text chunk is first divided into three segments (L, M, and R) and different features were extracted separately from each segment.

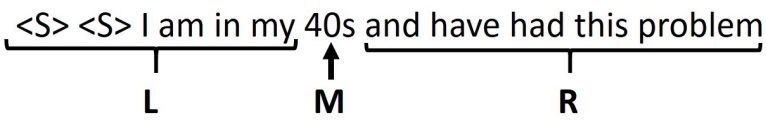


Figure 5.6: A text chunk and its segments L, M, and R

We engineered 29 features that are capable of differentiating age related text chunks from others. Table 5.4 presents a sample of the extracted features. Note that, the features that are extracted from multiple segments are included in the feature vector as multiple features (one for each segment).

Table 5.4: A Sample of features extracted from a text chunk

Feature	Applied segments	Representing terms
First-person pronouns	L	I, Im, Iam
Possessive pronouns	L	my, our
Family relationship mentions	L,R	mother, mom, father, dad, brother, sister, son, daughter
State of being verbs	L	am, is was, were, will be
Age related prepositions	L	at, until, around, under, about, abt
Old Mentions	M, R	old, yo, y/o
Year Mentions	M, R	year, years, yo, yrs, y/o, s
Time mentions	M, R	days, hours, hrs, minutes, min, weeks, wks
Dose mentions	M, R	mg, dose, tablets, mgs, micrograms, ug

Most of these features are constructed to capture age mentions. For example *First-person pronouns* are often used before an age mention. Some other features such as dose and time mentions are added to differentiate other key classes such as dosage (e.g., atenolol 50 mg) and duration (e.g., 10 years ago).

Classifier

Once the features were extracted from text chunks, we employed a classifier to identify the chunks that are most likely to be age mentions and assigned a confidence probability. To train this classifier we manually labelled 2,212 text chunks identified from the first step of this process. We labelled them as 'age' or 'other' based on whether it is an actual age mention or not. In the labelled dataset there are 1,186 'age' and 1,026 'other' sample.

The usual techniques for training a predictive classifier are Naïve Bayes, Support Vector Machines (SVM), Random Forest, and Multilayer Perceptron (MLP). We have excluded MLP as the dataset is small which could lead to overfit MLP models. Naïve Bayes, SVM and Random Forest classifiers are trained using the feature vectors of the above labelled dataset. The classifier implementations in WEKA (Hall et al., 2009) data mining toolkit are employed for this task. Each classifier is evaluated using 5-fold cross validation and

Naïve Bayes, SVM and Random Forest result averaged f-measures of 0.80, 0.83 and 0.85 respectively. Therefore, Random Forest is selected as the classifier for this task.

Classifier determines a confidence value ($c(t)$) for each text chunk to be an age mention and based on that text chunks that have higher confidence of being an age mention are selected for each author A.

$$\{t_a\}^A = \{t_a : t \in \{t\}^A \textbf{ and } c(t) > \tau\} \tag{5.1}$$

Where τ is a confidence threshold which is set to be 0.7.

Aggregate and resolve age

This step aims to resolve the age of each individual author from the selected high confidence age mention text chunks $\{t_a\}^A$ from the previous step. Note that the numerical value of age mentions might be different from the actual age value due to (i) age mentions of a past event, (ii) age mention of a different person, and (iii) misclassified text chunk. Hence, a weighted majority voting technique is employed to resolve the most probable age of each user A.

First, the numerical value of each age mention ($D(t_a)$) is determined. Note that age mentions are parsed to get the age value normalised to year (e.g., 6 months $\rightarrow$ 0.5 year). Also, two other parameters are derived from each age mention as follows: (i) tense of age mention (present, other), and (ii) subject type of the age mention (first-person, other).

Tense of the age mention is derived based on the tense of the verb in the text chunk. The text chunks with present tense verbs are marked as *present*, and other tenses or if the tense is not available it is marked as *other*). Subject type of the age mention is determined based on the subject in the chunk. The text chunks with first-person subjects are marked as *first-person*, and other subject types or if the subject is not available it is marked as *other*).

The rationale of these two parameters is that if the chunk is in the present tense then it is more likely to be the current age of author. Also, if the subject is first-person then it is more likely to be about the author. Therefore, a weight boost of 0.25 is added if the age mention is present tense and a weight boost of 0.25 boost is added if it has a first-person subject. The aggregated confidence value $C(a)$ determined as follows:

$$C(a) = \sum^{\{t_a\}^A} I(a \leq D(t_a) \leq a + \Delta a) \times 2^{-(D(t_a)-a)} \times W(t_a) \tag{5.2}$$

where I is the indicator function and the weight $W(t_a)$ is determined as $1 + 0.25 \times I(t_a \text{ present tense}) + 0.25 \times I(t_a \text{ first-person})$.

Equation 5.2 is formulated in a way that an age value $D(t_a)$ contributes to the confidence of a to $a + \Delta a$. This is because users often contributes to a particular OSG over a period of time. Hence, an age mention of a user can be different in older posts. For example, a user might mention age to be '56' two years ago and now mention that he is '58' . In such cases, age mention of '56' should contribute to the age value of 58. Note

that Δa is set to 2 to hypothesising that a user often contributes to OSG for a period of 2 years.

The aggregated confidence value is determined for all the age values $\{a\}^A = \{D(t_a) \forall t_a \in \{t_a\}^A\}$ mentioned for the user A, and the age value of the user is resolved based on the highest aggregated confidence value:

$$a_{resolved} = arg \max_{\{a\}^A}(C(a)) \tag{5.3}$$

5.5.3 Gender extraction

Gender is another important demographic information that is important for personalised retrieval. Some symptoms can relate to different diagnoses depending on whether the patient is a male or a female. Gender is sometimes explicitly self-disclosed (e.g., 'I'm a female'), and in some cases gender can be inferred based on gender mentions (e.g., 'my mother') or gender specific medical term mentions (e.g., 'pregnant'). However, similar to age extraction, this task is challenging due to the unstructured and diverse nature of gender mentions in OSG posts.

Cheng et al. (2011) focused on the language style of the text to predict the gender of the author. The idea is that males and females often follow different language styles for written communication. However, the accuracy of this approach will be low as OSG attract information seekers and providers with very diverse language styles from across the world. Another approach is to predict the gender by looking at the first name of the author (Herdadelen and Baroni, 2011). It keeps two lists of male and female first names and resolves the gender based on that. This approach does not work in our scenario as many people do not use their real name in OSG, they often use a nickname or a part of their name.

Zhu et al. (2012) look for gender clues: (i) gender specifying words such as 'men', 'woman', etc. and (ii) gender specific medical terms such as 'hot flashes', 'prostate cancer', etc. It was not mentioned that how they resolve the gender if terms related to both genders appear in posts of the patient. Also, another key issue is to resolve whether the gender clues are about that patient.

We extended the above approach of using gender clues to resolve gender, however, our technique is designed to be more robust to noisy gender cues, and also capable of fusing multiple gender cue to resolve the most probable gender of each author. Similar to (Zhu et al., 2012) this technique is two fold, (i) the first approach extracts gender specific narrations and processed them to resolve the gender, and (ii) the second approach looks for gender specific medical terms (body-parts, illnesses, symptoms). The results from both approaches are fused using a weighted majority voting method to resolve the highest confident gender of each author. This technique is presented in Figure 5.7 and described in detail below.

As shown in Figure 5.7, first the posts which are marked as advice by the narrative type identification are filtered out. This step is taken because advice posts often contain gender cues that are related to the receiver of the advice rather than the author of the

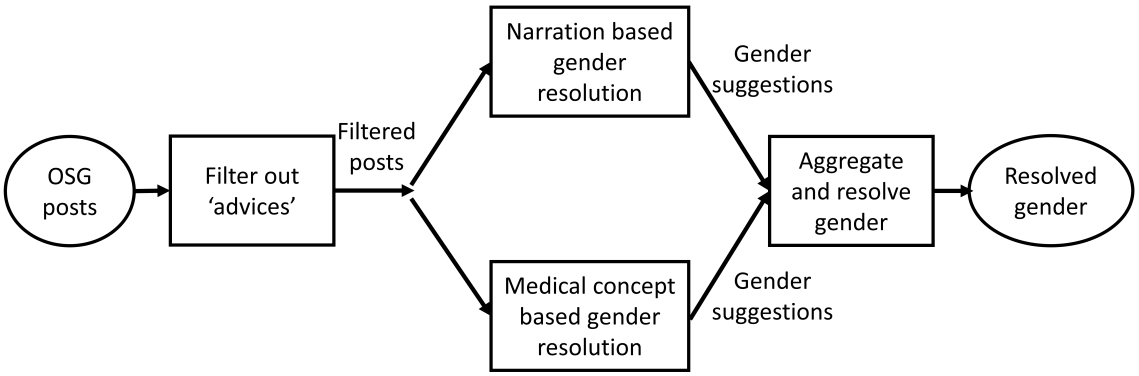


Figure 5.7: Gender resolution process

post. Hence, gender cues in advice posts can be misleading as they often not related to the author of the post.

Narration based gender resolution

The gender is often explicitly or implicitly mentioned in the narration of OSG posts. However, gender specified in narrations can be ambiguousness since such terms may be referring to some other actor other than the patient. The proposed technique is developed to overcome such ambiguities.

Firstly, the narration type is identified based on the technique mentioned in Section 5.5.1. If the narration type is a *first-person* narration (e.g., 'I', 'me'), then the narrator is the actual patient. Therefore, the narrator gender is resolved and assigned to the patient. The gender of the narrator is resolved using two types of gender mentions found in OSG posts:

1. Direct gender mentions: directly reveals the gender of the narrator (e.g., 'a male', 'a man', 'a mother', 'a widow'). Those mentions have to exist within close proximity of first-person pronoun i.e., 'I' in a sentence to verify that it is about the patient.

2. Indirect gender mentions: narrator gender can be inferred from gender specific relationships. For example, if OSG posts talk about 'my wife' or 'my fiance' that implies that the narrator is a male.

If the narration type is a *second-person* narration, then the narrator is a caregiver (e.g., partner, parent, child, friend) who is posting on behalf of the patient. In such cases, the relationship of patient to the narrator can be used to resolve the gender most of the time. For example, patient is a male if 'he' is husband, son, uncle etc. of the narrator; and female if 'she' is daughter, wife, mother, etc. of the narrator.

However, sometimes patients relationship to the narrator is gender-neutral (e.g., partner, friend) then the pronoun resolution process that is used to identify the narration type in Section 5.5.1 is examined to check if gender specific pronouns were resolved to that noun (e.g., he $\rightarrow$ friend) and gender is assigned accordingly. Table 5.5 shows three examples of the narration based gender resolution process.

Table 5.5: Few examples of narration based gender resolution process

Post	Pronoun resolved post	Narrative type	Inference of gender
My doctor ... yesterday. He ... daily. I took it ... and ... I felt ... Is this normal ...? I was ... because ... I haven't ... and I was ... with my fiance. I got paranoid about her ... even though I knew ... she ...	My doctor ... yesterday. <doctor> ... daily. I took it ... and ... I felt ... Is this normal ... ? I was ... because ... I haven't ... and I was ... my fiance. I got paranoid about <fiance> ... even though I knew ... <fiance> ...	first-person	Gender cue: *fiance* Patient: **male**
My mother is diagnosed ... She is suffering with ... I know that ... her heart. An endocrinologist suggested her to ... Also, he recommended her to ... But now she had ... She is suffering ...	My mother is diagnosed ... <mother> is suffering with ... I know ... <mother> heart. An endocrinologist suggested <mother> to .. . Also , <endocrinologist> recommended <mother> to ... But now <mother> had ... <mother> is suffering ...	second-person	Gender cue: *mother* Patient: **female**
A friend ... when he .... He suddenly ... hurt his back. He fell down .... Despite ... him to his feet and made him walk to ... away.	A friend ... when <friend> .... <friend> suddenly ... hurt <friend> back. <friend> fell down .... Despite ... <friend> to <friend> feet and made <friend> walk to ... away.	second-person	Gender cue: he (resolved to friend) Patient: **male**

Medical concept based gender resolution

There are medical concepts (e.g., body parts, symptoms, illnesses, procedures) that are often limited to a single gender (e.g., prostate cancer, pregnancy), which can be used to infer the gender of the patient. In this approach it is hypothesis that discussion of such a medical concept indicates that the patient is of that respective gender. For this approach it is immaterial to know the narration type, because the medical concept is about the patient, hence this approach is relatively straightforward than the narration based approach.

A list of seed medical terms (words or phrases) were constructed for both genders. These seeds lists were populated based on the gender related concept terms from UMLS Metathesaurus (Bodenreider, 2004). However, patients sometimes use lay terms for those medical concepts, which are not included in clinical ontologies such as UMLS Metathesaurus. Therefore, a set of samples from mens and womens sections of OSG were examined to include such lay terms of medical concepts that are deterministic of gender.

If the above formulated gender specific terms are present in OSG posts, relevant gender is assigned to the patient that is mentioned in that post.

Aggregate and resolve gender

Gender suggestions based on each post are aggregated to resolve the patient gender of each OSG user. The aggregation process is similar to the age resolution process specified in Section 5.5.2, where medical concept based gender suggestion is given a weight of 1.0 and narration based gender suggestion is given a weight of 2.0. Note that narration based

gender suggestions are given a higher weight as they are relatively more accurate than medical concept based. Once aggregated, gender of the patient is resolved based on the highest weighted gender, which is then re-assigned to all the posts of that author.

5.5.4 Medical concept extraction

Medical concept extraction from text is a further formidable task. Natural language processing (NLP) tools are used to extract terms from text that can be mapped to the medical concepts found in medical thesauruses such as the UMLS Metathesaurus (Bodenreider, 2004). There are several state-of-the-art tools that extracts medical concepts from text such as: MedLEE (Friedman et al., 1994), cTAKES (Savova et al., 2010), and MetaMap (Aronson and Lang, 2010).

Gupta et al. (2014), shows that better precision and recall can be achieved by developing a tool to support special characteristics of patient-authored text. Consumer health vocabularies extracted from community generated corpora (Vydiswaran et al., 2014) are employed to identify the relevant medical terms and a different stack to NLP processes is followed to extract key phrases from the text. However, as noted by the authors, the existing system is limited to certain subcategories of the OSG (Asthma and ENT). Therefore, for our task we adhere to the well-established medical concept extraction tools.

We employed cTAKES (Savova et al., 2010) tool for our medical concept extraction task. It identifies noun phrases in the text and then conducts a dictionary-look up in SNOMED CT and RxNORM (Nelson et al., 2011) medical concept databases. Each identified term is then mapped into five semantic types: disorder/disease, sign/symptoms, procedures, anatomical sites, and medications. In the proposed method, we subject each OSG post to this process and the identified medical concepts are extracted. As pointed-out by Gupta et al. (2014), some ambiguities exist in the identified concepts. For example, today is mapped to a drug named Today and web is mapped to the disorder congenital webbing. However, both these words are found frequently in OSG posts mostly referring to their usual meanings. To overcome this issue, we constructed a list of terms that are often mapped incorrectly and filtered out the identified concepts based on such terms.

5.5.5 Emotion extraction

As discussed in Section 2.4 emotions are an important element of social text which self-disclose the mental state of the author. In the context of OSG discussions, it can be hypothesised that the emotions expressed in OSG discussions are indicative of the mental state of OSG users (patents/caregivers), which can be used to infer their emotional well-being. Moreover, since the OSG user posts at different points of time, emotional fluctuations over time is indicative of their emotional journey during their participation in OSG.

Analysing emotions and emotion fluctuations over time of a OSG user is important for the medical researchers to understand the emotional well-being of patients/caregivers with a particular disease conditions, as well as to compare and contrast the emotional well-being

of patient cohorts with different disease conditions or other clinical factors (e.g., undergone different treatments).

As discussed in Section 2.4.2, multiple psychological emotional models are proposed to represent the emotional state, which ranges from the two dimensional valence-arousal model (Russell, 1980) to multi-dimensional models such as emotion wheel (Plutchik, 1991). While such models serve as theoretical basis for emotion representation, computational implementations are required to capture the emotional expression from discourse. For example, sentiment analysis techniques are the computational implementation of the valence-arousal model (Mohammad, 2016), which provide a signed real-value as the sentiment score, where the sign (positive/negative) represents the valence and the absolute value of score represents arousal.

Although sentiment analysis techniques are relatively mature and commonly used for capturing emotions, we found that the two dimensional model is too coarse grained for representing complex emotional states of OSG users. Therefore, we have developed a computational technique based on the Emotion Wheel (Plutchik, 1980b, 1991) to capture a multi-dimensional representation of emotions from OSG discourse.

As discussed in Section 2.4.2 Emotion Wheel has eight primary emotions (joy, trust, surprise, sadness, disgust, anger, anticipation and fear) and further eight secondary emotions which are derived using combinations of primary emotions (e.g., love: joy+ trust). These 16 emotions (primary and secondary) specified in the Emotion Wheel is incorporated as the emotional dimensions in the proposed computational model. Note that several modifications are applied to secondary emotions to suit the healthcare domain. The emotional intensity of each emotion is determined based on the proportion of relevant emotional terms present in each OSG post resulting a 16-dimensional real-valued emotion vector for each OSG post.

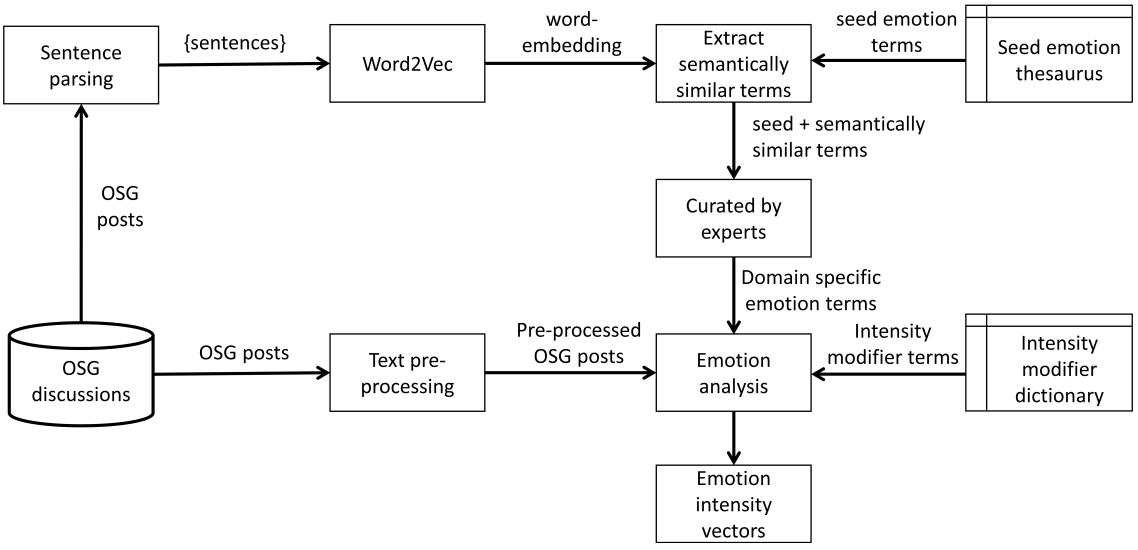


Figure 5.8: The proposed technique for emotion extraction

Figure 5.8 presents the proposed technique for emotion extraction. The relevant terms for each emotion is obtained using a two step process.

First, a seed emotion term thesaurus is constructed for each emotion. We constructed this thesaurus based on a list of feeling words used for mental status exams (Niolon, n.d.), which contains emotional terms for each of the 16 emotions. In Table 5.6 the second column provides the 16 emotions and the third column provides a sample of seed emotions terms employed for each category.

Table 5.6: Emotion categories and a sample of representative terms used for each emotion.

	Emotion	**Emotional terms from thesaurus**	**Emotional terms from word-embedding**
Positive	Happy	happy, great, joyous, glad, delighted	fab, chuffed, terrific, great news, looking forward, heart warming, uplifting, upbeat
	Good	good, pleased, comfortable, relaxed, content	comfy, nice, chill, chipper, ok, okay, clear headed, cool
	Alive	alive, playful, energetic, spirited, animated	chatty, perky, sociable, vibrant, vivacious, witty, easy going, peppy
	Love	love, attracted, warm, passionate, affectionate	romantic, cuddly, compassionate,intimate, adore, supportive, caring
	Positive	positive, eager, keen, bold, brave	smart, ambitious, proactive, cynical, insistent, willing, upbeat
	Open	open, understanding, accepting, satisfied, receptive	open minded, empathetic, cooperative, accommodating, approachable, forgiving, attuned, rational
	Interested	interested, fascinated, inquisitive, curious, intrigued	keen, impressed, cautious, leery, eager, intuitive, savvy, thoughtful
	Strong	strong, certain, dynamic, sure, tenacious	resilient, independent, adamant, fierce, self reliant, decisive, fighter, pragmatic
Negative	Sad	sad, tearful, grief, sorrowful, grief	heart break, teary, lonely, weepy, crying, despairing, hurtful
	Afraid	afraid, fearful, terrified, panic, worry	petrified, freaking out, apprehensive, dread, obsess, fret, nervous wreck
	Hurt	hurt, deprived, pained, dejected, agonised	traumatised, bruised, shattered, ached, exhausted, cramped, numb, fatigued, strained
	Angry	angry, annoy, provoke, aggressive, enraged	agitated, hostile, pissed off, argumentative, aggressive, rude, paranoid, ticked off, lashing out
	Depressed	depressed, disappointed, miserable, despair, powerless	despondent, distraught, suicidal, unloved, worthless, emotionally drained, snappy
	Helpless	helpless, incapable, alone, vulnerable, fatigued	insecure, tired, hopeless, powerless, defeated, overwhelmed, listless, incapacitated
	Confused	confused, upset, doubtful, uncertain, hesitant	unsure, perplexed, wary, leery freaked out, iffy, bummed, taken aback
	Indifferent	indifferent, insensitive, dull, reserved, lifeless	grumpy, apathetic, blunt, ignorant, emotionless, callous, crass, standoffish

Expanding a seed list of lexicons is a labours task which is often achieved using crowd-sourcing techniques such as Amazon Mechanical Turk (Buhrmester et al., 2011). However, recent research report techniques (Hamilton et al., 2016; Fast et al., 2016) to expand seed term list using a semi-supervised approach based on the word-embedding (Mikolov, Sutskever, Chen, Corrado and Dean, 2013) technique. Word-embeddings learn dense vector representations of words and phrases while automatically preserving the semantic relationships that exist in the text corpus by incorporating such relations into the vector space of the word-embedding. This enables the use of linear algebra to capture different semantic relationships within word-vectors in the word-embedding. The famous example by Mikolov, Yih and Zweig (2013) shows that the vector arithmetic of word vectors 'King -Man + Queen' results a word vector similar to the word vector of 'Woman'.

Developing such word-embedding using an OSG post corpus enables to capture terms used by the OSG users that are semantically similar to the seed emotional terms. We have developed a word-embedding from a large text corpus collected from two large OSG *patient.info* and *healingwell* mentioned in Table 5.1, which contained a total of 4,795,428 OSG posts. This corpus is preprocessed to remove any URLs, converted to lower case and then separated into sentences using the python NLTK sentence tokenizer [3], which has resulted a text corpus of 36,222,536 sentences. This text corpus is employed to train a 200 dimensional word-embedding using Word2Vec technique with skip-gram model (Mikolov, Sutskever, Chen, Corrado and Dean, 2013) and negative-sampling (Mnih and Teh, 2012). Note that common phrases are tagged up to three terms and used them as phrases during the training of the word-embedding. We employed the python Gensim (Řehůřek and Sojka, 2010) library for this implementation. The resulted word-embedding contains 312,196 unique terms (words or phrases).

Once the word-embedding is trained, most similar terms for each seed terms in the emotion thesaurus is identified using a nearest neighbour search in the embedding space using Cosine similarity. These identified terms are semantically similar terms to the seed emotion terms, in which some of the terms have the same emotional sense of the seed term while some others may not. For example, the top five nearest neighbours of *sorrowful* are *sadness, sincerity, joyful*, and *deeply saddened*, in which *joyful* is semantically similar but has the opposite emotional sense. Therefore, we manually looked at the selected terms and filtered out the terms that do not align with the emotional sense of the respective seed term. The third column of Table 5.6 presents a sample of emotional terms captured using the above technique. Note that, this emotions term identification is a one time task and the final emotion terms list is added to the emotion term thesaurus.

Intensity modifier terms are a set of terms that increase or decrease the intensity of the emotional term. For example, the term 'very' increases the intensity of the emotion 'good' when used together, whereas, the term 'kind of' decreases the intensity of the emotion 'okay' when used together. Moreover, some terms completely negate the emotions e.g., 'not okay' negates the emotion expressed by 'okay'. A thesaurus of such terms is often used in rule based sentiment analysis tools such as SentiStrength (Thelwall et al., 2012)

[3] http://www.nltk.org/api/nltk.tokenize.html

and VADER (Hutto and Gilbert, 2014) to improve the accuracy of the sentiment score. In this work, we have employed the intensity modifier term thesaurus used in VADER (Hutto and Gilbert, 2014).

When assessing the emotional expression of an OSG post, first it is sent through a preprocessing step which converts text lower case and removes URLs, alphanumeric characters, and punctuations. Stop words are not removed as stop words may contribute to the emotional expression. It is then tokenised into words. The pre-processed text is then used for the emotion analysis, which identifies the relevant emotion terms for each emotion and any intensity modifier terms that are associated with the emotion terms. Algorithm 1 presents the pseudo-code for this emotions vector E_P calculation of a particular OSG post P.

Algorithm 1: Determine emotion intensity vector

Input: P- OSG post, $\mathcal{T}_E$- Emotion thesaurus, $\mathcal{T}_m$- Intensity modifier thesaurus
Output: $E(P)$ - Emotion vector of OSG post P

```
E(P) ← ∅
P ← Preprocess(P)                    // remove URLs, numbers, punctuations
{w}_P ← Tokenize(P)                  // tokenise the post into words

for i ← 1 to 16 do
    T_Ei ← T_E
    e_i ← 0
    w_prev ← null
    foreach w in {w}_P do
        // lookup w in emotion i term list
        if w ∈ T_Ei then
            e_i ← e_i + 1

            // lookup w_prev in intensity modifier thesaurus
            if w_prev ≠ null and w_prev ∈ T_m then
                e_i ← e_i + T_m(w_prev)
        w_prev ← w
    e_i ← e_i ÷ |{w}_P|
    E(P) ← E(P) ∪ e_i
return E(P)
```

5.6 OSG user timeline construction

This section describes the automatic construction of user timelines from their self-disclosed narratives in OSG posts. A timeline consists of time sensitive events organised in chronological order. Apart from demographics which are not time sensitive, all other information extracted in the previous section can be employed to the construction of the user timeline. However, this work is limited to the timeline construction based on clinical events and emotional events (emotional trends) only.

A timeline of events helps to examine the user behaviour over time as well as temporal associations between multiple events. For example, such a timeline is an important resource for a health professional to identify the temporal ordering of clinical events over the disease progression of a patient. Such insights can be employed during prognosis as well as therapy planning (Augusto, 2005). Also, the medical researchers can use such information to investigate disease progression of certain illnesses across different cohorts of the patients in the population. Having emotions alongside clinical events enables the medical researchers to analyse the associations between clinical events and emotions, and find insights on the emotional burden of a disease condition over time.

5.6.1 Clinical event timeline

Clinical event is defined in literature as a clinically relevant symptom, state, perception, procedure or occurrence (Tao et al., 2010a,b). Clinical event extraction is a specified event extraction task as the event related information such as relevant key terms is priorly known. As discussed in Section 2.3.2, specified event detection uses an engineered feature dictionary or labelled datasets to train a classifier to identify documents/texts that contain relevant event information, and then capture the relevant named entities (such as event time) that are related to the identified event.

Existing work on clinical event extraction is mostly developed for the clinical narratives in Electronic Health Records (EHR) and discharge summaries. There is a recent interest on such techniques due to the availability of several human annotated datasets such as the dataset released for i2ib challenge (Sun et al., 2013) and SemEval-2015 Task-6 (Bethard et al., 2015). Tao et al. (2010b) developed CNTRO (Clinical Narrative Temporal Relation Ontology) which uses an ontology based approach to model time sensitive information in clinical narratives. For each identified event it extracts temporal information such as time and duration by parsing the time indicative terms/phrases (e.g., '2 week 'for duration of two weeks) using time ontologies (Hobbs and Pan, 2006).

These existing techniques are specifically designed for the narratives generated by health professionals, where such narratives often adhere to certain guidelines and professional practices. Therefore, vocabulary, reporting of clinical events and expression of temporarily is fairly uniform across those narratives. On the other hand, OSG content is user generated (patient/caregiver). As discussed in Section 2.7 such user generated content is so diverse and contain different language patterns prevalent among different socio-geographic groups (Eisenstein et al., 2014). Therefore, the existing techniques need to be extended to facilitate timeline extraction from OSG.

Wen and Rose (2012) have developed such an extended technique to construct clinical timeline i.e., cancer trajectories from the OSG discussions in breast cancer support groups. They have engineered a consumer health vocabulary of breast cancer related keywords which can be phrases or certain abbreviations (e.g., diagnosed: dx, dx'd). From the OSG posts, the sentences with the keywords in the above developed vocabulary are extracted and further parsed to extract the temporal expression related to the time of that event. Time is inferred using both specific time expressions (e.g., 4th May 2018) and relative

time expressions (e.g., a week ago). The relative time expressions are converted to a specific time with the support of the post date of the OSG post (e.g., 'a week ago'→post date - 7 days). This approach is further extended by Naik et al. (2017) with a semi-automated approach for engineering relevant keyword dictionaries and also have employed HEidleTime (Strötgen and Gertz, 2010) for time resolution.

We employed Wen and Rose (2012) approach for clinical event extraction from OSG. The medical concept extraction discussed in Section 5.5.4 already extracts clinical concepts such as symptoms and procedures based on the terms found in medical thesauruses. However, as discussed in Section 5.5.4 such thesauruses are optimised for the vocabulary in clinical discourse used by health professionals, therefore often lack the terms used by general health consumers i.e., patient or caregiver. This issue highlights the need for a consumer health vocabulary to more effectively extract clinical concepts from consumer generated discourse such as OSG.

Developing a generic consumer health vocabulary is a laborious task which is beyond the scope of this work. However, we have developed a consumer health vocabulary for several key cancer related clinical concepts, in order to demonstrate its utility as well as for the further application of this work in cancer domain (presented in Chapter 6). We have selected five key clinical events of a cancer patient as follows:

1. diagnosis
2. medical tests (e.g., biopsies, other pathology tests)
3. surgery
4. radiotherapy
5. recurrence

A consumer health vocabulary was engineered based on the terms (words/phrases) used by the users in OSG that are indicative of the above designated event types. In order to capture such terms from a very large text corpus, we employed the same approach used in Section 5.5.5 for expanding emotion thesaurus, which is a semi-supervised approach based on the word-embedding (Mikolov, Sutskever, Chen, Corrado and Dean, 2013) technique. Similar to the emotion term capturing, such word-embedding based on OSG post corpus enables to capture terms used by the OSG users that are semantically similar to the clinical terms that are indicative of the above designated event types.

We employed the same word-embedding developed in Section 5.5.5 and looked up semantically similar terms for a given term in the embedding space using Cosine similarity. For example, the ten most similar terms of 'radiotherapy' ordered based on Cosine similarity score are 'radiation', 'chemotherapy', 'hormone_therapy', 'radio_therapy', 'salvage_radiation', 'external_beam_radiation', 'chemo', 'brachytherapy', 'external_radiation', 'brachy' and 'ebrt'. radiation and radio therapy are synonyms while salvage radiation, external beam radiation and brachytherapy are different variants of radiotherapy. EBRT is a abbreviation for external beam radiation therapy, while chemo and brachy are shortened forms for chemotherapy and brachytherapy respectively.

The inclusion of the phrase 'hormone therapy' as a similar term for 'radiotherapy' is interesting as it is neither a synonym nor a variant, but just used in a semantically similar manner in OSG discussions. This inclusion highlights the fact that the candidate terms selected using word-embeddings need to further validated for their meaning by an expert. Therefore, we employed the support of a cancer surgeon to further curate the automatically generated term sets of each event category. Table 5.7 shows ten term samples of this curated consumer health vocabulary for each event category delineated above.

Table 5.7: A sample of representative terms of each clinical event type. Note that some of the terms are common misspelled words.

Clinical event	Representative terms
Diagnosis	diagnosed, dxed, diagonsed, diagnosedwith, diag, dx, clinically diagnosed, officially diagnosed
Medical tests	biopsy, biopsies, colonscopy, biop, fna, endoscopy, scope, mammogram, mpmri, cystoscopy
Surgery	surgery, operation, op, sugery, surgury, opp, corrective surgery, ops, surgical procedure, opperation, keyhole surgery, bunion surgery
Radiotherapy	radiotherapy, radiation, chemotherapy, radio therapy, salvage radiation, external beam radiation, chemo, brachytherapy, external radiation, brachy
Recurrence	recurrence, reoccurrence, recurrance, bcr, biochemical recurrence, biochemical failure, local recurrence, biochemical relapse, secondary cancers

The identified event sentences are then further processed to identify any temporal expression mentioned in the free-text form. We employed SUTime (Chang and Manning, 2012) to detect such temporal expressions. SUTime is a rule based classifier which contains a set of hand-crafted rules and dictionary of time sensitive terms. It first identifies the time related terms and then extend them to chunks and apply the rules to resolve the time. Note that we extended the time sensitive term list of SUTime by including some of the relevant terms identified based on the above created word-embedding. For example, 'weeks' is often referred to using terms such as 'wks', 'wk', and 'months' as 'mnths'. Such additions improve the recall of temporal expression. SUTime resolves both specific time expressions and relative time expressions (e.g., '2 wks ago' →'OFFSET P-2W'). The relative time expression is then resolved based on the post date of the OSG post.

OSG users, often mention the same event multiple times in different OSG posts. Such duplication can be resolved with slightly different specific times mainly due to the differences in the granularity of the temporal expression. In such instances, mentions of same clinical event are aggregated and its representative time is resolved using the median time of the respective cluster.

5.6.2 Emotion timeline

The emotions timeline is an emotional representation of the state of mind of a user based on his discourse over time. Unlike clinical events, emotions are better represented as a continues flow over time i.e., an emotional trajectory of a user.

For the construction of the emotion timeline, initially, the OSG posts of a user are sampled into fixed-time bins of time Δt (e.g., week, month etc.). The emotions of the OSG posts in a single time-bin is aggregated to determine the emotional expression over that time period. We hypothesis that the approach of aggregating emotional expressions over a time-period would provide a more unbiased account of the emotional state of that user over that time, as the averaging smoothed-out frequent fluctuations in the emotions but retain macro emotional trends that correlate with the clinical events.

The emotions are captured as a 16-dimensional vector for each OSG post based on the technique introduced in Section 5.5.5. The emotional representation of user U for the time period $t - t + \Delta t$ is obtained by averaging the emotional vectors of the OSG posts that has the time stamp within that time period.

5.7 Evaluation of the information extraction techniques

This section evaluates the three knowledge extraction modules described in Section 5.5: (i) narrative type classification, (ii) age extraction, and (iii) gender extraction. We obtained the services of qualified domain experts for manual classification of test datasets. Narrative type classification is evaluated using a labelled set of posts as advice or experience. Age and gender resolution is evaluated using a labelled set of OSG post authors using their published posts.

5.7.1 Narrative type classification performance

Narrative type classification is evaluated using 500 posts labelled by domain experts as experience or advice. Note that we ignored the sub-classification of experience (first person or second person) for this evaluation, because second person experiences are relatively rare in the dataset. It is evaluated as a classification problem where the two classes are Experience and Advice. Table 5.8 presents the evaluation results.

Table 5.8: Evaluation results of the narrative type extraction technique

Label	Number of posts	Precision	Recall
Experience	329	0.92	0.96
Advice	171	0.91	0.81
Combined	500	0.92	0.89

The results show that both *experience* and *advice* are identified with a precision above 0.9. Recall of *advice* is relatively low mainly because some advising posts are mixed with the authors experience and therefore hard to identify them as advice.

5.7.2 Age and gender resolution performance

Age and gender resolution was evaluated using a set of 300 labelled author profiles. Posts of each author were examined to identify the age or gender of the author if such information is present. Each author profile is annotated based on the identified age and gender.

The labelled data is then compared to the output of age and gender resolution modules. We employed precision and recall statistics for this evaluation. Note that age is often mentioned in incremental values for some authors as a result of prolonged contribution to the OSG over several years. Therefore, age resolution is considered correct if it falls within two integer values of the labelled age.

Similar to the previous evaluation, performances of the gender and resolution modules are evaluated as classification problems. For gender, the classes are *Female*, *Male*, and *Unknown* and for age, classes are *Age mentioned* and *Age not mentioned*. Note that, in *Age mentioned* class the classifier has to correctly resolve the age value (within two integer values to the labelled age value) in order to be a true positive.
Table 5.9 and Table 5.10 present the age and gender classification results respectively.

Table 5.9: Evaluation results of the age extraction technique

Label	Number of profiles	Precision	Recall
Age mentioned	131	0.78	0.89
Age not mentioned	169	0.95	0.84
Combined	300	0.86	0.87

Table 5.10: Evaluation results of the gender extraction technique

Label	Number of profiles	Precision	Recall
Female	109	0.91	0.87
Male	35	0.90	0.77
Unknown	156	0.87	0.94
Combined	300	0.90	0.86

Both age and gender resolution have average precision and recall over 0.85. Precision in *Age mentioned* class is relatively low because when a profile does not have actual age mentions, the classifier tends to pick up low confidence age mentions that are often age mentions in past incidents or age mentions about other people. The same issue resulted a relatively low recall in *Age not mentioned* class as well.

Recall for *Male* is relatively low. Most of those misses are classified as *Unknown* as the classifier misses the gender specific clues. This is mainly because males tend to expose very few clues about their gender compared to females.

5.8 Demonstration

This section demonstrates the capabilities of the proposed platform in addressing the information needs of patients (end-users) and researchers. Patient information needs are based on finding similar cases that are more relevant to them (e.g., patients with similar

conditions). Researchers aim to extract aggregated insights on medical conditions based on different dimensions such as age, gender and time. Two use cases are presented in the following subsections, one each for a patient and a researcher, to demonstrate how their information needs can be addressed using knowledge extraction layer generated by the proposed method.

The rest of this section is organised as follows. First subsection describes the test dataset collected from an active OSG and then present the implementation of the proposed knowledge extraction layer. Next subsection delineates the implementation of personalised search with a patient use case. Finally, the last subsection demonstrates OSG analytic capabilities for the researcher use case.

5.8.1 OSG data collection

The popular OSG *patient.info*[4] and *healingwell*[5] mentioned in Table 5.1 were selected for data collection because of their high volume of posts and large number of participants.

A web scraping tool is developed to automatically traverse through each topic and collect all the threads. From each thread, title, first post and subsequent reply posts are collected. Each OSG post is collected with its time-stamp and author id. As shown in Figure 5.4, the posts are then aggregated by the author id. We have filtered-out the users who have less than five posts in the OSG, because the information exposed by such users are often not sufficient to understand their demographic and clinical factors. The final dataset contains 4,469,107 posts from 79,829 OSG users, where 1,662,312 posts (47,712 users) are from *patient.info* and 2,806,795 posts (32,117 users) from *healingwell*.

The collected OSG posts are stored as JSON documents using a setup of the open source search platform Elasticsearch[6]. Elasticsearch search is employed both as the document store and search platform since it handles both full-text and structured search. Elasticsearch is a distributable full-text search engine, designed to be scalable to handle very large datasets (Kononenko et al., 2014).

5.8.2 Information extraction

The information extraction techniques developed in Section 5.5 are employed to enrich the extracted OSG posts.

The narrative type of each post is identified individually for each post, where each OSG post is categorised as *experience: first person*, *experience: second person*, and *advice*.

The age and gender are resolved for each author. All the posts of an author are aggregated using the associated author-id of each post. These aggregated posts are then processed to resolve age and gender of each author.

Table 5.11 provides a summary of the demographics identified. The proposed demographic extraction modules managed to resolve age and gender of 42% users which

[4]http://patient.info/forums
[5]https://www.healingwell.com/community/default.aspx
[6]http://www.elastic.co/products/elasticsearch

Table 5.11: Counts and percentages of extracted age, gender and narrative type

	Users (percentage)	OSG posts (percentage)
Total	79,829 (100)	4,469,107 (100)
Age resolved	48,126 (60.3)	4,045,781 (90.6)
< 20	10,380 (13.0)	824,405 (18.5)
21-30	10,303 (12.9)	664,950 (14.9)
31-40	6,849 (8.6)	534,157 (12.0)
41-50	6,939 (8.7)	569,000 (12.7)
51-60	6,581 (8.2)	732,864 (16.4)
61-70	4,536 (5.7)	519,317 (11.6)
> 70	2,538 (3.2)	201088 (4.5)
Gender resolved	44,700 (56.0)	3,963,523 (88.7)
female	32,268 (40.4)	2,943,776 (65.9)
male	12,432 (15.6)	1019747 (22.8)
Age and gender resolved	33501 (42.0)	3770618 (84.4)

comprises of 84.4% of the posts in the collected dataset. It is a 12% increase compared to the reported 30% success by Cho et al. (2013).

Among the gender resolved users, 72.2% are females whereas only 27.8% are males. The gender resolved OSG posts are further skewed, where among the gender resolved posts 74.3% from females and 25.7% from males. We have further investigated this in the research literature. Kummervold et al. (2002) report that women are more participative in OSG than men. Li et al. (2015) state that women tend to self-disclose more information than men. Therefore, we assume there is less male participation in OSG and also even the participating males tend to expose less information to identify their gender.

5.8.3 Personalised search to extract relevant information

One of the key driver to participate in OSG is the ability to engage with individuals who has similar conditions (Tanis, 2008). In fact, Fox (2011c) reports that in 2011, 20% of adult internet users have used internet to find individuals with similar health concerns.

As discussed in Section 5.4, stage slowromancapv@ of the proposed framework provides personalised (relevant and reliable) information in response to patient search queries based on their medical and demographic information.

We first employed the medical information extracted from the query to identify the posts that are experiences and contain matching medical information. Such posts are then ranked based on the demographic similarity of the author to the demographics of the patient. For this ranking, we use a custom relevance measure based on age and gender.

Let age and gender of an information seeking patient P^i be P^i_a and P^i_g respectively. Using the same notation, let age and gender of an existing patient P^e be P^e_a and P^e_g. The relevance measure r_E is defined as follows:

$$r_E = W_a \times \exp(\frac{-|P^i_a - P^e_a|}{\sigma^2_a}) + W_g \times \mathrm{I}(P^i_g = P^e_g) \tag{5.4}$$

where W_a and W_g are the weights for age and gender respectively. I is the indicator function which is 1 if P_g^i equals P_g^e and 0 otherwise. Weight of age is associated with a Gaussian decay function which is 1 if P_a^i equals P_a^e. σ_a^2 is used to control the granularity of age matching where smaller values make the decay function to decrease rapidly with the age difference and vice versa. W_a and W_g are set to 0 if the patient does not provide their demographic details. The OSG posts are ranked based on this relevance measure and presented to the patient.

In this patient use case, we demonstrate how personalised retrieval can provide relevant and reliable information to the patient compared to the existing full-text search. We use the same query: "Im a 40 year old woman taking Nexium for heartburn". Contextual information is initially extracted from the query and used to identify medical information (symptom: heartburn, medication: Nexium) and demographic information (age: 40, gender: female). Most relevant experiences are retrieved from the database using the abovementioned method.

For comparison of the results we employed two approaches of full-text search: (i) default search in the OSG, and (ii) search key terms with the Boolean aggregation AND (retrieves the posts that contain all the search terms). Excerpts from the top five results from our method and the two approaches of the full-text search are provided in Table 5.12. The key terms that are relevant to the query is highlighted in each excerpt. Note that, some experiences retrieved by the proposed method does not have age or gender mentioned in that post. This is because age and gender is resolved for each author using all posts by that author, so age or gender of that author is inferred from other posts and not the retrieved post.

Above results show that the proposed method retrieves more relevant posts for the given query. Instead of taking key terms of the query as-is, the proposed method identifies the patient is a female and her age is 40. Therefore, it retrieves similar experiences from females who are aged close to 40. Also, the posts do not necessarily need to have age and gender mentions in the post itself as they were resolved for each author.

In comparison, the second query (full-text search in the OSG) attempts a direct string matching with the search terms and retrieves partially matched results that contain any (unknown) combination of search terms. It is apparent that the last match of this query is irrelevant, because it is only matching woman and 40 but does not have the symptom heartburn or the medication 'Nexium'. The third query has a very low recall with only one retrieved post, as it is rare to have all four terms in a single post. Also, it is clearly noticeable that the match term '40' is not an age mention.

5.8.4 OSG analytics for researchers

Medical research is often conducted using small samples of patients due to the associated cost (both time and money) of such research. On the other hand, such information is accumulated in OSG, crowd sourced by real patients. These untapped resources are

Table 5.12: Excerpts from the top five results obtained using the three search approaches (including the proposed method) to retrieve similar experiences for the query "I'm a 40 year old woman taking Nexium for heartburn"

Querying method	Excerpts from top five results
The proposed structured search Breaks the query into the following structure: symptom:heartburn medication: Nexium age: 40 gender: female	(*female,40*): How I cured my gastritis ... side effect of fish oil is **heartburn**... my doctor said I could try **Nexium** as well I decided not to... **My husband** made gluten-free banana bread... **I am 40**... (*female,40*): I had a very severe attack during a 24 hour ph probe test. I used to have **heartburn**... I am on nexium, ranitadine and donperidone... **I too am 40 years old** but I feel 80 ... (female,43): My ...it's been in a long time and the heartburn seems to be easing off... inhibitors are medication for reducing acid in your tummy like nexium and zoton... I'm only 43 and feel my life is... (female,50): I'm a 50yo Aussie female with Barrett's... I still have heartburn if I eat/drink the wrong things or forget to take my medication for a few hours... I was on Nexium 40 forever until I saw a different GP... (female,30): i am bloated and get heartburn all the time...iam 30yrs old wiv 4kids...take them today along with nexium. Nexium in my opinion is of no use at all...
Full-text search using 'heartburn Nexium woman 40' (searched in the actual forum search of patient.info website)	Nexium and side effect, anyone here while taking Nexium suffer diarrhea... I was taking another PPI tablet for heartburn for about 2 years (Nexium) Constant burping ...no pain or heartburn. I have been taking nexium for the past two weeks to relieve constant burping... just wondering if ... if you have been on Nexium for a long time? Women would be the ones who might find their iron levels low... I am a 40 year old woman with no notable health problems aside from the nephrotic syndrome. At 10 years old I was diagnosed with primary focal segmental glomerulosclerosis
Full-text search using 'heartburn AND Nexium AND woman AND 40' (searched within the collected posts using Elasticsearch)	the meds slowly over two years improved symtoms, during the symptom phase l started with different pain, chronic heartburn, leading to gall bladder removal...bought nexium as ld read theyre same as omp, they are expensive... they use 40-80gm.... but tomorrow will try cabbage juice, just told woman on cfs site who has chronic nausea to try it

inaccessible to researchers due to inherent noise, unstructured nature and diversity of information representation. Researchers have to attempt the formidable task of executing full-text queries and manually extract information from the resulting posts.

As discussed in Section 5.4, stages slowromancapiii@ and slowromancapiv@ of the proposed framework builds a structured layer on top of the unstructured text of OSG posts which can be utilised for OSG analytics. As shown in stage slowromancapiii@ of Figure 5.4, each OSG user can be represented using demographic, clinical and emotional dimensions that enables researchers to conduct OSG analytics and gain insights. It provides unprecedented access to OSG data from different viewpoints.

In order to demonstrate the OSG analytics capability, we performed several analyses on patients who report the symptom heartburn. Note that this attempt is solely to showcase the analytical capabilities and not a comprehensive medical research on heartburn.

Dimensional analysis: In this analysis, we combine age and gender dimensions and present demographic distribution of patients who report the symptom heartburn.

Figure 5.9 shows the demographic distribution of posts that mention the symptom heartburn. This type of analysis is useful to identify the age groups that are more affected by a particular symptom and also observe potential demographic biases.

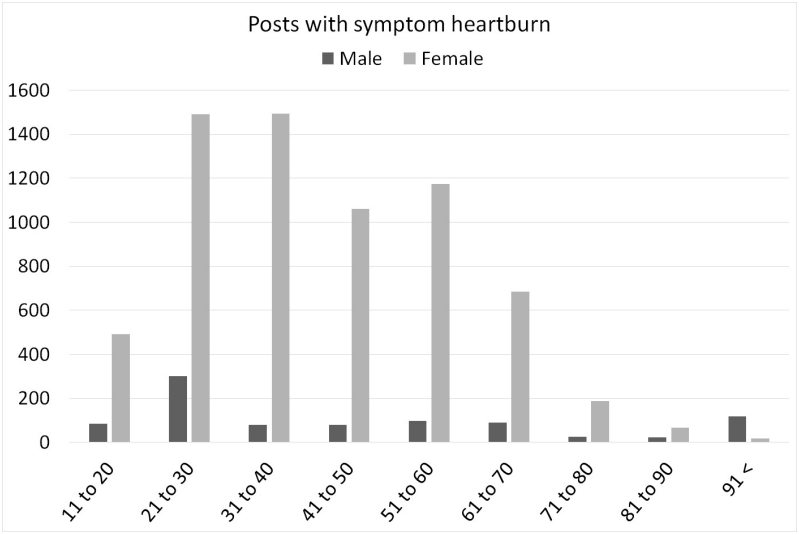


Figure 5.9: Demographic distribution of the posts that mention symptom 'heartburn'

Association mining: The OSG analytics layer is also useful for association mining. It can be used to analyse associations between different symptoms in order to identify co-existing symptoms. Table 5.13 presents the top five other symptoms that co-exists with the symptom heartburn in different age groups.

Table 5.13: Top five symptoms co-exists with the symptom 'heartburn' in different age-groups

Age group	Top five symptoms co-exist with 'heartburn'
< 20	less sleep, depressed, tiredness, anxiety, stress, living alone
21 to 40	reflux, anxiety, less sleep, nausea, stress
41 to 60	anxiety, stress, depression, reflux, indigestion
61 to 80	reflux, indigestion, anxiety, constipation, less sleep

Temporal analysis: The Date dimension can be used to perform temporal analysis to identify seasonal patterns in the OSG. Figure 5.10 shows the temporal distribution of the post counts that report the symptom heartburn drawn for each month for a period of three years. It shows that over the three-year period reported heartburns are relatively high during March and April.

5.9 Chapter Summary

This chapter presents a multi-stage information structuring platform for online support groups which facilitate the information needs to of OSG stakeholders: consumers, researchers and health professionals. This platform process unstructured OSG messages using a suite of machine learning and natural language processing techniques to retrieve age, gender, narrative type, medical entities, emotions and patient timeline. Age, gender,

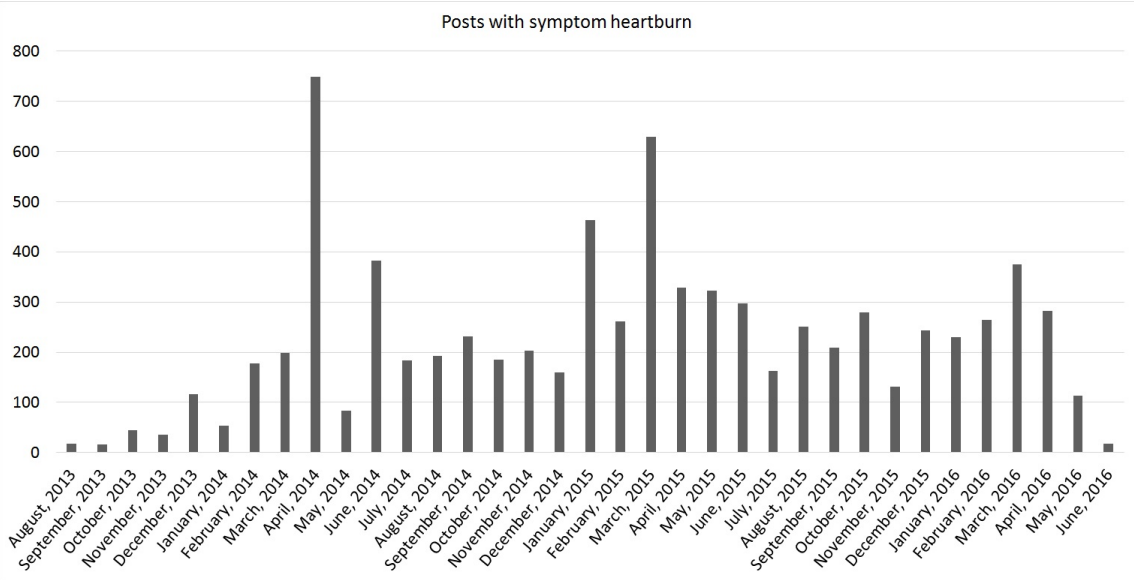


Figure 5.10: Temporal distribution of the posts that mention symptom 'heartburn' over a three-year period

and narrative type extraction modules show high precision and recall when evaluated with labelled data sets. The demonstration shows that the extracted attributes provide relevant and reliable information to consumers. Also, aggregated information is useful for research purposes. Next chapter extends this platform to structure OSG related to prostate cancer.

Chapter 6

A digital health platform

When you come to the end of your rope, tie a knot and hang on.
Franklin D. Roosevelt

This chapter employs the information extraction and structuring platform presented in the previous chapter to analyse the free-text discussions in Online Support Groups (OSG) related to prostate cancer and conduct a multi-dimensional exploration on the emotional expression of various groups and the underlying causalities for such expressions. This analysis informs the key groups that are in need of support as well as the aspects upon which such support needs to be delivered. As the number of cancer survivors is increasing globally within finite healthcare systems, this understanding is pivotal to optimise the care schemes and also to optimally provide care to vulnerable groups.

The platform presented in Chapter 5 is a platform designed for OSG on any health topic and it extracts demographic, clinical and emotional information. In this chapter, this platform is further extended to capture prostate cancer specific clinical and decision making related information. This extended platform is applied on a large corpus of prostate cancer related OSG discussions collected from ten active OSG and the collected insights are presented as two case studies where one explores the groups that are in need of more psychological support and the other investigates different decision making behaviours and decision factors.

The rest of the chapter is organised as follows. Section 6.1 briefs on cancer incidence, unmet needs of cancer survivorship and OSG as a potential resource to understand such needs. Section 6.2 justifies the selection of prostate cancer for the analysis due to its higher survivorship and unique characteristics related to treatment choices and side-effects. Section 6.3 presents the extended platform to analyse prostate cancer OSG discussions. Section 6.4 discusses the prostate cancer OSG data collection. Sections 6.5 and 6.6 presents the two case studies on the emotional expression of different groups and treatment decision making. Section 6.7 concludes the chapter with a discussion.

6.1 Cancer, cancer burden and cancer care

Cancer is a diverse group of diseases in which cells in some part of the body begin to multiply abnormally. In contrast to normal cells, cancer cells do not die when they should and new cells are produced when they are not needed, thereby developing a lump of cancer tissues i.e., tumour. This abnormal and out of control growth of cancer cells damage nearby tissue and also reduce the supply of nutrients to nearby normal cells. Such damage hinders the functionality of the body part with the abnormal growth and may fail to function eventually causing death. Also, cancer cells are *malignant* which means that they can spread into other body parts using the blood or lymph systems [1].

According to Global Cancer Observatory, in 2018 there are an estimated 18.1 million new cancer cases diagnosed and 9.6 million cancer related deaths worldwide (Bray et al., 2018). In US, American Cancer Society estimated that there are 1.7 million new cancer cases diagnosed and 0.6 million cancer related deaths in 2018 (American Cancer Society, 2018; Siegel et al., 2018). Cancer incident rates are rising as the growth of population and increase of life expectancy, as well as changes of lifestyles (e.g., obesity, lack of exercises) in developing countries due to socioeconomic development (Bray et al., 2018).

Although cancer incident rates increases, cancer related deaths are decreasing due to the advances of treatment options as well as the advances in screening and early detection programs (Bray et al., 2018). Therefore, *cancer* which was once uniformly fatal in a short period of time has been transformed into a diverse set of diseases with different survival rates. Long-term survival is possible in the majority of major cancers types when diagnosed earlier. In US alone by 2018 it is estimated that there are 15.5 million cancer survivors, which is expected to rise to 20.3 million (10 million males and 10.3 million females) in 2026 (Miller et al., 2016).

Besides for the cancer patient/survivor; partners, family and friends often provide the necessary care during the primary or adjuvant treatments as well as during the longterm management of cancer as a chronic disease. It is reported that in 2015, at least 2.8 million Americans provided care to a cancer patient (National Alliance for Caregiving, 2016).

6.1.1 Unmet needs of cancer

The transformation of most cancers from a fatal to chronic disease and the growing number of cancer survivors have become a key challenge to the modern health care systems as it struggles to understand and allocate resources to support the key needs of cancer patients and survivors.

Cancer patients are increasingly involved in the treatment related decision making and are in need of better support on selecting treatment options. The cancer survivors, although being relived of the primary treatment, are unprepared to live with cancer and its treatment related chronic outcomes (e.g., changes to the body image, infertility, stroke, intimacy issues), and also often worry about cancer recurrence (Council et al., 2005; Alfano and Rowland, 2006). Moreover, cancer caregivers report that long-term caring of the

[1] https://www.cancer.gov/about-cancer/understanding/what-is-cancer

loved one is physically, emotionally, and financially demanding (Northouse et al., 2010; Adelman et al., 2014; Kim and Schulz, 2008) often leading to high levels of stress and anxiety (Alfano and Rowland, 2006) and in need of help to manage their physical and emotional issues (National Alliance for Caregiving, 2016).

The healthcare providers and institutions are progressively limited in their scope of reach and service, due to increased demand, financial constraints and resource limitations (Aiken et al., 2002; Luxford et al., 2011). Therefore, it is important to understand the core needs properly, prioritise the allocation of resources based on the importance, and identify the groups that essentially in need of support so that the support can be optimally delivered with the available resources.

6.1.2 Understanding the psychological burden of cancer

Although there have been so many studies that led to advances in cancer preventing, screening and treatment which treats cancer clinically, the studies on the social and psychological burden of cancer have been limited. This lack of studies is one of the key reason for lack of resource allocation in the current health care system to provide the required support to overcome the social and psychological challenges of the cancer patients, survivors and caregivers.

The studies on social and psychological burden of cancer are often carried out as randomised control trials and cohort studies where the social and psychological well-being is assessed using survey instruments either designed for general quality of life or specialised for certain moods such as depression and anxiety (Groth-Marnat, 2009). For example, some generic instruments are Rotterdam Symptom Check List (RSCL) (De Haes et al., 1990), Functional Assessment of Cancer Therapy-General (FACT-G) (Cella et al., 1993) and European Organization for Research Treatment in Cancer: Quality of Life Questionnaire (EORTC-QLQ-C30) (Aaronson et al., 1993). Some depression and anxiety specific instruments are The Distress Thermometer (DT) (Roth et al., 1998), The Hospital Anxiety and Depression Scale (HADS) (Zigmond and Snaith, 1983), and The Beck Depression Inventory (BDI) (Beck et al., 1996).

These studies require the individuals (patients, survivors, caregivers) to self-report their social and psychological well-being time-to-time and correlate to their clinical outcomes from cancer and cancer treatments, thereby developing an understanding of the social and psychological issues at different time points of the cancer journey for different cohorts individuals. However, such long-term analysis is cumbersome, expensive and often suffered from high drop-out rates as the individuals are less keen to support such studies that stretch for a long period of time. Moreover, as discussed in Section 3.1, such studies are often subjected to issues in generalisability due to small sample sizes, and various recall biases.

6.1.3 OSG as a complementary approach to understand the psychological burden of cancer

As discussed in Section 5.2.1, OSG are a form of a social media platform that provides anonymous comfortable virtual spaces for patients survivors and carers to share experiences, seek advice, express emotions and provide emotional support. Participants with similar experiences provide each other with informational and emotional support.

As pointed-out in Section 5.2.3, OSG contain unsolicited accounts of self-reported first person experiences at different time points of the illness journey. Those self-reported expressions are scattered across the OSG as parts of different discussions. However, once aggregated they provide significant insights about the patient/caregiver and their illness journey over time. This resource is seen to be instrumental in addressing the limitations and challenges of the current approach to understanding the unmet issues of cancer. The free-text corpus in OSG discussions can be mined to uncover the needs and challenges of the cancer patients, survivors and caregivers based on the issues that they discuss at different stages of their cancer journey. The emotions expressed when discussing the issues can be employed as a proxy to the quality of life to understand the severity of the issues. The individuals can be categorised based on their demographics (age, gender), role (patient, caregiver), and other clinical factors (e.g., treatment, cancer aggressiveness). The issues expressed by different categories of individuals can be prioritised based on their emotional expression and employed in the decision making process to identify the pressing issues and the groups that essentially in need of support so that the support can be optimally delivered with the available resources.

This approach requires the relevant information about the individuals to be extracted from the free-text OSG discussions, which is challenging due to the noisy and unstructured nature of OSG content.

In order to validate this premise that OSG can be better utilised to gain insights about the diverse needs of cancer care among the patients, survivors and caregivers, a use case of prostate cancer has been selected. The next section provides an overview of prostate cancer and justifies its use as a case study to analyse the needs and challenges of prostate cancer patients, survivors and caregivers, discussed in prostate cancer OSG.

6.2 Prostate cancer and related OSG usage

Prostate cancer is the development of cancer in the prostate gland, a part of the male reproductive system. It is often slow growing and relatively less fatal compared to other types of cancers, but it may spread into other parts of the body (e.g., bones, lymph nodes). The key incidence factors of prostate cancer are age (mainly on males > 50) and family history.

6.2.1 Incidence

It is the second most frequent cancer among men (after lung cancer), and fifth leading cancer for mortality among men (Bray et al., 2018). In 2018 there has been an estimated

1.3 million new incidents and 359,000 mortalities of prostate cancer across the globe (Bray et al., 2018). It is the most frequently diagnosed cancer among men in 105 countries (out of 185 countries), including most of the high income countries such as US, UK, Canada, Australia, New Zealand and most of the northern and western Europe (Bray et al., 2018).

Despite being one of the most frequent cancers among men, its 5-year survival rate is highest in the developed countries, which is 99% (2017) in US (American Cancer Society, 2018), 84% (2010) in UK [2], and 95% (2014) Australia[3]. This increased survival is mainly due to the introduction of early detection schemes such as abnormal prostate-specific antigen (PSA) testing (Kvåle et al., 2007) so that cancer can be removed or contained during early stages of cancer. In US it is estimated that there are over 3.3 million men with a history of prostate cancer (Miller et al., 2016).

6.2.2 Diagnose

PSA and Gleason score are the key determinants of the prostate cancer stage. PSA which stands for *prostate specific antigen* is a protein made in the prostate gland. Increase of PSA level than normal in the blood is an indication of having prostate cancer, although such an effect can be due to various reasons, which may often lead to misdiagnose (false negatives). Due to the simplicity of the procedure, PSA test is often used as a routine test in older adults where abnormalities were directed to further tests.

Gleason grading system measures the development of cancer by analysing the cell morphology of cancer cells and grading them 1-5 where 5 being most aggressive (Gleason et al., 1974). The final score is determined by combining the score of the most common cell pattern and non-dominant cell pattern with the highest grade. The final score ranges between 2-10 which is indicative of the growth of cancer where 10 being the most aggressive.

6.2.3 Treatment and side-effects

There are three major treatment options prostate cancer:

1. **Surgery:** a surgical procedure known as radical prostatectomy that completely removes the prostate gland. It is the most widely used treatment for prostate cancer and often administrated as open surgery, laparoscopically or robotic-assisted.

2. **Radiation therapy:** using ionizing radiation destroy the malignant cells of cancer, it uses as a primary treatment method or as a secondary treatment following a surgery to remove any remaining cancer cells.

3. **Active surveillance:** for men with a less aggressive cancer (low-risk) or older individuals who have other serious conditions that prevent them from a curative treatment. It involves continues monitoring of cancer for its progression with the intent of switching to a curative treatment if cancer becomes aggressive.

[2] https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/prostate-cancer/survival

[3] https://prostate-cancer.canceraustralia.gov.au/statistics

As these treatments lead to different outcomes in terms of curing (or containing cancer) as well as the subsequent treatment related side-effects (Albertsen, 2016; Cooperberg and Carroll, 2015). Therefore, the patients are eager to chose the best treatment for them, which is often selected based on multiple factors such as the stage of the cancer, age, other medical conditions, financial conditions as well as the personal preference of the individual (Berry et al., 2003; Xu, Dailey, Eggly, Neale and Schwartz, 2011). In fact, there have been reports of decision regrets among the post-treatment survivors (Clark et al., 2001; Gwede et al., 2005) mainly attributed to treatment related side-effects. This complexity makes patients to seek multiple opinions on the treatment options from clinicians, family, friends and other prostate cancer survivors (Denberg et al., 2006; Berry et al., 2003). Also, patients acquire more information by reading relevant books, magazines and web pages (Berry et al., 2003) as well as interacting online with other individuals in OSG (Huber et al., 2011; Ihrig et al., 2011a).

The treatments of prostate cancer come with chronic side effects which are mainly urinary, sexual and bowel impairments. Among them urinary side effects such as urinary incontinence and sexual side-effect erectile dysfunction are intimate in nature, thus patients are reluctant or rather embarrassed to self-disclose and talks about these issues with their physicians, family and friends (Weber et al., 2000). This issue is more prevalent among men as they are less eager to seek support for social and psychological issues (Wang et al., 2013; Lintz et al., 2003) and discuss intimate health issues with physician or peers (Lintz et al., 2003; Weber et al., 2000). In contrast, online social media platforms such as OSG provide patients with the ability to communicate under a pseudo name without revealing their true identity. As discussed in Section 3.2.4 this anonymous nature leads to *online disinhibition effect* (Suler, 2004) where individuals are comfortable in self-disclose more frequently and intensely in online communication in contrast to face-to-face communication. Therefore, prostate cancer patients (or survivors) are more comfortable and likely to use OSG to discuss their intimate side-effect related problems (Nanton et al., 2018).

6.2.4 Prostate cancer OSG

The discussions related to prostate cancer happens in OSG either dedicated for prostate cancer discussions (e.g, Prostatecanceruk), or in generic OSG with a specific section for prostate cancer (e.g, Healingwell sub-forum on Prostate Cancer).

The above discussed high incidence rates and the high survival rates globally result a significantly large population of prostate cancer patients, survivors as well as caregivers of prostate cancer patients participating in prostate cancer OSG. Especially given that prostate cancer is prominent in the developed countries where most of the patient are frequent users of internet and online social media platforms. Therefore, a large volume of prostate cancer related discussions are accumulated in OSG platforms.

The participants use prostate cancer OSG at four key stages of their cancer journey which are diagnosis, pre-treatment, post-treatment side-effects and recurrence. Those stages of cancer are being identified as the most in need of psychological support (Weis, 2003) due to significant transitions of the quality of life as well as the significant decisions

to be made during those stages. The participants use OSG discussions to express their concerns and receive information and emotional support on how to cope with the situation. This process result accumulation of the information as well as the emotional state of the individuals during those stages.

Figure 6.1 presents an anonymised sample of five OSG posts by a patient, from diagnosis of cancer to four months post-surgery.

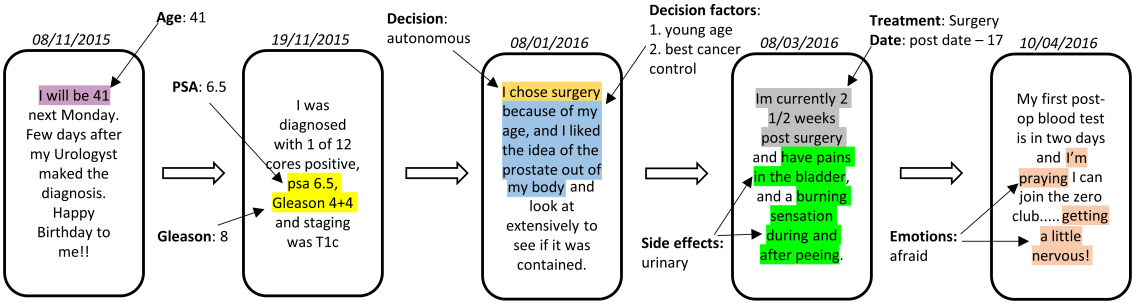


Figure 6.1: An anonymised sample (parts are omitted and rephrased to preserve privacy) of five posts by a prostate cancer patient. The highlighted excerpts are demographic, clinical, emotion expressions and decision making process related information, stated in the form of free-text.

Figure 6.1 demonstrates the wealth of implicit information contained within OSG posts. Patients begin by self-disclosing demographic and clinical information, followed by their decision-making process, relevant decision factors and emotions, in order to seek validation from other patients (Jayles et al., 2017; Mishra et al., 2013). The timeline of clinical and emotion information is implicit in the time-stamp of the post and in addition often explicitly mentioned in the post content. However, this entire body of information is encapsulated within large volumes of unstructured text data which lacks a domain-specific structure required for investigation or intervention and support by primary care providers (Murdoch and Detsky, 2013).

In nutshell, the prostate cancer OSG are voluminous due to its scale of occurrence and high survival rate, as well as information rich due to the unique set of characteristics of the prostate cancer patient trajectory. The next section presents the Patient reported information multidimensional exploration platform (PRIME) which is an extended version of the proposed information structuring framework in Section 5.4 to analyse prostate cancer specific OSG discussions.

6.3 PRIME: Patient reported information multidimensional exploration platform

This section presents the proposed Patient Reported Information Multidimensional Exploration (PRIME) platform which is developed to automatically analyse a large collection of cancer related OSG posts and extract demographic, clinical and emotional factors with their associated temporality. These factors were employed for automated investigation of

patient behaviours, clinical factors and patient emotions, across the temporalities of diagnosis, treatment and recovery. More specifically, we focus on the automated multi-granular extraction, analysis, classification and aggregation of decision-making behaviours, decision factors, the temporality of patient interactions, the temporality of clinical information and side effects, and trajectory of positive and negative emotions, in the context of decision groups, demographics and treatment type.

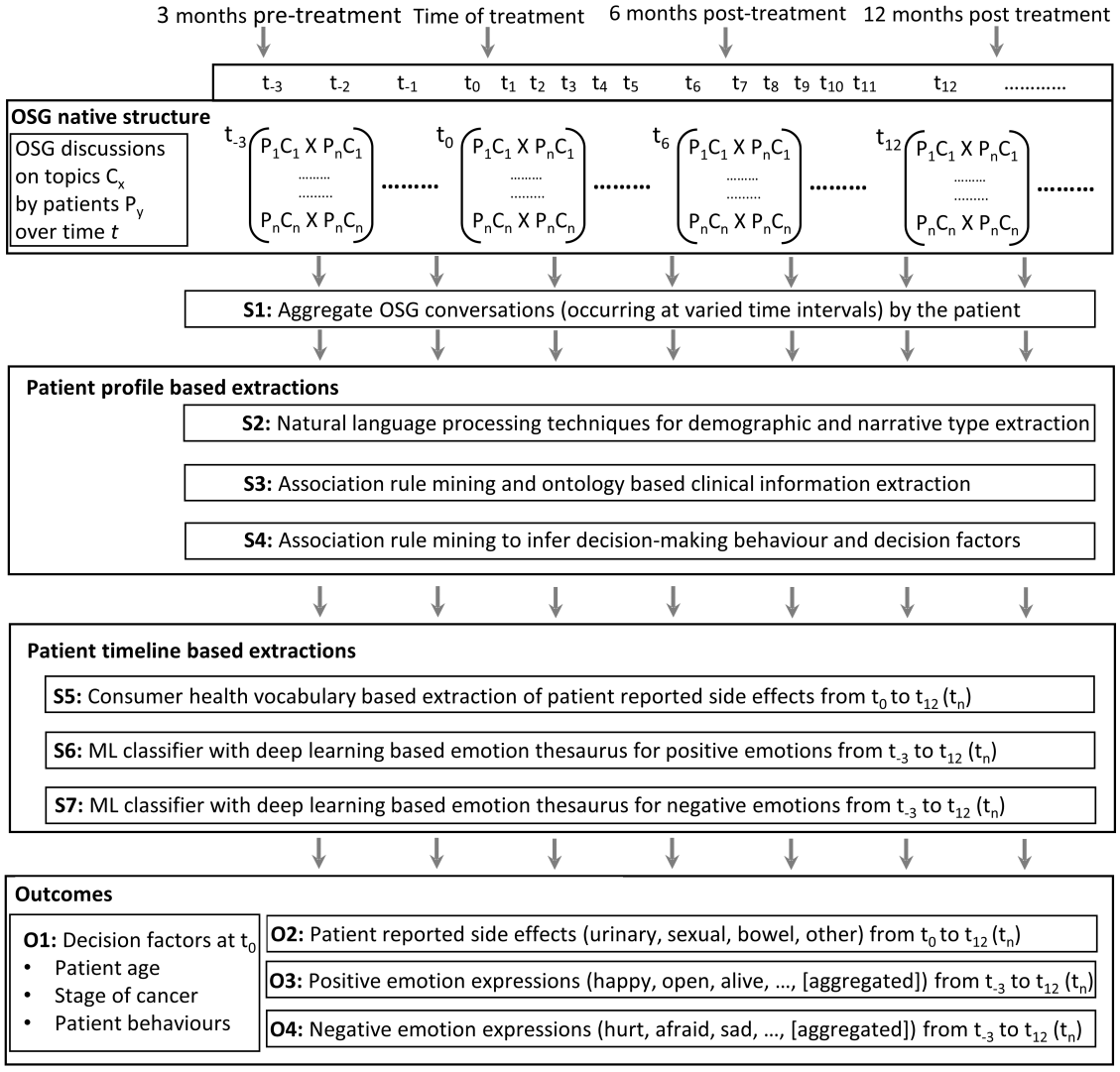


Figure 6.2: Patient Reported Information Multidimensional Exploration (PRIME) platform, which contains a suite of machine learning and natural language processing based information extraction modules to process cancer related OSG data.

As shown in Figure 6.2 the PRIME framework functions in seven *Stages S1-S7* to capture multiple layers of information from free-text OSG posts. As discussed in Section 5.1, OSG are organised as discussion threads where each thread is initiated by a user (often a patient, partner of a patient or a caregiver) with a particular question or health concern. Other participants respond to that with answers, suggestions, expression of their own experience or a follow-up question for further information. This process creates a discussion thread in OSG.

An OSG comprises a large number of discussions where patients contribute their decisions, experiences and opinions at different stages of their cancer journey from diagnosis to post-treatment and potential recurrence of cancer. Hence OSG contain a multitude of granular and aggregate information on patient behaviours, side effects and emotion expressions over time.

6.3.1 Aggregating OSG posts by the patient

The patient behaviour in an OSG over time can be represented as t_x $(P_1C_1 \times P_yC_z)$, where t_x denotes time, P_y denotes a contributing patient, C_z denotes an OSG thread and P_yC_z denotes an OSG post by patient P_y as a contribution to the discussion in thread C_z. In this naturally occurring order of OSG discussions, the posts by a single patient P_y are scattered over multiple discussions $(C_1, \ldots, C_z)$. Hence, as the initial step in the *Stage 1* the OSG conversations by a single patient are collocated and chronologically ordered based on the timestamp.

The timestamp of OSG post depicts its published time. However, as different cancer patients start their cancer journey at different times, published time cannot be used to compare and contrast patients. Therefore, the time of each OSG post is normalised based on a key cancer related event (e.g., diagnosis, treatment) of each patient, which is experimentally set to treatment event as it was found that many patients join OSG close to their treatment. Note that the cancer events and their temporalities are automatically extracted during the construction of patient timeline which will be discussed later in *Stage3*.

6.3.2 Patient profile based extractions

Stages 2-4 extracts various aspects of the patient profile such as demographics, clinical information and treatment decision making.

Capturing demographics

Stage 2 is based on natural language processing and machine learning based information retrieval techniques presented in Section 5.5, which extracts age, gender and narrative type. Based on the narrative type being *first-person* or *second-person* the role of the participant has being identified as the patient or caregiver, where female caregivers are predominantly the partners of patients. If the gender cannot be inferred with sufficient confidence, such individuals were marked as male patients as prostate cancer is a male only cancer. Patients are grouped based on age into five age groups (<40, 40-50, 50-60, 60-70 and >70) to be in line with the standard prostate cancer cohort studies.

Capturing clinical information

Stage 3 enriches patient profiles with clinical information, which are important to categorise patients based on their treatment type and the stage of cancer.

The treatment type that the patient has undergone is often disclosed during the discussions about treatment choice and their respective consequences. As mentioned in Section 6.2, multi-modality of treatment options makes the selection of treatment a complex process. Hence, it is important to compare and contrast the outcomes of a patient with different modalities.

The treatment information is extracted using the clinical event extraction technique (as part of the clinical timeline extraction) presented in Section 5.6 which is a specified event detection technique (See Section 2.3.2 on specified event detection). It extracts the OSG posts that mention any of the treatment modalities and then determines the temporality of that event based on the explicitly mentioned time (e.g., '2 weeks ago') and the implicit time based on the timestamp of the OSG post. Note that if the patient had undergone multiple treatments during their cancer journey (e.g., surgery and then radiotherapy) the first treatment is considered to represent his modality as it is the primary treatment in his cancer journey.

In OSG, participants disclose their PSA and Gleason scores as a means of informing their stage of cancer to the OSG community. However, such mentions employ diverse narrative styles. For example, Gleason score is mentioned as 'gleason', 'gleson', 'gleeson', 'g' and 'gs 'followed by the value in either as a total score (e.g., '7') or as a combination of the two components (e.g., '4+3'). In *Stage 3*, association rules and extracts from clinical ontologies are employed to develop regular expressions to identify these mentions of Gleason and PSA and subsequently, capture the numerical details of Gleason and PSA scores.

Capturing treatment decision making behaviour and decision factors

Stage 4 infers the decision making behaviour of each patient and the corresponding decision factors considered for the decision making. As discussed in Section 6.2, multiple modalities in prostate cancer treatment process make the treatment decision making a complicated process. Due to this complexity, patients are often eager to get involved in the decision making process to select what is best for them. Information of this decision making process is often posted in OSG with the intention of receiving peer-validation from those who had similar circumstances.

These decision making behaviours are captured based on the established patient decision making models (Charles et al., 1999; Emanuel and Emanuel, 1992; Veatch, 1972) which highlights different types of decision making behaviours depending on the degree of involvement by clinician and patient. The three prominent categorises of the decision making behaviours are as follows:

1. **Paternalistic**: also known as *priestly* which is the traditional behaviour in clinical decision making where clinician assumes a dominant role in making the decision based on his expertise and experience. In this approach patient delegates the decision making authority on the treatment to the clinician and provide consent to the recommendations of the clinician, assuming that the clinician would use his expertise

and experience to consider all potential possibilities and select the treatment that is the best option for the patient.

2. **Autonomous**: also known as *informative* or *consumer model* is the decision making behaviour where the patient assumes the dominant role. In this behaviour, the clinicians provide the relevant information, let the patient make the decision, and subsequently executes the suggested treatment option. In this approach, it is assumed that the patient with sufficient knowledge is capable of selecting the treatment option that best suits his condition.

3. **Shared**: also known as *deliberative* is the decision making behaviour where both patient and the clinician jointly work to find the best treatment options by weighing all options using multiple aspects such as clinical, social and financial.

These decision making behaviours are reflected in self-disclosed statements that announce the treatment decision of the individual (see third OSG post in Figure 6.1). In order to capture such statements, a set of template patterns was engineered to capture sentences that describe that either individual has taken the decision (Autonomous) or the treatment option was recommended by a clinician (Paternalistic). The template patterns are as follows:

- Autonomous template (T_A): `<I/We> <words>* <DECIDE> <words>* <TREATMENT>`
- Paternalistic template (T_P): `<DOCTOR> <words>* <RECOMMEND> <words>* <TREATMENT>`

Note that `<words>*` denotes zero or multiple words in-between, and upper case terms are *template terms* which consider a set of synonym terms (word or phrases). Table 6.1 shows a selected sample of terms employed as candidates for each template term.

Table 6.1: A sample of candidate terms for each template term used in the template sentences of decision making behaviour. Note that some of the terms are abbreviations or shorten versions of the actual terms or phrases.

Template term	Sample candidate terms
DECIDE	decided, chosen, wind up going, made the call, settled, opted, went for, took the option, end up
RECOMMEND	recommend, recommended, prescribe, prescribed, advised, advise, endorse, endorsed, advocate
DOCTOR	doctor, doc, surgeon, urologist, uro, specialist, consultant, radiologist, oncologist, radiotherapist
TREATMENT	surgery, davinci, da vinci, robotic, prostatectomy, ralp, rrp, lrp, rpp, key hole, open op
	radiation, imrt, brachytherapy, radiotherapy, seed therapy, brachy, seed implant, ebrt
	surveillance, AS, watch and wait

These template sentences from both templates T_A and T_P were used to capture the matching sentences in the OSG posts of each individual. The proportion $P(T_X) =$

$T_X/(T_A + T_P)$ (where X is A or P) appeared in the OSG posts of a user is used to determine the decision making behaviour of that user as follows:

1. **Paternalistic**: if $P(T_P) >= 0.25$ and $P(T_A) < 0.25$
2. **Autonomous**: if $P(T_P) < 0.25$ and $P(T_A) >= 0.25$
3. **Shared**: if $P(T_P) >= 0.25$ and $P(T_A) >= 0.25$

For example, if a user only has T_P or T_A, his decision making behaviour is categorised as Paternalistic or Autonomous respectively. However, if he has both T_P and T_A significantly ($>= 0.25$), then his decision making behaviour is Shared.

Multiple decision factors, both clinical/non-clinical affect the treatment decision-making process. As shown in the third OSG post of Figure 6.1, patients often mention underlying decision factors alongside with the mentions of the treatment decision often as a way of justifying their decision to the OSG community and seeking peer-validation.

The decision factors were searched in the OSG posts which are identified from the above technique as containing sentences related to decision making and captured using a thesaurus of consumer health terms. Initially, a list of common decision-making factors related to prostate cancer was created based on existing literature (Berry et al., 2003; Gwede et al., 2005; Huber et al., 2011, 2017; Ihrig et al., 2011b) and further validated by clinicians. This list includes medical concerns such as side effects, doctor skills, and best cancer control as well as socio-demographic reasons such as age, fast recovery and financial concerns. Once the decision factors were identified, a thesaurus of consumer health terms was developed by analysing a set of sample OSG posts identified from the above technique as containing information related to treatment decision making.

6.3.3 Patient timeline based extractions

Stage 5 onwards, PRIME framework incorporates the time dimension of OSG discussions and patient interactions. The patient timeline extraction technique presented in Section 5.6 is adapted with prostate cancer specifics to automatically generated a clinical event and emotion timeline of each individual based on the self-disclosed side effects captured in *Stage 5* and positive/negative emotions captured in *Stages 6-7*.

Stage 5 captures the self-disclosure of side effects and grouped them into four key categories: *urinary*, *sexual*, *bowel* and *other* in which *urinary*, *sexual*, and *bowel* are the key side effect categories of prostate cancer treatments (Hamdy et al., 2016; Donovan et al., 2016), and *other* represents the miscellaneous side effects arises due to prostate cancer and its treatments.

The relevant side-effects for each category were selected with the support of clinicians and a thesaurus of relevant terms (words/phrases) for each side-effect were engineered based on the medical concepts found in the UMLS Metathesaurus (Bodenreider, 2004). However as discussed in Section 5.5.4, health consumers (patients and caregivers) discuss side effects using everyday layman language (e.g., urinary incontinence described as leakage, leak, drip). Therefore, the above thesaurus is extended by adding consumer health

terms identified using a semi-supervised approach where a candidate set of terms were identified using a word-embedding (Mikolov, Sutskever, Chen, Corrado and Dean, 2013) created from prostate cancer OSG and further validated by a team of clinical experts. This developed thesaurus is employed to capture any mentions of an occurrence of side effects and map such mentions to the timeline based on the associated timestamp.

In *Stages 5-6* the positive and negative emotions were captured using the emotion timeline generation technique presented in Section 5.6.2. It provides a 16-dimensional real-valued emotion vector for each OSG post which included 8 positive and 8 negative emotions extracted based on the psychological emotional model Emotion Wheel (Plutchik, 1980b, 1991). The value of each emotion is a representation of the strength of that emotion expressed in the respective OSG post.

Each patient timeline is time-normalised based on the treatment month captured in *Stage 4* as t_0. The events (side effects and emotions) are aggregated monthly based on the reported timestamp, and the timeline is generated from three months pre-treatment (t_{-3}) to 12 months post-treatment (t_{12}) based on the available information.

The PRIME platform has been implemented using JAVA language and available at `https://github.com/tharindurb/PRIME`.

The following sections empirically evaluate the PRIME platform using a large corpus of prostate cancer related OSG discussions collected from 10 active OSG. The evaluation is carried as two case studies where the first case-study evaluates the emotional expression of those who have undergone treatment for low-intermediate risk prostate cancer and second case study evaluates decision making behaviours and their associated decision factors and emotional expression.

6.4 Prostate related OSG data collection

The social data related to online prostate cancer discussions were collected from high volume active OSG on prostate cancer discussions. An OSG is considered active if it has at least 100 new conversations per week. From these active OSG, conversations were selected using the specific topic 'prostate cancer'. Table 6.2 presents the ten OSG selected for the data collection and their URLs. Note that, some of these OSG are dedicated for prostate cancer discussions (e.g, Prostatecanceruk), while others are generic OSG with a specific section for prostate cancer discussions (e.g, Healingwell sub-forum on Prostate Cancer).

The identified discussions were extracted using a set of web scrapers developed for each OSG platform. The collected dataset contains 609,960 conversations from 22,233 OSG users, comprising a text corpus of 93,606,581 word tokens. Table 6.2 shows the number of users and number of posts collected from each OSG platform.

6.4.1 Ethical considerations

This study has been conducted under the ethics approval from La Trobe University Research Ethics Committee. All patient-reported data used in this study are non-identifying

Table 6.2: Population and volume statistics of the ten selected OSG on prostate cancer discussions.

OSG	URL	Users	Posts
Healingwell	healingwell.com/community	6,829	401,325
Cancer Survivors Network	csn.cancer.org/forum	2,775	33,166
Cancerforums	cancerforums.net	2,761	52,840
Cancercompass	cancercompass.com	2,524	12,141
Healthboards	healthboards.com/boards	1,938	17,144
Patientinfo	patient.info/forums	1,727	33,304
Macmillanuk	community.macmillan.org.uk	1,178	9,210
Prostatecanceruk	community.prostatecanceruk.org	1,119	28,041
Prostatecancerinfolink	prostatecancerinfolink.ning.com/forum	862	7,187
Ustoo	inspire.com/groups/us-too-prostate-cancer	519	15,602

and publicly available from the corresponding OSG. The OSG do not provide access to identifying information of patients, and PRIME does not process any identifying information. This work only publishes aggregates of the analysed data, which cannot be reverse engineered using any means for any form of re-identification.

6.5 Analysis of the quality of life of prostate cancer patients

As discussed in Section 6.2, the treatment of low-intermediate risk clinically localised PCa is becoming increasingly complex due to comparable cure rates of the available treatment options; radical prostatectomy (RP), external beam radiation therapy (EBRT), and active surveillance (AS). Therefore, in addition to the tumour characteristics, significant emphasis is placed on customising the side effect profiles of each treatment option to the patient, based on their preference, anxiety levels and experience of the medical professional (Zeliadt et al., 2006).

Recent studies (Barocas et al., 2017; Chen et al., 2017; Donovan et al., 2016) compare patient quality of life (QoL) in men randomised to AS, RP and EBRT demonstrating that QoL post-treatment mirrored reported changes in function, yet no significant differences were observed among the groups in measures of anxiety, depression, and general health-related or cancer-related QoL. While study settings are controlled for many variables, they are dependent on questionnaires completed in a *trial setting* and may not accurately capture *real-life* issues experienced by patients in diverse circumstances undergoing different treatment options for localised prostate cancer (Bowling, 1995; Carr and Higginson, 2001).

In contrast, the complex phenomenon as QoL is effectively communicated as free flowing text containing expressions of emotions instead of predetermined responses in a questionnaire (Carr and Higginson, 2001; Seale et al., 2010). A complementary approach to conventional studies is provided in this case study which employs the emotional expression extracted from PRIME as a proxy to patient reported QoL, and compare/contrast the expression of emotions across the three modalities and other demographic factors. Note that

the *time* factor is not considered in this study, thus, side-effects and emotions captured in the patient timeline are aggregated.

6.5.1 Inclusion criteria

Following the completion of the automated intelligent extraction of demographic, clinical information and expressions of emotion, inclusion criteria set in order to focus on the analysis of a specific cohort. Thereby, this study focuses on those who have undergone treatment (RP, EBRT, AS) for low-intermediate risk PCa. Low risk PCa (Gleason $\leq$6) and intermediate risk PCa (Gleason 7) according to D'Amico risk classification (D'Amico et al., 1998). This inclusion criteria is applied to all the individuals with known modality and Gleason/PSA scores, which resulted in a participant cohort of 6,084 individuals (27.3% of the total profiles extracted).

6.5.2 Participant characteristics

The selected participant cohort has RP as the highest modality with 4,241 (69.7%) participants, followed by 1,528 (25.1%) EBRT and 315 (5.2%) AS. Table 6.3 presents demographic and clinical information of participants (patient or partners of patients) in total and across each modality. The percentages in each column are determined relative to the total participants in each cohort (total or each modality). The *p-value* presented is from chi-square tests that evaluated the statistical significance across the three modalities.

Table 6.3 presents a comparison of different participant characteristics across the three modalities. Among the participants who selected AS, 85% have a Gleason score of less than 6, while only 15% have 7. This is because AS is preferred for patients who are *low-risk* (have a slow growing cancer). The age distribution among EBRT and AS are skewed towards older participants as some are not suitable for surgical procedures in RP due to their old age. The partner cohort across the modalities are around 10% while AS has lowest of 9%.

Side effect results show that those undergoing RP had comparatively high urinary symptoms and sexual side effects compared to the other groups, while those who had EBRT had comparatively high bowel symptoms. These findings on side effects are corroborated by key outcomes from a large randomised controlled study and two large population based prospective cohort studies (Barocas et al., 2017; Chen et al., 2017; Donovan et al., 2016).

6.5.3 Emotions by age

Figure 6.3a depicts the positive emotion categories of five key age groups. The categories; *open*, *happy*, *positive* and *good* are consistently close to 20% whereas *alive* is close to 15% and *interested* is close to 10%. The cohort aged <40 has on or above average emotion levels across all positive emotions.

Figure 6.3b presents the negative emotions of the same age groups. In general, negative emotions are less expressed compared to positive emotions. The expression of emotion *afraid* is significantly high for individuals aged <40 which is 10% higher than other age

Table 6.3: Demographics and clinical characteristics of the identified low-intermediate risk prostate cancer patients and partners of patients from the ten selected OCSGs. Note that, the percentages in each column is determined relative to the total participants in each cohort (total or each modality). The *p-value* presented in the last column is from chi-square tests that evaluated the statistical significance across the three modalities.

	RP	EBRT	AS	Total	
	n (% in RP)	n (% in EBRT)	n (% in AS)	n (% in Total)	p value
Total participants	4241	1528	315	6084	
Gleason score					
<=6	2543 (60)	831 (54)	269 (85)	3643 (60)	<0.001
7	1698 (40)	697 (46)	46 (15)	2441 (40)	<0.001
Median PSA	6	7.8	6.5		
Age					
<40	110 (3)	37 (2)	8 (2)	155 (3)	0.937
41-50	688 (16)	144 (9)	32 (10)	864 (14)	<0.001
51-60	1638 (39)	411 (27)	89 (28)	2138 (35)	<0.001
61-70	933 (22)	456 (30)	91 (29)	1480 (24)	<0.001
>70	197 (5)	192 (13)	39 (12)	428 (7)	<0.001
Unknown				1019 (17)	
Participant role					
Patient	3702 (87)	1319 (86)	288 (91)	5309 (87)	0.676
Partner	539 (13)	209 (14)	27 (9)	775 (13)	0.069
Side effects					
Urinary symptoms	2229 (53)	625 (41)	75 (24)	2929 (48)	<0.01
Bladder Irritation	804 (19)	407 (27)	32 (10)	1243 (20)	<0.01
Bladder neck contracture	586 (14)	196 (13)	17 (5)	799 (13)	<0.01
Urinary incontinence	1953 (46)	390 (25)	41 (13)	2384 (39)	<0.01
Urethral strictures	312 (7)	108 (7)	11 (3)	431 (7)	0.045
Bowel symptoms	84 (2)	98 (6)	1 (0)	183 (3)	<0.01
Rectal irritation	33 (1)	42 (3)	0 (0)	75 (1)	<0.01
Rectal bleeding	66 (2)	72 (5)	1 (0)	139 (2.28)	<0.01
Sexual symptoms	1679 (40)	424 (28)	40 (13)	2143 (35)	<0.01
Erectile dysfunction	1679 (40)	424 (28)	40 (13)	2143 (35)	<0.01
Other	579 (14)	254 (17)	14 (4)	847 (14)	<0.01
Bleeding	453 (11)	202 (13)	14 (4)	669 (11)	<0.01
Clot	221 (5)	66 (4)	3 (1)	290 (5)	0.002
Hernia	497 (12)	115 (7)	14 (4)	626 (10)	<0.01
Infection	796 (19)	297 (19)	40 (13)	1133 (19)	0.038
Stroke	119 (3)	54 (3)	3 (1)	176 (3)	0.041
Tiredness	411 (10)	333 (22)	5 (2)	749 (12)	<0.01

groups. Also, this group is close to 5% higher than others in *depressed* and *helpless*. However, the emotion *angry* is less expressed by <40 group in comparison to all other

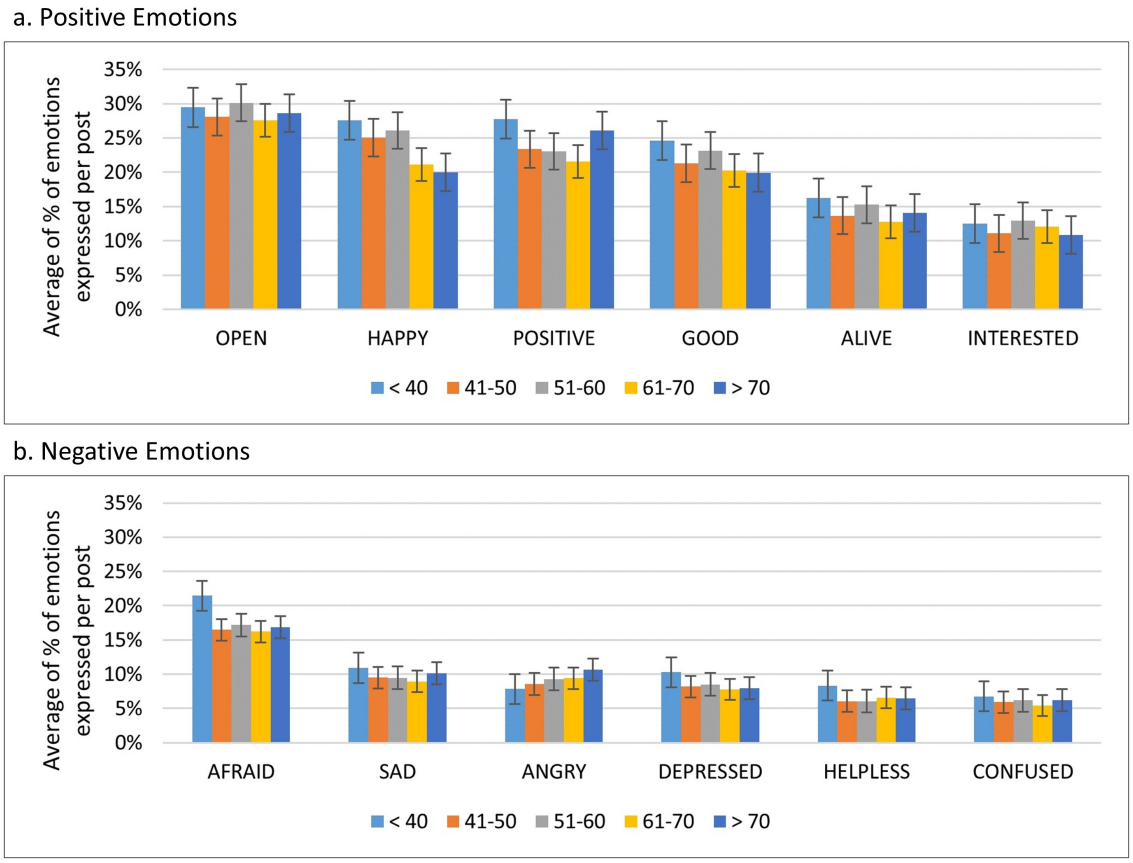


Figure 6.3: Positive and negative emotions by age group.

groups. The remaining two negative emotions *sad* and *confused* are expressed consistently across all age groups.

6.5.4 Emotions by modality

Figure 6.4 presents positive and negative emotions by modality. The positive emotion *open* is the most significant, while *alive* and *interested* are the least significant.

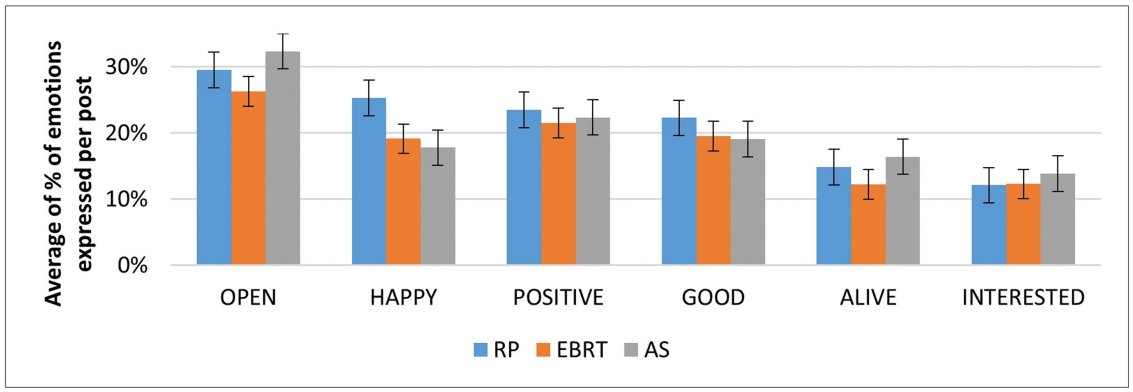
Figure 6.4b depicts negative emotions by modality and *afraid* is the most significant with an on average of 10% higher across all modalities. Noteworthy fluctuations are observed for AS; where *sad*, *depressed* and *helpless* are significantly low compared to the other two modalities.

6.5.5 Emotions by patient and partner

Figure 6.5 presents positive and negative emotions by patients and partners. Both groups were more *open*, and less *alive* and *interested*. Partners were slightly more *positive*, *good* and *alive*. On average, partners have expressed more negative emotions than patients, where *afraid* is expressed 7% more than patients.

Figure 6.5b presents partner negative emotions by treatment modality. Partners of participants under active surveillance are more *afraid*, more *angry*, and less *hurt* than other two modalities.

a. Positive Emotions



b. Negative Emotions

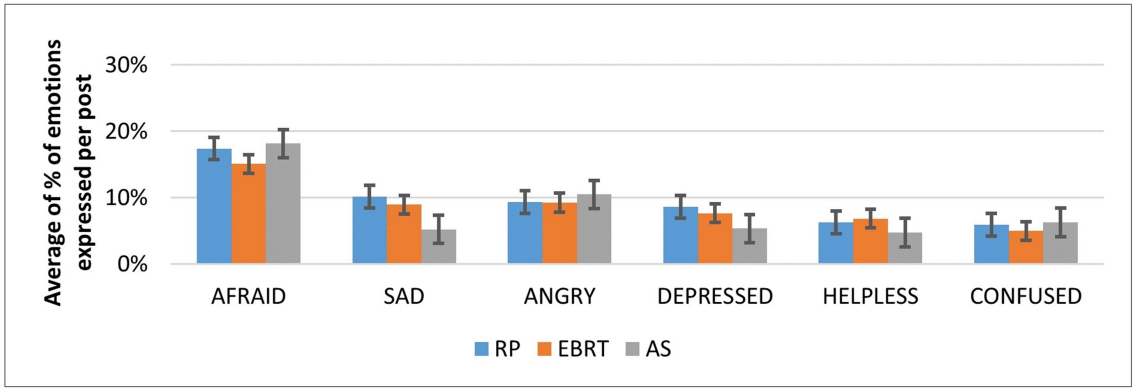


Figure 6.4: Positive and negative emotions by the treatment modality (RP, EBRT and AS).

6.5.6 Significant emotion terms

Table 6.4 shows the frequently used emotion terms by the three selected cohorts (<40, >70, partners), which were determined based on the percentage increase of frequency of each emotion term in the selected cohort compared to all the participants.

Table 6.4: Frequently-used emotion terms for three selected cohorts: (i) patients aged <40, (ii) patients aged >70 and (iii) partners of patients.

Group	Terms expressing positive emotions	Terms expressing negative emotions
Patients aged <40 (30% above baseline)	merry, lucky, intent, challenged, content, affected, loved, peaceful, brave, determined, pleased, tender, secure, reassured, wonderful	sore, fatigued, nervous, suspicious, frustrated, alarmed, bad, worried, frightened, scared
Patients aged >70 (10% above baseline)	reliable, accepting, touched, brave, courageous, satisfied, optimistic	unpleasant, threatened, frightened, fatigued
Partners of patients (20% above baseline)	quiet, spirited, reassured, loved, calm, admiration, brave, amazed, kind	tearful, upset, frightened, scared, guilty, worried, miserable

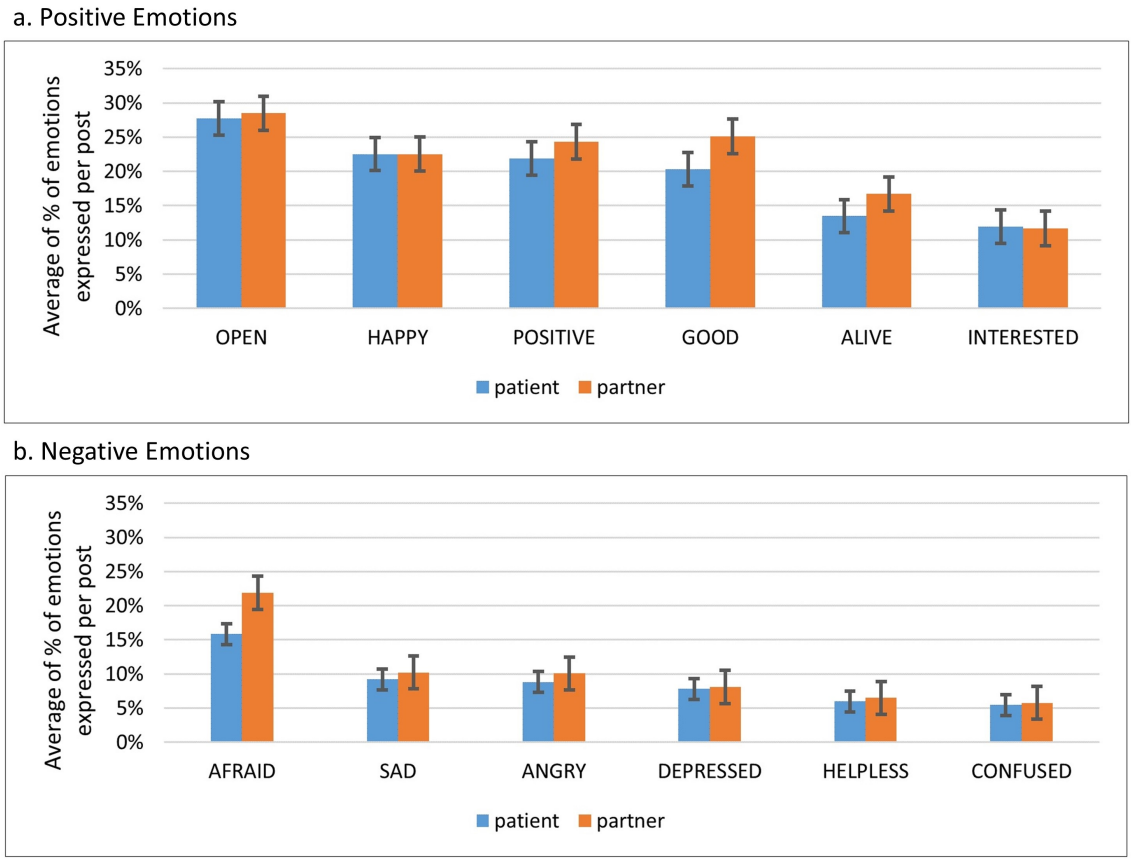


Figure 6.5: Positive and negative emotions by patient and partner.

Patients aged <40 have expressed significantly high positive and negative emotions compared to other age groups. Aligning with studies that report high levels of psychological distress in young cancer patients (Compas et al., 1999; Mosher and Danoff-Burg, 2005), this group demonstrates significantly high negative emotions and emotion terms (Table 6.4). Positive categories (*happy*, *positive*) and positive emotion terms (Table 6.4) are also high in this group. This is noted as post-traumatic growth in trials on other types of cancer patients (Blank and Bellizzi, 2008; Manne et al., 2004; Pudrovska, 2010). As such, this cohort appears to be the ideal beneficiaries of OSG and could gain from additional health care resources and discussions focused on treatment decision making and cultivating positivity.

Emotions of patients and partners differ significantly in several categories. Partners express higher levels of *afraid*, *positive*, these differences can be explained by (1) female partners being emotionally expressive than males (Kring and Gordon, 1998), thereby actively use OSG to share emotional experiences (Seale, 2006) and (2) caregiving partners report more emotional distress and anxiety than the patient (Given et al., 2001; Northouse et al., 2007). Negative terms significantly used by partners e.g., 'tearful', 'frightened', 'miserable' further confirms this. While the cancer care process is directed towards the patient with regular interactions with healthcare professionals and support groups; partner's burden of caregiving and anxiety of potentially losing their loved one, are often overlooked, highlighting the need for better support for partners of patients.

This study reports that significantly higher and more negative emotions were expressed in younger age cohorts and partners of patients who may benefit from increased psychological support from healthcare providers.

6.6 Analysis of decision making behaviour and decision factors

As discussed in Section 6.2.3, treatment decision making in prostate cancer is a complex process due to the availability of multiple treatment options and comparable outcomes. Therefore, the selection of treatment is not only based on clinical factors such as the stage of cancer, age, other medical conditions, but also other factors such as financial conditions and personal preference of the individual (Berry et al., 2003; Xu, Dailey, Eggly, Neale and Schwartz, 2011). Moreover, as discussed in Section 6.3.2, over the last few decades patients have assumed a more dominant role in decision making deviating from clinician dominated *paternalistic* behaviour to a more patient involved *autonomous* and *shared* decision making behaviours.

The PRIME framework has captured the patient decision making behaviours and the associated decision factors mentioned in the OSG by the OSG participants, which enables a large scale analysis of the decision making behaviours and key decision factors across different cohorts based on demographic and clinical factors.

6.6.1 Inclusion criteria

In this case study, the inclusion criteria is set to included patients those who self-disclosed their primary PCa treatment and discussed the decision-making process that led to the particular treatment. A total of 6,457 participants matched this criteria which is 29.0% of the total profiles extracted. Among them 420 (6.5%) have shown *paternalistic*, 3883 (60.1%) *autonomous* and 2154 (33.4%) *shared* decision making behaviours. This distribution shows that OSG participants predominantly get involved in the treatment decision making process (*autonomous* or *shared*). Note that there may be a representative bias on *paternalistic* individuals, as they are more reliant on clinical sources and comparatively less likely to participate in OSG discussions.

6.6.2 Decision making behaviour groups

Figure 6.6 presents a comparison of the distribution of decision behaviour groups across treatment modality, age group and Gleason score. The three Gleason score groups (< 7,7,$>$ 7) are based on D'Amico risk classification (D'Amico et al., 1998) as low-risk, intermediate-risk and high-risk respectively.

The results highlight that the *autonomous* group dominates each cohort, followed by the *shared* group leaving less than 10% for the *paternalistic* group. This shows that patient involvement in decision making is significant across all the cohorts. Active Surveillance (AS) has the lowest percentage on the *autonomous* group, but highest in *paternalistic* and

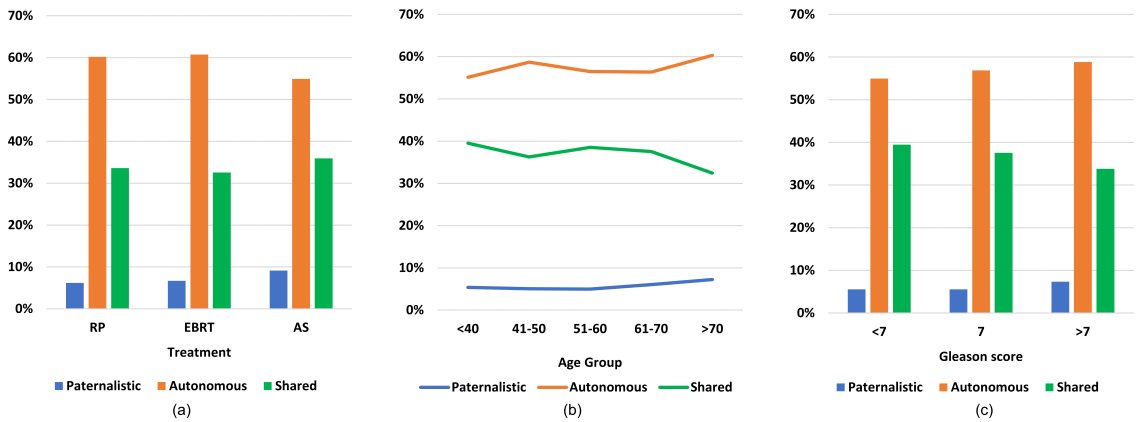


Figure 6.6: The distribution of decision making behaviour groups across: (a) Treatment modality, (b) Age group, and (c) Gleason score. Note that the percentage value for a behaviour group in a particular cohort is determined using the total (across the three behaviour groups) in each cohort as the denominator.

shared. The *paternalistic* group has a fairly uniform distribution across the age groups. The percentage of the *shared* group is highest among <40 and the *autonomous* group is highest among > 70, which shows that young individuals are more keen to jointly work with the clinician to make the treatment decision, while older individuals are more keen to take the decision themselves. The *shared* group is highest among Gleason <7 (low-risk) group. They may be more keen for shared decision making as they often have all three treatment options.

The decision making behaviour groups exhibit different behaviours in OSG activities over time in their patient trajectory such as actively participating to OSG discussions and type of activity (e.g., seeking support or providing support). The Figure 6.7 presents insights on these participant behaviours over time for the decision making behaviour groups.

Figure 6.7a shows the number of active participants over time which highlights that the participation peaked for all three groups during the month of treatment decision-making which is an indication of active information seeking by all groups. *Paternalistic* and *autonomous* groups reduce activity soon afterwards, but the *shared* group more consistently participated in OSG discussions throughout the next 12 months. This trend is reflected in Figure 6.7b as well, where the average number of posts by an active participant is significantly high for the *shared* group. Figure 6.7c shows that in all groups the proportion of *advice* posts increases over time which shows that the participants gradually become support providers to others who need. However, the shared group with a consistently high proportion of *advice* is shown to be more collaborative in OSG discussions. This transformation of the initial information/support seeker gradually becoming an information/support provider happens as the participants gain more relevant knowledge and wisdom (often by receiving support from peers), which is very important for the ecosystem in any support group as it provides the continues supply of information and support for those who need (Ziebland and Wyke, 2012).

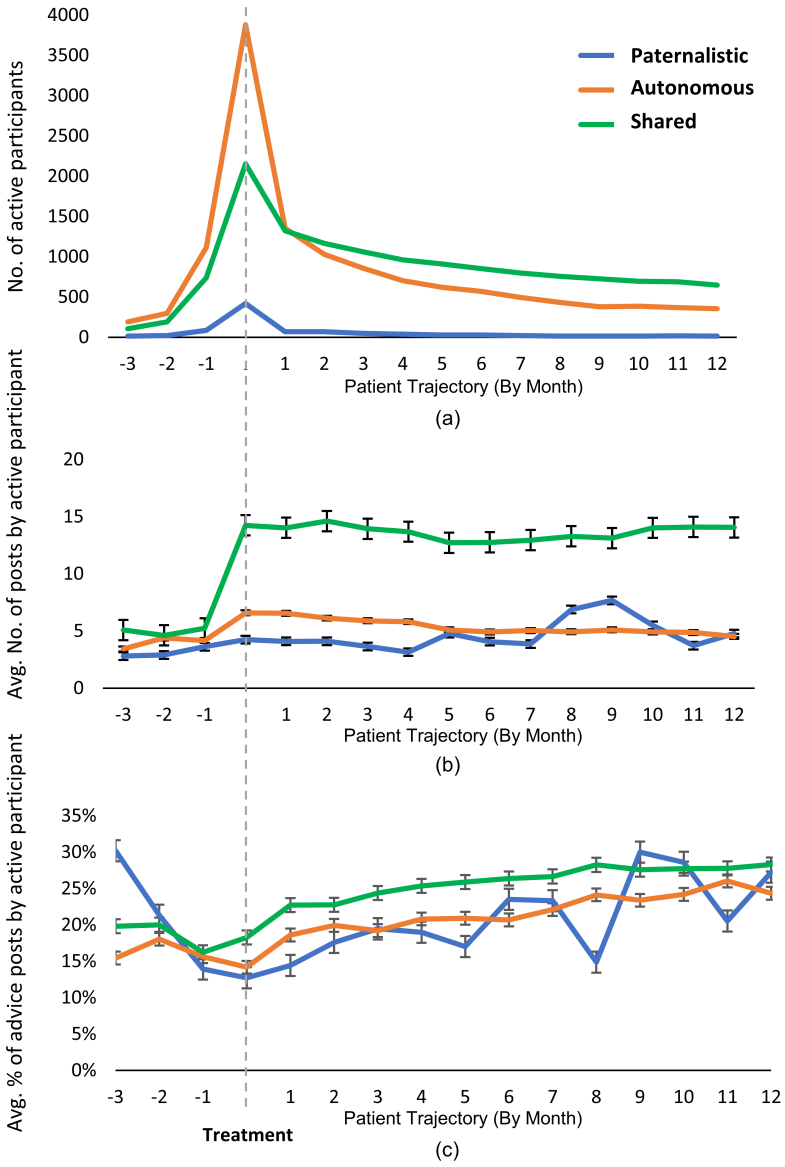


Figure 6.7: The OSG participation activities by the three decision making behaviour groups over the patient trajectory from 3 months pre-treatment to 12 months post-treatment. Note that an active participant during a particular month is an individual who has participated in OSG discussions during that month. The OSG posts are classified as an expression of experience or providing advice by the classifier presented in Section 5.5.1.

6.6.3 Decision factors

It is important to understand the decision factors that influenced the treatment decisions of the individuals. *Stage 4* of PRIME delineated in Section 6.3.2 extracted these self-disclosed decision factors alongside the selected modality and the decision making behaviour. This section analyse such decision factors against the selected modality and decision making behaviour shown by the participant.

Table 6.5 present the discussed decision factors against the selected modality, and Table 6.6 present the discussed decision factors against decision making behaviour. The *p-value* in the last column of both tables denotes the statistical significance of the mentioned decision factors among the cohorts.

Table 6.5: The decision factors employed for the treatment decision against the primary treatment modality of the patient.

	RP n (% in RP)	EBRT n (% in EBRT)	AS n (% in AS)	Total n (% in Total)	p value
Doctor skill discussed	2617 (66.76)	1392 (61.78)	190 (66.9)	4199 (65.03)	0.06
Surgeon mentioned	1975 (50.38)	880 (39.06)	141 (49.65)	2996 (46.4)	<0.01
Radiologist mentioned	371 (9.46)	488 (21.66)	27 (9.51)	886 (13.72)	<0.01
GP mentioned	142 (3.62)	75 (3.33)	11 (3.87)	228 (3.53)	0.80
Medical Indication	377 (9.62)	162 (7.19)	16 (5.63)	555 (8.6)	<0.01
Nerve mentioned	675 (17.22)	124 (5.5)	10 (3.52)	809 (12.53)	<0.01
Young age	272 (6.94)	93 (4.13)	16 (5.63)	381 (5.9)	<0.01
Bowel symptoms	13 (0.33)	46 (2.04)	1 (0.35)	60 (0.93)	<0.01
Urinary incontinence	624 (15.92)	196 (8.7)	8 (2.82)	828 (12.82)	<0.01
Erectile dysfunction	535 (13.65)	151 (6.7)	14 (4.93)	700 (10.84)	<0.01
Recovery time	214 (5.46)	50 (2.22)	0 (0)	264 (4.09)	<0.01
Best cancer Control	611 (15.59)	295 (13.09)	34 (11.97)	940 (14.56)	0.02
Financial concerns	336 (8.57)	179 (7.94)	17 (5.99)	532 (8.24)	0.29
Less invasive	117 (2.98)	31 (1.38)	1 (0.35)	149 (2.31)	<0.01

As shown in Table 6.5 'doctors skill' has been the most discussed decision factor and equally important across all the modalities. Similarly, best cancer control and financial concerns are equally discussed across all modalities reflecting they are equally important to all modalities. RP and AS groups talk about the surgeon more often while EBRT group is more interested on the radiologist. In terms of side effects preserving nerves, urinary incontinence and erectile dysfunction are mostly discussed by the RP group as they are the prominent side-effects of surgical procedures. EBRT group discuss bowel symptoms more often as they are the prominent side effects for radiotherapy. Young age has been given as a reason by mostly RP group followed by AS, since both are the most preferred by young patients.

The results in Table 6.6 shows that the shared groups overwhelmingly discussed all the decision factors than the other two groups ($p < 0.001$). This is because the *shared* group actively utilised the OSG discussions in treatment decision making process to weight the available options with their outcomes. Thus, they tend to talk more about the decision factors. In contrast, paternalistic and autonomous groups either less comprehensive on

Table 6.6: The decision factors employed for the treatment decision against the decision making behaviour group.

	Paternalistic	Autonomous	Shared	Total	p value
	n (% in Paternalistic)	n (% in Autonomous)	n (% in Shared)	n (% in Total)	
Doctor skill discussed	200 (47.62)	2162 (55.68)	1837 (85.28)	4199 (65.03)	<0.01
Surgeon mentioned	219 (52.14)	1047 (26.96)	1730 (80.32)	2996 (46.4)	<0.01
Radiologist mentioned	51 (12.14)	186 (4.79)	649 (30.13)	886 (13.72)	<0.01
GP mentioned	12 (2.86)	51 (1.31)	165 (7.66)	228 (3.53)	<0.01
Medical indication	8 (1.9)	187 (4.82)	360 (16.71)	555 (8.6)	<0.01
Nerve mentioned	20 (4.76)	299 (7.7)	490 (22.75)	809 (12.53)	<0.01
Young age	6 (1.43)	110 (2.83)	265 (12.3)	381 (5.9)	<0.01
Bowel symptoms	0 (0)	18 (0.46)	42 (1.95)	60 (0.93)	<0.01
Urinary incontinence	20 (4.76)	301 (7.75)	507 (23.54)	828 (12.82)	<0.01
Erectile dysfunction	13 (3.1)	205 (5.28)	482 (22.38)	700 (10.84)	<0.01
Recovery time	3 (0.71)	77 (1.98)	184 (8.54)	264 (4.09)	<0.01
Best Cancer Control	22 (5.24)	331 (8.52)	587 (27.25)	940 (14.56)	<0.01
Financial	10 (2.38)	136 (3.5)	386 (17.92)	532 (8.24)	<0.01
Less Invasive	2 (0.48)	42 (1.08)	105 (4.87)	149 (2.31)	<0.01

decision factors or may not tend to engage on OSG discussions about the considered decision factors.

When comparing the autonomous and paternalistic groups, it can be seen that the paternalistic group has high mention rate of clinicians (surgeon, radiologist, and GP) than the autonomous group, as they tend to discuss more about clinicians that took the decision on behalf of them. In contrast, the autonomous group discuss relatively more on side effects aspects (e.g., nerve mentions, urinary, bowel and sexual symptoms) than the paternalistic groups as they tend to discuss about their decision factors more than the paternalistic group.

The results generated the PRIME framework strongly correlate with the existing studies on decision making behaviour (Charles et al., 1999) and decision factors (Berry et al., 2003; Gwede et al., 2005). All groups actively sought information on OSG: The Shared group provided consistent, prolonged interactions and contributed more as support providers. The Autonomous group only sought advice mainly around the treatment and subsequently contributed minimally to conversations on OSG.

6.6.4 Patient timeline analysis

It is important to analyse the patient timeline to understand the temporalities of the clinical events and emotions of the different behaviour groups. Such analysis enables to compare and contrast the characteristics of the behaviour groups.

Figure 6.8 presents the aggregated positive and negative emotions timelines of the three decision behaviour groups. Note that the aggregated emotions are obtained by accumulating all positive emotions (e.g., *open*, *alive*) into an aggregated positive emotion and accumulating all negative emotions (e.g., *afraid*, *hurt*) into an aggregated negative emotion.

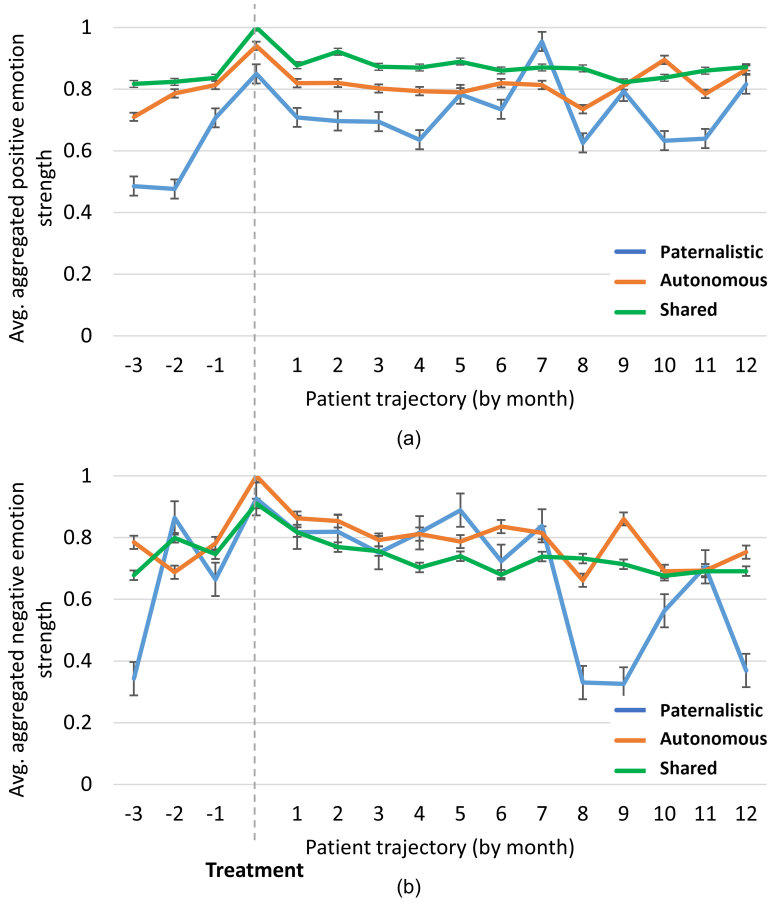


Figure 6.8: The positive and negative emotion trajectories of the three decision making behaviour groups over the patient trajectory from 3 months pre-treatment to 12 months post-treatment.

As shown in Figure 6.8 both positive and negative emotions peaked at the treatment month, which due to the elevated emotional state associated with a traumatic event like surgery. The *shared* group is the most positive and *autonomous* group is the most negative during the treatment month. After the treatment, there is a rapid drop in positive and negative emotions across all groups which stabilised in subsequent months. However, the *shared* group consistently expressive in positive emotions than the other two groups. The paternalistic group have shown high variance in the latter part of the timeline in both positive and negative emotions. It seems to be due to the drop of active *paternalistic*

participants as shown in Figure 6.7a resulting in the averaged emotions values impacted by a few emotional individuals.

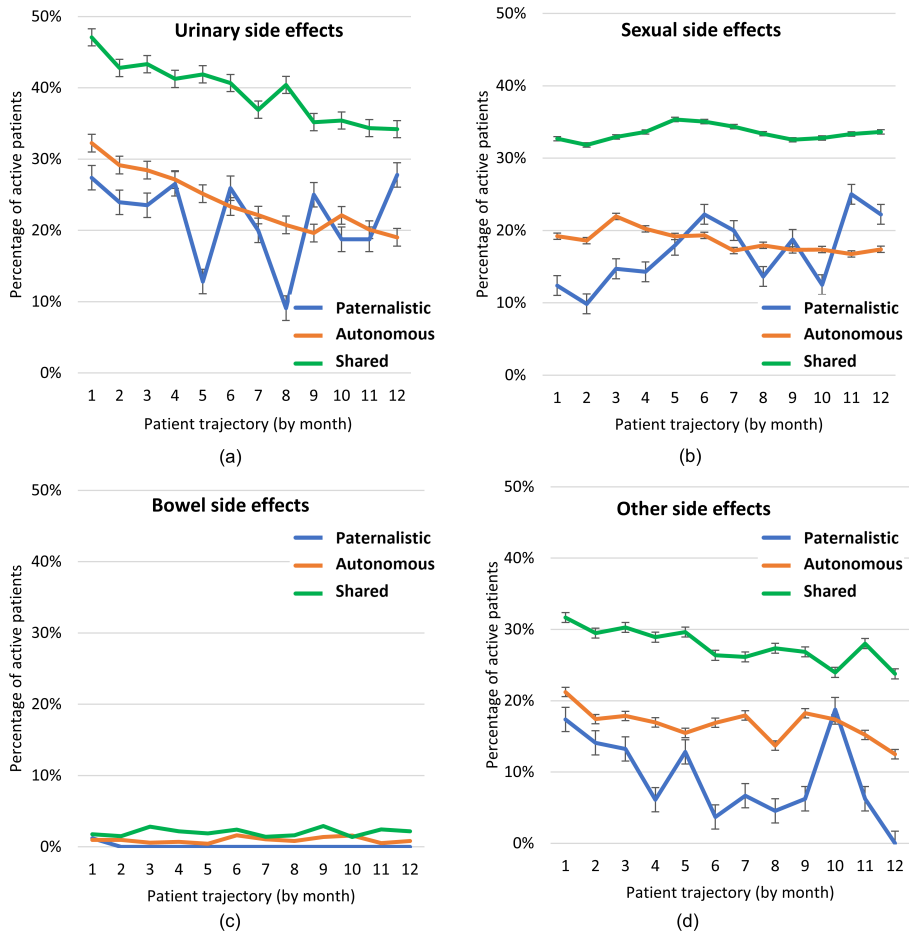


Figure 6.9: The self-disclosed side effect timeline aggregated monthly across the three decision making behaviour groups from 1st month post-treatment to 12 months post-treatment.

Although treatment decision making does not impact the occurrence of treatment related side-effects, it impacts the awareness and acceptance of the treatment related side-effects. Figure 6.9 presents the self-disclosed side effect timeline aggregated monthly across the three decision making behaviour groups. It highlights that *shared* group reports more on side effects than the other two groups across all side effect categories. *Shared* and *autonomous* groups are initially reported more urinary side effects (Figure 6.9a) which gradually declined over time, in contrast to the *paternalistic* group which has fluctuations over time. *shared* and *autonomous* groups are consistently reported sexual side effects (Figure 6.9b) while *paternalistic* group shows an increased reporting over time, reaching a level equivalent to that of the *autonomous* group by month 12. The bowel side effects (Figure 6.9c) are the least mentioned across all groups as they are relatively less prominent. This significantly more reporting of side-effects by the *shared* group is mainly due to their openness to discuss such issues and high collaborativeness in OSG. Similarly, the *paternalistic* group is less open and collaborative to discuss side-effects in OSG. However, it is interesting to observe the increase of the report of sexual side effects by the *paternalistic*

group over time. It may indicate that over time they become more open about the chronic sexual side effects and use OSG discussion to seek emotional support from OSG as a coping mechanism.

These results highlight the characteristics of different decision making behaviours, the decision factors considered, reported side effects and the emotions by the individuals. The multiplicity of the decision factors including both clinical, psychological and financial aspects highlights the complex nature of the treatment decision making which has deviated from strictly clinical outcome based to a more holistic approach. The *shared* group has shown to discuss significantly more decision factors indicating that they are more aware of decision factors and effectively use OSG discussions in the decision making process. The *shared* group reported more side effects. However, they have demonstrated a better quality of life (increased positive and decreased negative emotions) over time more than the other two groups. These characteristics of the shared group align with two coping strategies of traumatic events such as cancer. First is being more informed about the expected outcomes leads to an informed decision that develops acceptance to the outcomes of the treatment. The second is, discussing about a traumatic event among individuals with similar or worse outcomes leads to positive social comparisons (see Section 3.2.4) highlighting that the individual is not alone reduces the emotional distress (Taylor and Lobel, 1989).

These results also highlight the importance of the shared decision-making. It is argued that patients should be provided necessary tools to gather information, know their decision options, scenarios and consequences for shared decision-making to be effective (Elwyn et al., 2012). The significance of emotional support that allows patients to freely express values and preferences and ask questions without clinician obstruction is also highlighted (Stacey et al., 2011). The proliferation of OSG is a clear indication that patients and carers are bridging this gap by seeking (and providing) this service extraneous to healthcare providers and institutions. Further, OSG provide information, decision options and emotional support with the added advantage of a geographically dispersed community of individuals who are undergoing/have undergone similar circumstances.

6.7 Discussion

The above two case studies have shown the wealth of information encapsulated in OSG discussions. The first case-study investigated the positive and negative emotional expression of different groups and highlights several highly emotionally expressive groups as more in need of support. The second case-study explored different decision making behaviours and evaluated their decision factors, emotional expression and report of side-effects; and identified that *shared* group reports better emotional expression compared which could be attributed to their extensive consideration of decision factors and active role in decision making. Also, *shared* groups is pivotal for the continuation of OSG as they are more likely to stay longer with the OSG and become information/support providers over time. These insights help to shape-up optimum delivery of necessary care to the vulnerable prostate cancer patients/survivors and caregivers.

With the increasing presence of social media empowered patients, OSG make a paramount contribution as a medium for discussions on information exchange and emotional support. This is seen to be instrumental in addressing the 'out of sight out of mind' dilemma that arises due to periodic or occasional clinician consultations during the cancer journey. The participants who have undergone similar circumstances (e.g., similar treatment) are willing to share their experience, offer advice and emotional support to those who seek in OSG. Hence, the medical support network for care in any type of cancer should look into relevant OSG platforms in order to understand the unmet needs and provide optimal and individualised care that is clinically appropriate for patients with cancer.

As explicated in this study, the PRIME platform provides this required functionality to automatically analyse OSG discussions and transform into a multi-dimensional information source which can be explored to understand the unmet needs of prostate cancer. PRIME can be succinctly extended into any form of cancer by adapting the clinical information extraction layers of the platform to incorporate the clinical information of the respective cancer.

6.8 Chapter Summary

This chapter extends the multi-stage information structuring platform (presented in previous chapter) to capture insights specific to prostate cancer related online support groups (OSG). It structures and transforms prostate cancer related online support group discussions into a multidimensional representation based on demographics, emotions, clinical factors. This representation is employed to investigate the self disclosed quality of life against time, demographics and clinical factors. Moreover, it was used to assess different decision making behaviours and decision factors related to treatment decision making and their associated emotions over time pre- and post-treatment. These investigations have provided insights on the emotional expression of different groups and highlight several highly emotionally expressive groups as more in need of support. Also, differences in decision making behaviours and decision factors across distinct decision making groups. These insights help to shape-up optimum delivery of necessary care to the vulnerable prostate cancer patients/survivors and caregivers.

Chapter 7

Conclusion

Begin thus from the first act, and proceed; and, in conclusion, at the ill which thou hast done, be troubled, and rejoice for the good.

Pythagoras

This chapter concludes the thesis by presenting a summary of the research contributions, as well as directions for future work. Section 7.1 provides a summary of conceptual, technical and application contributions, Section 7.2 describes how those contributions have addressed the key research questions, and Section 7.3 concludes this chapter by providing future research directions arising from this thesis.

7.1 Summary of contributions

This thesis is an exploration to develop techniques that can be used to gain deeper insights from social data, which has yield theoretical, technical and application contributions.

As the conceptual contribution, a comprehensive literature study has been conducted on current state-of-the-art machine learning and natural language processing techniques employed to generate insights from social data in Chapter 2. This study is focused around three types of insights often captured from social data which are topics, events and emotions. Section 2.2 investigates current techniques used for topic extraction which include topics modelling and text clustering. Section 2.3 assessed two types of event detection techniques: specified event detection techniques to capture known event types using event related taxonomies and unspecified event detection techniques to capture unknown/unforeseen event types using various change indicators. Section 2.4 first reviews theories emotions and emotional models developed in social sciences and subsequently review how those models are incorporated into computational techniques to extract emotions from text. Finally, Section 2.7 provides an exhaustive analysis of the challenges and limitations of existing techniques in relation to gaining insights from social data.

As the second theoretical contribution, this thesis developed a layered conceptual framework for interpreting data generated in online social media platforms a.k.a. social data. This framework is presented in Chapter 3. It is built upon existing theories of cognition, social needs and social behaviour, which depicts that social data is generated

by social interactions happening in online social media platforms, and such interactions reflect different social behaviours driven by different social needs. This conceptual framework provides a deeper meaning to social data as archived traces of online social behaviours rather than just a corpus of unstructured text. In addition, four key social behaviours were conceptualised based on existing literature and a comprehensive comparison is provided comparing and contrasting exhibition of each behaviour in the physical world and online social media platforms.

On the technical front, two suites of algorithms were developed to harness insights from different types of online social media platforms. First is two unsupervised incremental learning algorithms presented in Chapter 4. These algorithms were designed to capture insights from fast-paced social data streams. The first is a new unsupervised incremental machine learning algorithm that automatically structures a social data stream into topics and extends those across time into topic pathways. This algorithm is developed based on GSOM self-structuring algorithm and IKASL incremental learning algorithm. It captures changes in topics over time as well as distinct new topics as new topic pathways at different points of time. Also, it was designed to overcome the challenges present in social data with respect to its brevity, unstructuredness, and diversity. The second algorithm is a multi faceted event detection algorithm developed to monitor topic pathways for significant changes in online social behaviours over time, and identify such changes as potential events of interest. The changes in social behaviours were identified using automatic event indicators such as changes in volume, positive sentiment and negative sentiment. These algorithms are presented in the journal publication entitled *Automatic event detection in microblogs using incremental machine learning* (Bandaragoda et al., 2017).

The second technical contribution is another a set of algorithms presented in Chapter 5, built with a suite of machine learning and natural language processing algorithms developed to structure slow-paced social data streams. These algorithms are specifically developed for online support groups which are social discussions related to health issues by patients and caregivers. These algorithms extract demographics, deeper emotions and clinical events from users self-disclosed information encapsulated in unstructured text. Subsequently, the extracted information is further structured by formulating patient event timelines based on clinical events and emotions with their associated time. This extracted information is used to build a multi-dimensional database that can be used to query relevant and reliable information for different use cases. Moreover, multiple aggregates of information can be obtained for research purposes. These algorithms are presented in journal publications *Text mining for personalized knowledge extraction from online support groups* (Bandaragoda, De Silva, Alahakoon, Ranasinghe and Bolton, 2018) and *Machine learning to support social media empowered patients in cancer care and cancer treatment decisions* (De Silva et al., 2018).

As application contributions, the above developed algorithms have been successfully trialled with social data streams from two online social media platforms. The suites of algorithms developed for fast-paced text streams has been tested with two Twitter datasets containing 6 months of tweets on two entities– a politician and an organisation.

As shown in the experiment results in Section 4.7, the topic pathway separation algorithm successfully captured the key topic pathways representing ongoing discussions related to the respective entities. Also, shifts in the discussions represented by new key terms were successfully learned and associated with the relevant topic pathway. Moreover, new distinct topics were automatically captured as new topics pathways. The event detection algorithm monitors those topic pathways and automatically captured significant changes in human behaviour using changes in volume and sentiment. Those captured events were aligned with contemporary news articles that discussed those events. These results are published in the journal publication entitled *Automatic event detection in microblogs using incremental machine learning* (Bandaragoda et al., 2017).

The second application contribution is based on the second suite of algorithms developed to structure slow-paced social data streams. It was applied to a large text corpus collected from two large online support groups (OSG) which consist of over 4 million posts. It successfully structures those posts pivoted by user id, automatically building user profiles and user timelines from the extracted information. The use of these algorithms is demonstrated using two uses cases where first shows how it can be used by patient/caregivers to look for relevant and reliable information based on different criteria, better than the standard methods available. Also, the researcher use case demonstrated that researchers can use this structure to capture aggregated information on different study cohorts which can be used to generate population level insights. These results are published in the journal publication entitled *Text mining for personalized knowledge extraction from online support groups* (Bandaragoda, De Silva, Alahakoon, Ranasinghe and Bolton, 2018).

The third application contribution is presented in Chapter 6, which is an extension to the researcher use case discussed above. In this application, the platform developed to structure slow-paced social data stream is extended to capture insights specific to prostate cancer related online support groups (OSG). It structures and transforms prostate cancer related online support group discussions into a multidimensional representation based on demographics, emotions, clinical factors and investigated the self disclosed quality of life against time, demographics and clinical factors. Moreover, it assessed different decision making behaviours and decision factors related to treatment decision making and their associated emotions over time pre- and post-treatment. These investigations have provided insights on the emotional expression of different groups and highlight several highly emotionally expressive groups as more in need of support. Also, differences in decision making behaviours and decision factors across distinct decision making groups. These insights help to shape-up optimum delivery of necessary care to the vulnerable prostate cancer patients/survivors and caregivers. These investigations were carried out in collaboration with researchers from a key urology institute and the results were published in three journals papers: (i) *The patient-reported information multidimensional exploration (PRIME) framework for investigating emotions and other factors of prostate cancer patients with low intermediate risk based on online cancer support group discussions* (Bandaragoda, Ranasinghe, Adikari, de Silva, Lawrentschuk, Alahakoon, Persad and Bolton, 2018), (ii) *Machine learning to support social media empowered patients in cancer care and cancer treatment*

decisions (De Silva et al., 2018), and (iii) *Robotic-assisted vs. open radical prostatectomy: A machine learning framework for intelligent analysis of patient-reported outcomes from online cancer support groups* (Ranasinghe et al., 2018).

7.2 Addressing the research questions

This section describes how the above contributions have addressed the research questions delineates in Chapter 1.

1. *What are the limitations of existing artificial intelligence algorithms and natural language processing techniques in the study of social interactions and social behaviours using representative online social data in digital environments?*

   This research question investigates applications of existing techniques on social data and consolidates key challenges and limitations faced in such applications. Existing machine learning and natural language processing techniques applied to social data is assessed in Chapter 2. This assessment captures key applications such as topic extraction, event detection, emotion extraction, self-structuring and incremental learning. It discussed different supervised and unsupervised state-of-the-art techniques and what are their limitations when applied to social data. Subsequently section 2.7 consolidates the key challenges of social data leading to the limitations in existing techniques.

2. *How can theories of social behaviour from social sciences contribute towards a conceptual model of enhanced understanding of social interactions in digital environments, as well as the representative online social data?*

   This research question is formulated to investigate how existing social theories on human social behaviour, social needs and cognition can be applied to social data. In order to address this research question, a conceptual framework is developed in Section 3.2 based on existing theories in social sciences to consider social data as representing the surface layer of a hierarchy of human social behaviour, social needs and cognition. This model assumes social data is generated as a result of social interactions that happens in online social media platforms, social behaviours are abstractions of social interactions, and cognition drives social interactions to achieve social needs. The Section 3.2 further explains four key behaviours that present in most of the social interactions and compared those behaviours in natural and digital environments. Based on this conceptual framework a platform is proposed in Section 3.3 which provides structure to unstructured social data for machine processing and also as a means for coupling established social theories with the new forms of digital social behaviour and interactions.

3. *How can new incremental machine learning algorithms, founded on the principles of self-structuring artificial intelligence, address the challenges of using social data to understand social behaviours?*

This research question is formulated to investigate how existing conventional machine learning and natural language processing techniques can be extended using the concepts of self-structuring AI and incremental machine learning developed to address question 1 can be employed to address the limitations identified from question 2. Two suites of algorithms specified in the previous section as technical contributions were developed to address this question. Both were developed by extending existing machine learning and natural language processing algorithms using the above concepts to overcome the identified challenges in social data and capture deeper insights.

The set of algorithms presented in Chapter 4 is a new platform developed by extending existing unsupervised techniques to self-structure and incrementally learn from fast-paced social data streams. As discussed in previous section, the first algorithm is capable of structuring a social data stream into coherent groups i.e., topics and extend them across time as topic pathways. It is specifically designed for the highly unstructured nature of the text present in fast-paced social data streams and is capable of handling sparse features due to brevity as well as dynamically changing vocabulary due to changes in social interactions. Also, a new multi-faceted event detection algorithm is developed to monitor changes in human behaviours inside topic pathways using multiple changes indicators.

The second suite of algorithms presented in Chapter 5 is designed to captured deeper insights from slow-paced social data streams. It consists of machine learning and natural language processing algorithms specifically designed to capture self-disclosed information from unstructured text corpus in online support groups. Those algorithms capture self-disclosed demographics, author role (patient, caregiver), clinical events and deeper emotions. The emotion capturing algorithm capture higher granular emotional expression using 16 emotions derived from Plutchik's emotions wheel. Further extensions added to these algorithms in Chapter 6 where algorithms were developed to capture treatment choice related decision making behaviours and decision factors.

4. *How can the algorithms developed in 3. advanced into technology platforms that deliver actionable insights for societal advancement?*

   This research question is formulated to investigate how the developed algorithms can be employed to develop technology platforms that are capable of generating actionable insights which can be used for societal advancements. These technology platforms were specifically designed to facilitate distinct and contrasting characteristics/challenges of different online social media platforms. As detailed in the previous section as application contributions, this thesis presents three applications where the developed platforms are employed to generate valuable insights. In these applications, large text corpora collected from different online social media platforms were processed using the developed platforms. In the first application, two datasets collected from Twitter are used to extract topic pathways and significant events using

the platform developed in Chapter 4. The second application is structuring information present as self-disclosed unstructured text in online support groups. This structured information provides relevant and reliable information to other users and aggregates pivoted by different variables to health researchers. Chapter 6 expands the researcher application into gain insights from prostate cancer related online support groups.

7.3 Future directions

The platforms and techniques developed in this thesis have addressed numerous problems of gaining insights from social data accumulated in online social media platforms, and have successfully applied to different applications. However, considering the scale of data and the amount of insights encapsulated, gaining insights from social data on underlying social behaviours is still in its infancy. As discussed, swaths of social data accumulated in social media platforms every day and people from all walks of life are rapidly embracing online social media platforms to perform online social interactions. Hence, there exists substantial challenges to overcome and massive potential to gain.

One of the key avenue for future research is the predictive aspect, where predictive techniques can be integrated into the developed self-learning platforms to predict future dynamics of the human behaviours. These predictions can be individualistic where predictions can be generated for an individual based on different layers of the conceptual model presented in this thesis. For instance, predictions can be generated related to the cognitive process of an individual to predict the state of the mind and transition of emotions overtime based on the past indicators of such transitions captured from social data. Such predictions are invaluable to manage chronic mental disorders such as depression and anxiety which spreads among the society at epidemic levels. A predictive model trained to capture warning signs of mental disorders from social data/social interactions can be deployed to social media platforms to identify individuals who are in need of support.

Also, predictions can be at group or population level to predict different group dynamics of behaviour. For instance, the event detection technique developed in Chapter 4 can be extended for event prediction where predictive models can be trained to learn transitions of event indicators to learn whether or when a discussion gains popularity and generate predictions on future popularity. Such predictions are useful in many applications such as in law-enforcement to predict potentially disruptive movements as well as in marketing to predict gain of a marketing campaign.

Another direction for future work is on how to extend this work to a fully automated system that can self-adapt and scale to generate insights from social media relevant to any domain. The self-structuring AI framework presented in Chapter 3 is generalisable to any domain. Also, the unsupervised self-structuring algorithms presented in Chapter 4 and 5 are scalable and applicable to any domain. However, some of the natural language processing and machine learning algorithms developed to gather insights are specifically developed and fine-tuned for the applications presented in the thesis. As future work, this

can be extended with a further layer of self-structuring and self-adaptive AI to make the insight generation fully automated.

References

Aaronson, N. K., Ahmedzai, S., Bergman, B., Bullinger, M., Cull, A., Duez, N. J., Filiberti, A., Flechtner, H., Fleishman, S. B., de Haes, J. C. et al. (1993). The european organization for research and treatment of cancer qlq-c30: a quality-of-life instrument for use in international clinical trials in oncology, *JNCI: Journal of the National Cancer Institute* **85**(5): 365–376.

Abel, F., Hauff, C., Houben, G.-J., Stronkman, R. and Tao, K. (2012). Twitcident: Fighting Fire with Information from Social Web Streams, *Proceedings of the 21st International Conference on World Wide Web*, ACM, pp. 305–308.

Adelman, R. D., Tmanova, L. L., Delgado, D., Dion, S. and Lachs, M. S. (2014). Caregiver burden: a clinical review, *Jama* **311**(10): 1052–1060.

Adler, P. S. and Kwon, S.-W. (2002). Social capital: Prospects for a new concept, *The Academy of Managemenet Review* **27**: 17–40.

Aggarwal, C. C., Han, J., Wang, J. and Yu, P. S. (2003). A framework for clustering evolving data streams, *Proceedings of the 29th international conference on Very large data bases-Volume 29*, VLDB Endowment, pp. 81–92.

Aggarwal, C. C., Hinneburg, A. and Keim, D. A. (2001). On the Surprising Behavior of Distance Metrics in High Dimensional Space, *International conference on database theory*, Springer, Berlin, Heidelberg, pp. 420–434.

Aggarwal, C. C. and Zhai, C. (2012a). *A Survey of Text Clustering Algorithms*.

Aggarwal, C. C. and Zhai, C. (2012b). *Mining text data*, Springer Science & Business Media.

Aiken, L. H., Clarke, S. P., Sloane, D. M., Sochalski, J. and Silber, J. H. (2002). Hospital nurse staffing and patient mortality, nurse burnout, and job dissatisfaction, *Jama* **288**(16): 1987–1993.

Ajzen, I. (1985). From intentions to actions: A theory of planned behavior, *Action control*, Springer, pp. 11–39.

Ajzen, I. (1991). The theory of planned behavior, *Organizational behavior and human decision processes* **50**(2): 179–211.

Alahakoon, D., Halgamuge, S. K. and Srinivasan, B. (2000). Dynamic Self-Organizing Maps with Controlled Growth for Knoledge Discovery, *IEEE Transactions on Neural Networks* **11**(3): 601—-614.

Albertsen, P. (2016). Who Should Consider Active Surveillance?, *The Journal of Urology* **196**(6): 1604–1605.

Alderfer, C. P. (1969). An empirical test of a new theory of human needs, *Organizational behavior and human performance* **4**(2): 142–175.

Alfano, C. M. and Rowland, J. H. (2006). Recovery issues in cancer survivorship: a new challenge for supportive care, *The Cancer Journal* **12**(5): 432–443.

Allan, J., Carbonell, J., Doddington, G., Yamron, J., Yang, Y. et al. (1998). Topic detection and tracking pilot study: Final report, *Proceedings of the DARPA broadcast news transcription and understanding workshop*, Vol. 1998, Citeseer, pp. 194–218.

Allan, J., Papka, R. and Lavrenko, V. (1998). On-line new event detection and tracking, *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 37–45.

Alm, C. O., Roth, D. and Sproat, R. (2005). Emotions from text: machine learning for text-based emotion prediction, *Proceedings of the conference on human language technology and empirical methods in natural language processing*, Association for Computational Linguistics, pp. 579–586.

AlSumait, L., Barbará, D. and Domeniconi, C. (2008). On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking, *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, IEEE, pp. 3–12.

Altman, I. and Taylor, D. A. (1973). *Social penetration: The development of interpersonal relationships.*, Holt, Rinehart & Winston.

American Cancer Society (2018). Cancer facts & figures 2018, *Technical report*.

Aronson, A. R. and Lang, F.-M. (2010). An overview of MetaMap: historical perspective and recent advances., *Journal of the American Medical Informatics Association : JAMIA* **17**(3): 229–36.

Aronson, A. R., Lang, F.-M., Aronson, A., Aronson, A., Rindflesch, T., Browne, A., Aronson, A., Rindflesch, T., Rindflesch, T., Aronson, A., Rindflesch, T., Aronson, A., Aronson, A., McCray, A., Aronson, A., Browne, A., Humphrey, S., Rogers, W., Kilicoglu, H., Aronson, A., Masseroli, M., Kilicoglu, H., Lang, F., Fiszman, M., Demner-Fushman, D., Lang, F., Ahlers, C., Fiszman, M., Demner-Fushman, D., Demner-Fushman, D., Humphrey, S., Ide, N., Weeber, M., Klein, H., Aronson, A., Aronson, A., Bodenreider, O., Chang, H., Kim, G., Aronson, A., Mork, J., Aronson, A., Mork, J., Gay, C., Gay, C., Kayaalp, M., Aronson, A., Neveol, A., Shooshan, S., Humphrey, S., Butte, A., Kohane, I., Hsieh, Y., Hardardottir, G., Brennan, P., Baud, R., Ruch, P., Gaudinat, A.,

Yetisgen-Yildiz, M., Pratt, W., Zhou, L., Srinivasan, P., Niu, Y., Zhu, X., Li, J., Ruiz, M., Mougin, F., Burgun, A., Bodenreider, O., McInnes, B., Pedersen, T., Carlis, J., Pakhomov, S., Shah, N., Hanson, P., Fan, J., Friedman, C., Patterson, O., Hurdle, J., Lowe, H., Barnett, G., Miller, R., Gieszczykiewicz, F., Vries, J., Evans, D., Ginther-Webster, K., Hart, M., Hersh, W., Greenes, R., Nadkarni, P., Chen, R., Brandt, C., Denny, J., Smithers, J., Miller, R., Srinivasan, P., Harman, D., Kim, W., Aronson, A., Wilbur, W., Shah, N., Bhatia, N., Jonquet, C., Schwartz, A., Hearst, M., Humphreys, B., McCray, A., Cheh, M., Aronson, A., Bodenreider, O., Demner-Fushman, D., Chapman, W., Bridewell, W., Hanbury, P., Shooshan, S., Aronson, A., Weeber, M., Mork, J., Aronson, A., Aronson, A., Mork, J., Neveol, A., Wilbur, W., Hazard, G. and Divita, G. (2010). An overview of MetaMap: historical perspective and recent advances., *Journal of the American Medical Informatics Association : JAMIA* **17**(3): 229–36.

Asur, S. and Huberman, B. A. (2010). Predicting the future with social media, pp. 492–499.

Atefeh, F. and Khreich, W. (2015). A survey of techniques for event detection in twitter, *Computational Intelligence* **31**(1): 132–164.

Attik, M., Bougrain, L. and Alexandre, F. (2005). Self-organizing map initialization, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 3696 LNCS, pp. 357–362.

Augusto, J. C. (2005). Temporal reasoning for decision support in medicine, *Artificial Intelligence in Medicine* **33**(1): 1–24.

Baldridge, J. (2005). The opennlp project, *URL: http://opennlp. apache. org/index. html,(accessed 2 February 2012)* p. 1.

Baldwin, B. (1997). CogNIAC: High precision coreference with limited knowledge and linguistic resources, *Proceedings of the ACL Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Text, Madrid, Spain, July 1997* pp. 38–45.

Baldwin, T., Cook, P., Lui, M., Mackinlay, A. and Wang, L. (2013). How Noisy Social Media Text , How Diffrnt Social Media Sources ?, *Proc. IJCNLP 2013*, number October, pp. 356–364.

Bandaragoda, T. R., De Silva, D. and Alahakoon, D. (2017). Automatic event detection in microblogs using incremental machine learning, *Journal of the Association for Information Science and Technology* **68**(10): 2394–2411.

Bandaragoda, T. R., De Silva, D., Alahakoon, D., Ranasinghe, W. and Bolton, D. (2018). Text mining for personalized knowledge extraction from online support groups, *Journal of the Association for Information Science and Technology* **69**(12): 1446–1459.

Bandaragoda, T., Ranasinghe, W., Adikari, A., de Silva, D., Lawrentschuk, N., Alahakoon, D., Persad, R. and Bolton, D. (2018). The Patient-Reported Information Multidimensional Exploration (PRIME) Framework for investigating emotions and other factors

of prostate cancer patients with low intermediate risk based on online cancer support group discussions, *Annals of surgical oncology* **25**(6): 1737–1745.

Banerjee, S., Ramanathan, K. and Gupta, A. (2007). Clustering short texts using wikipedia, *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '07*, ACM Press, New York, New York, USA, p. 787.

Bar-Lev, S. (2008). "We are here to give you emotional support": Performing emotions in an online HIV/AIDS support group, *Qualitative Health Research* **18**(4): 509–521.

Barabasi, A.-L. (2005). The origin of bursts and heavy tails in human dynamics, *Nature* **435**(7039): 207.

Barocas, D. A., Alvarez, J., Resnick, M. J., Koyama, T., Hoffman, K. E., Tyson, M. D., Conwill, R., Mccollum, D., Cooperberg, M. R., Goodman, M., Greenfield, S., Hamilton, A. S., Hashibe, M., Kaplan, S. H., Paddock, L. E., Stroup, A. M., Wu, X.-C. and Penson, D. F. (2017). Association Between Radiation Therapy, Surgery, or Observation for Localized Prostate Cancer and Patient-Reported Outcomes After 3 Years, *JAMA* **317**(11): 1126–1140.

Bazarova, N. N. and Choi, Y. H. (2014). Self-disclosure in social media: Extending the functional approach to disclosure motivations and characteristics on social network sites, *Journal of Communication* **64**(4): 635–657.

Beck, A. T., Steer, R. A. and Brown, G. K. (1996). Beck depression inventory-ii, *San Antonio* **78**(2): 490–498.

Becker, H., Iter, D., Naaman, M. and Gravano, L. (2012). Identifying content for planned events across social media sites, *Proceedings of the fifth ACM international conference on Web search and data mining*, ACM, pp. 533–542.

Becker, H., Naaman, M. and Gravano, L. (2011). Beyond Trending Topics: Real-World Event Identification on Twitter., *Proceedings of the Fifth International Conference on Weblogs and Social Media (ICWSM)*, pp. 1–17.

Bender, J. L. (2011). *The Web of Care : A Multi-Method Study Examining the Role of Online Communities as a Source of Peer-to-Peer Supportive Care for Breast Cancer Survivors By The Web of Care : A Multi-Method Study Examining the Role of Online Communities as a Source of Peer-*, PhD thesis.
URL: *https://tspace.library.utoronto.ca/handle/1807/31690*

Berger, C. R. and Bradac, J. J. (1982). *Language and social knowledge: Uncertainty in interpersonal relations*, Vol. 2, Hodder Education.

Berger, C. R. and Calabrese, R. J. (1975). Some explorations in initital interaction and beyod: Toward a developmental theory of interpersonal communication, *Human Communciation Research* **1**(2): 99–112.

Berry, D. L., Ellis, W. J., Woods, N. F., Schwien, C., Mullen, K. H. and Yang, C. (2003). Treatment decision-making by men with localized prostate cancer: The influence of personal factors, *Urologic Oncology: Seminars and Original Investigations* **21**(2): 93–100.

Bethard, S., Derczynski, L., Savova, G., Pustejovsky, J. and Verhagen, M. (2015). SemEval-2015 Task 6: Clinical TempEval, *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pp. 806–814.

Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U., Beeri, C. and Buneman, P. (1999). When Is "Nearest Neighbor" Meaningful?, *Database TheoryICDT'99* **1540**: 217–235.

Blank, T. O. and Bellizzi, K. M. (2008). A gerontologic perspective on cancer and aging. **URL:** *http://doi.wiley.com/10.1002/cncr.23444*

Blank, T. O., Schmidt, S. D., Vangsness, S. A., Monteiro, A. K. and Santagata, P. V. (2010). Differences among breast and prostate cancer online support groups, *Computers in Human Behavior* **26**(6): 1400–1404.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent dirichlet allocation, *Journal of machine Learning research* **3**(Jan): 993–1022.

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology, *Nucleic Acids Research* **32**(90001): 267D–270.

Bollen, J., Mao, H. and Pepe, A. (2011). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena, *Fifth International AAAI Conference on Weblogs and Social Media*.

Bollen, J., Mao, H. and Zeng, X. (2011). Twitter mood predicts the stock market, *Journal of computational science* **2**(1): 1–8.

Bourdieu, P. (2011). The forms of capital.(1986), *Cultural theory: An anthology* **1**: 81–93.

Boureau, Y.-L., Ponce, J. and LeCun, Y. (2010). A theoretical analysis of feature pooling in visual recognition, *International Conference on Machine Learning (ICML)*, pp. 111–118.

Bowlby, J. (1969). *Attachment, Vol. 1 of Attachment and loss*, New York: Basic Books.

Bowling, A. (1995). What things are important in people's lives? A survey of the public's judgements to inform scales of health related quality of life, *Social Science and Medicine* **41**(10): 1447–1462.

Boyd, D. and Ellison, N. B. (2007). Social Network Sites: Definition, History, and Scholarship, *Journal of Computer-Mediated Communication* **13**: 210–230.

Boyd, R., Gintis, H. and Bowles, S. (2010). Coordinated punishment of defectors sustains cooperation and can proliferate when rare, *Science* **328**(5978): 617–620.

Boyd, R., Gintis, H., Bowles, S. and Richerson, P. J. (2003). The evolution of altruistic punishment, *Proceedings of the National Academy of Sciences* **100**(6): 3531–3535.

Boyd, R. and Richerson, P. J. (2009). Culture and the evolution of human cooperation, *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **364**(1533): 3281–3288.

Bradley, M. M., Greenwald, M. K., Petry, M. C. and Lang, P. J. (1992). Remembering pictures: pleasure and arousal in memory., *Journal of experimental psychology: Learning, Memory, and Cognition* **18**(2): 379.

Bradley, M. M. and Lang, P. J. (1999). Affective norms for english words (anew): Instruction manual and affective ratings, *Technical report*, Citeseer.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A. and Jemal, A. (2018). Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries, *CA: a cancer journal for clinicians* .

Bretherton, I. (1992). The origins of attachment theory: John bowlby and mary ainsworth., *Developmental psychology* **28**(5): 759.

Brody, S. and Diakopoulos, N. (2011). Cooooooooooooooollllllllllllll !!!!!!!!!!!!!! Using Word Lengthening to Detect Sentiment in Microblogs *** Preprint Version, *Computational Linguistics* (2010): 562–570.

Brooks, M., Kuksenok, K., Torkildson, M. K., Perry, D., Robinson, J. J., Scott, T. J., Anicello, O., Zukowski, A., Harris, P. and Aragon, C. R. (2013). Statistical affect detection in collaborative chat, *Proceedings of the 2013 conference on Computer supported cooperative work*, ACM, pp. 317–328.

Buhrmester, M., Kwang, T. and Gosling, S. D. (2011). Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data?, *Perspectives on psychological science* **6**(1): 3–5.

Burke, M., Kraut, R. and Marlow, C. (2011). Social capital on facebook: Differentiating uses and users, *Proceedings of the SIGCHI conference on human factors in computing systems*, ACM, pp. 571–580.

Bury, M. (1982). Chronic illness as biographical disruption, *Sociology of health & illness* **4**(2): 167–182.

Cannon, W. B. (1927). The james-lange theory of emotions: A critical examination and an alternative theory, *The American journal of psychology* **39**(1/4): 106–124.

Carlson, J. R. and Zmud, R. W. (1999). Channel expansion theory and the experiential nature of media richness perceptions, *Academy of Management Journal* **42**(2): 153–170.

Carpenter, G. A. and Grossberg, S. (1988). The ART of Adaptive Pattern Recognition by a Self-Organizing Neural Network, *Computer* **21**(3): 77–88.

Carr, A. J. and Higginson, I. J. (2001). Are quality of life measures patient centred?, *BMJ (Clinical research ed.)* **322**(7298): 1357–60.

Cecchini, M., Aytug, H., Koehler, G. J. and Pathak, P. (2010). Making words work: Using financial text as a predictor of financial events, *Decision Support Systems* **50**(1): 164–175.

Cella, D. F., Tulsky, D. S., Gray, G., Sarafian, B., Linn, E., Bonomi, A., Silberman, M., Yellen, S. B., Winicour, P., Brannon, J. et al. (1993). The functional assessment of cancer therapy scale: development and validation of the general measure, *J Clin Oncol* **11**(3): 570–579.

Chang, A. X. and Manning, C. D. (2012). SUTime: A library for recognizing and normalizing time expressions., *LREC*, number iii, pp. 3735–3740.

Charles, C., Gafni, A. and Whelan, T. (1999). Decision-making in the physician-patient encounter: Revisiting the shared treatment decision-making model, *Social Science and Medicine* **49**(5): 651–661.

Chen, E. S., Hripcsak, G., Xu, H., Markatou, M. and Friedman, C. (2008). Automated acquisition of disease–drug knowledge from biomedical and clinical documents: an initial study, *Journal of the American Medical Informatics Association* **15**(1): 87–98.

Chen, R. C., Basak, R., Meyer, A.-M., Kuo, T.-M., Carpenter, W. R., Agans, R. P., Broughman, J. R., Reeve, B. B., Nielsen, M. E., Usinger, D. S., Spearman, K. C., Walden, S., Kaleel, D., Anderson, M., Stürmer, T. and Godley, P. A. (2017). Association Between Choice of Radical Prostatectomy, External Beam Radiotherapy, Brachytherapy, or Active Surveillance and Patient-Reported Quality of Life Among Men With Localized Prostate Cancer, *JAMA* **317**(11): 1141–1150.

Cheng, N., Chandramouli, R. and Subbalakshmi, K. (2011). Author gender identification from text, *Digital Investigation* **8**(1): 78–88.

Chikersal, P., Poria, S. and Cambria, E. (2015). Sentu: sentiment analysis of tweets by combining a rule-based classifier with supervised learning, *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pp. 647–651.

Chiu, C., Hsu, M. and Wang, E. (2006). Understanding knowledge sharing in virtual communities: An integration of social capital and social cognitive theories, *Decision Support Systems* **42**(3): 1872–1888.

Cho, J. H., Liao, V. Q., Jiang, Y. and Schatz, B. R. (2013). Aggregating Personal Health Messages for Scalable Comparative Effectiveness Research, *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics - BCB'13*, number September 2013, pp. 907–916.

Chou, H.-T. G. and Edge, N. (2012). they are happier and having better lives than i am: the impact of using facebook on perceptions of others' lives, *Cyberpsychology, Behavior, and Social Networking* **15**(2): 117–121.

Clark, J. A., Wray, N. P. and Ashton, C. M. (2001). Living with treatment decisions: regrets and quality of life among men treated for metastatic prostate cancer, *Journal of Clinical Oncology* **19**(1): 72–80.

Cole, J. et al. (2017). Surveying the digital future, *UCLA Center for Communication Policy, University of California–Los Angeles* .

Collins, R. L. (1996). For better or worse: The impact of upward social comparison on self-evaluations., *Psychological bulletin* **119**(1): 51.

Compas, B. E., Stoll, M. F., Thomsen, A. H., Oppedisano, G., EppingJordan, J. E. and Krag, D. N. (1999). Adjustment to breast cancer: agerelated differences in coping and emotional distress, *Breast Cancer Research and Treatment* **54**(3): 195–203.

Cooperberg, M. R. and Carroll, P. R. (2015). Trends in management for patients with localized prostate cancer, 1990-2013, *Jama* **314**(1): 80–82.

Cortes, C. and Vapnik, V. (1995). Support-vector networks, *Machine learning* **20**(3): 273–297.

Council, N. R. et al. (2005). *From cancer patient to cancer survivor: lost in transition*, National Academies Press.

Cozby, P. C. (1973). Self-disclosure: a literature review., *Psychological bulletin* **79**(2): 73.

Cutting, D. R., Karger, D. R., Pedersen, J. O. and Tukey, J. W. (1992). Scatter/gather: A cluster-based approach to browsing large document collections, *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 318–329.

Da Silva, N. F., Hruschka, E. R. and Hruschka Jr, E. R. (2014). Tweet sentiment analysis with classifier ensembles, *Decision Support Systems* **66**: 170–179.

Daft, R. L. and Lengel, R. (1986). Organizational information requirements, media richness and structural design, *Management Science* **32**(5): 554–571.

Daft, R. L., Lengel, R. H. and Trevino, L. K. (1987). Message equivocality, media selection, and manager performance: Implications for information systems, *MIS quarterly* pp. 355–366.

Dalva, M. and Katz, L. (1994). Rearrangements of synaptic connections in visual cortex revealed by laser photostimulation, *Science* **265**(5169): 255–258.

D'Amico, A., Whittington, R., Malkowicz, S. B., Schultz, D., Blank, K., Broderick, G. A., Tomaszewski, J. E., Renshaw, A. A., Kaplan, I., Beard, C. J. and Wein, A. (1998). Biochemical outcome after radical prostatectomy, external beam radiation therapy, or interstitial radiation therapy for clinically localized prostate cancer., *Jama* **280**(11): 969–74.

D'Andrea, E., Ducange, P., Lazzerini, B. and Marcelloni, F. (2015). Real-time detection of traffic from twitter stream analysis, *IEEE transactions on intelligent transportation systems* **16**(4): 2269–2283.

Darwin, C. (1859). On the origin of species by means of natural selection, *Murray, London* .

Darwin, C. (1872). The expression of the emotions in man and animal.

Das, S. R. and Chen, M. Y. (2007). Yahoo! for amazon: Sentiment extraction from small talk on the web, *Management science* **53**(9): 1375–1388.

Davis, K. (2012). Friendship 2.0: Adolescents' experiences of belonging and self-disclosure online, *Journal of adolescence* **35**(6): 1527–1536.

De Haes, J., Van Knippenberg, F. and Neijt, J. (1990). Measuring psychological and physical distress in cancer patients: structure and application of the rotterdam symptom checklist, *British journal of cancer* **62**(6): 1034.

De Silva, D. (2010). *A Cognitive Approach to Autonomous Incremental Learning*, PhD thesis, Monash University.

De Silva, D. and Alahakoon, D. (2010). Incremental knowledge acquisition and self learning from text, *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, pp. 1—-8.

De Silva, D., Ranasinghe, W., Bandaragoda, T., Adikari, A., Mills, N., Iddamalgoda, L., Alahakoon, D., Lawrentschuk, N., Persad, R., Osipov, E. et al. (2018). Machine learning to support social media empowered patients in cancer care and cancer treatment decisions, *PloS one* **13**(10): e0205855.

Demner-Fushman, D., Chapman, W. W. and McDonald, C. J. (2009). What can natural language processing do for clinical decision support?, *Journal of biomedical informatics* **42**(5): 760–772.

Denberg, T. D., Melhado, T. V. and Steiner, J. F. (2006). Patient treatment preferences in localized prostate carcinoma: The influence of emotion, misconception, and anecdote, *Cancer* **107**(3): 620–630.

Derks, D., Bos, A. E. and Von Grumbkow, J. (2007). Emoticons and social interaction on the internet: the importance of social context, *Computers in human behavior* **23**(1): 842–849.

Derlaga, V. J. and Berg, J. H. (1987). *Self-disclosure: Theory, research, and therapy*, Springer Science & Business Media.

Donovan, J. L., Hamdy, F. C., Lane, J. A., Mason, M., Metcalfe, C., Walsh, E., Blazeby, J. M., Peters, T. J., Holding, P., Bonnington, S., Lennon, T., Bradshaw, L., Cooper, D., Herbert, P., Howson, J., Jones, A., Lyons, N., Salter, E., Thompson, P., Tidball,

S., Blaikie, J., Gray, C., Bollina, P., Catto, J., Doble, A., Doherty, A., Gillatt, D., Kockelbergh, R., Kynaston, H., Paul, A., Powell, P., Prescott, S., Rosario, D. J., Rowe, E., Davis, M., Turner, E. L., Martin, R. M. and Neal, D. E. (2016). Patient-Reported Outcomes after Monitoring, Surgery, or Radiotherapy for Prostate Cancer, *New England Journal of Medicine* **375**(15): 1425–1437.

Drost, E. A. et al. (2011). Validity and reliability in social science research, *Education Research and perspectives* **38**(1): 105.

Dunbar, R. I. (1998). The social brain hypothesis, *Evolutionary Anthropology* **6**(5): 178–190.

Dunbar, R. I. M. (1993). Coevolution of neocortical size, group size and language in humans, *Behavioral and Brain Sciences* **16**(04): 681.

Dunbar, R. I. M. and Shultz, S. (2007). Evolution in the Social Brain, *Science* **317**(5843): 1344–1347.

Dunbar, R. I., Marriott, A. and Duncan, N. D. (1997). Human conversational behavior, *Human nature* **8**(3): 231–246.

Eisenstein, J. (2013). What to do about bad language on the internet, *NAACL-HLT 2013*, pp. 359–369.

Eisenstein, J., O'Connor, B., Smith, N. A. and Xing, E. P. (2014). Diffusion of lexical change in social media, *PLoS ONE* **9**(11).

Ekman, P. (1992). An argument for basic emotions, *Cognition & emotion* **6**(3-4): 169–200.

Ekman, P. and Oster, H. (1979). Facial expressions of emotion, *Annual review of psychology* **30**(1): 527–554.

Ellison, N. B., Steinfield, C. and Lampe, C. (2007). The benefits of facebook friends: social capital and college students use of online social network sites, *Journal of Computer-Mediated Communication* **12**(4): 1143–1168.

Elwyn, G., Frosch, D., Thomson, R., Joseph-Williams, N., Lloyd, A., Kinnersley, P., Cording, E., Tomson, D., Dodd, C., Rollnick, S. et al. (2012). Shared decision making: a model for clinical practice, *Journal of general internal medicine* **27**(10): 1361–1367.

Emanuel, E. J. and Emanuel, L. L. (1992). Four models of the physician-patient relationship, *Jama* **267**(16): 2221–2226.

Ertöz, L., Steinbach, M. and Kumar, V. (2003). Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data, *Proceedings of Second SIAM International Conference on Data Mining*, pp. 47–59.

Ester, M., Kriegel, H.-P., Sander, J. and Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, *KDD*, pp. 226–231.

Esuli, A. and Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining., *LREC*, Vol. 6, Citeseer, pp. 417–422.

Eysenbach, G. (2008). Medicine 2.0: Social networking, collaboration, participation, apomediation, and openness.
URL: *http://www.jmir.org/article/viewFile/jmir_v10i3e22/2*

Farnstrom, F., Lewis, J. and Elkan, C. (2000). Scalability for clustering algorithms revisited, *SIGKDD explorations* **2**(1): 51–57.

Fast, E., Chen, B. and Bernstein, M. (2016). Empath: Understanding Topic Signals in Large-Scale Text, *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, ACM Press, New York, New York, USA, pp. 4647–4657.

Fehr, E. and Fischbacher, U. (2004). Social norms and human cooperation, *Trends in cognitive sciences* **8**(4): 185–190.

Festinger, L. (1954). A theory of social comparison processes, *Human relations* **7**(2): 117–140.

Finkel, J. R., Grenager, T. and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling, *Proceedings of the 43rd annual meeting on association for computational linguistics*, Association for Computational Linguistics, pp. 363–370.

Fox, S. (2011a). Peer-to-peer Health Care — Pew Research Center.
URL: *http://www.pewinternet.org/2011/02/28/peer-to-peer-health-care-2/*

Fox, S. (2011b). *The social life of health information, 2011*, Pew Internet & American Life Project Washington, DC.

Fox, S. (2011c). *The Social Life of Health Information*, Pew Internet \& American Life Project Washington, DC.

Fox, S. (2013). After dr google: peer-to-peer health care, *Pediatrics* **131**(Supplement 4): S224–S225.

French, R. M. (1999). Catastrophic forgetting in connectionist networks, *Trends in cognitive sciences* **3**(4): 128–135.

Friedenberg, J. and Silverman, G. (2011). *Cognitive science: An introduction to the study of mind*, Sage.

Friedman, C., Alderson, P. O., Austin, J. H., Cimino, J. J. and Johnson, S. B. (1994). A general natural-language text processor for clinical radiology, *Journal of the American Medical Informatics Association* **1**(2): 161–174.

Fritzke, B. (1995a). Growing grida self-organizing network with constant neighborhood range and adaptation strength, *Neural processing letters* **2**(5): 9–13.

Fritzke, B. (1995b). A growing neural gas network learns topologies, *Advances in neural information processing systems*, pp. 625–632.

Frost, J. H. and Massagli, M. P. (2008). Social uses of personal health information within PatientsLikeMe, an online patient community: What can happen when patients have access to one another's data, *Journal of Medical Internet Research* **10**(3): 15–3.

Fung, G. P. C., Yu, J. X., Yu, P. S. and Lu, H. (2005). Parameter free bursty events detection in text streams, *Proceedings of the 31st international conference on Very large data bases*, VLDB Endowment, pp. 181–192.

Furao, S. and Hasegawa, O. (2006). An incremental network for on-line unsupervised classification and topology learning, *Neural Networks* **19**(1): 90–106.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M. and Bouchachia, A. (2014). A survey on concept drift adaptation, *ACM computing surveys (CSUR)* **46**(4): 44.

Given, B. A., Given, C. W. and Kozachik, S. (2001). Family support in advanced cancer, *CA: a cancer journal for clinicians* **51**(4): 213–231.

Glaeser, E. L., Laibson, D., Sacerdote, B., The, S., Journal, E. and Nov, F. (2002). An Economic Approach to Social Capital, *The Economic Journal* **112**(483): F437–F458.

Gleason, D. F., Mellinger, G. T., Arduino, L. J., Bailar III, J. C., Becker III, L. E., Berman III, H. I., Bischoff, A. J., Byar, D. P., Blackard, C. E., Doe, R. P. et al. (1974). Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging, *The Journal of urology* **111**(1): 58–64.

Golder, S. A. and Macy, M. W. (2014). Digital footprints: Opportunities and challenges for online social research, *Annual Review of Sociology* **40**: 129–152.

Gooden, R. J. and Winefield, H. R. (2007). Breast and prostate cancer online discussion boards: A thematic analysis of gender differences and similarities, *Journal of Health Psychology* **12**(1): 103–114.

Granovetter, M. (1973). The Strength of Weak Ties, *The American Journal of Sociology* **78**(6): 1360–1380.

Greene, K., Derlega, V. J. and Mathews, A. (2006). Self-disclosure in personal relationships, *The Cambridge handbook of personal relationships* pp. 409–427.

Grishman, R. and Sundheim, B. (1996). Message understanding conference-6: A brief history, *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, Vol. 1.

Groth-Marnat, G. (2009). *Handbook of psychological assessment*, John Wiley & Sons.

Gupta, S., Maclean, D. L., Heer, J. and Manning, C. D. (2014). Induced lexico-syntactic patterns improve information extraction from online medical forums, *Journal of the American Medical Informatics Association* **21**(5): 902—-909.

Gwede, C. K., Pow-Sang, J., Seigne, J., Heysek, R., Helal, M., Shade, K., Cantor, A. and Jacobsen, P. B. (2005). Treatment decision-making strategies and influences in patients with localized prostate carcinoma, *Cancer* **104**(7): 1381–1390.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I. H. (2009). The WEKA data mining software, *ACM SIGKDD Explorations Newsletter* **11**(1): 10.

Hamdy, F. C., Donovan, J. L., Lane, J. A., Mason, M., Metcalfe, C., Holding, P., Davis, M., Peters, T. J., Turner, E. L., Martin, R. M., Oxley, J., Robinson, M., Staffurth, J., Walsh, E., Bollina, P., Catto, J., Doble, A., Doherty, A., Gillatt, D., Kockelbergh, R., Kynaston, H., Paul, A., Powell, P., Prescott, S., Rosario, D. J., Rowe, E. and Neal, D. E. (2016). 10-Year Outcomes after Monitoring, Surgery, or Radiotherapy for Localized Prostate Cancer, *New England Journal of Medicine* pp. 1–10.

Hamilton, W. D. (1964). The genetical evolution of social behaviour. ii, *Journal of theoretical biology* **7**(1): 17–52.

Hamilton, W. L., Clark, K., Leskovec, J. and Jurafsky, D. (2016). Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora.

Harvey, K. J., Brown, B., Crawford, P., Macfarlane, A. and McPherson, A. (2007). 'Am I normal?' Teenagers, sexual health and the internet, *Social Science and Medicine* **65**(4): 771–781.

Haykin, S. (1994). *Neural networks: a comprehensive foundation*, Prentice Hall PTR.

He, Q., Chang, K. and Lim, E.-P. (2007). Analyzing feature trajectories for event detection, *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 207–214.

Hebb, D. O. (1949). *The Organization of Behavior*.

Hebb, D. O. (2005). *The organization of behavior: A neuropsychological theory*, Psychology Press.

Herdadelen, A. and Baroni, M. (2011). Stereotypical gender actions can be extracted from web text, *Journal of the American Society for Information Science and Technology* **62**(9): 1741–1749.

Herrmann, E., Call, J., Hernández-Lloreda, M. V., Hare, B. and Tomasello, M. (2007). Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis, *science* **317**(5843): 1360–1366.

Hill, C. E. and Knox, S. (2001). Self-disclosure., *Psychotherapy: Theory, Research, Practice, Training* **38**(4): 413.

Hobbs, J. R. and Pan, F. (2006). Time ontology in owl, *W3C working draft* **27**: 133.

Hofmann, T. (1999). Probabilistic latent semantic indexing, *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 50–57.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis, *Machine learning* **42**(1-2): 177–196.

Hong, L. and Davison, B. (2010). Empirical study of topic modeling in twitter, *Proceedings of the First Workshop on Social Media Analytics*, pp. 80–88.

Høybye, M. T., Johansen, C. and Tjørnhøj-Thomsen, T. (2005). Online interaction. Effects of storytelling in an internet breast cancer support group, *Psycho-Oncology* **14**(3): 211–220.

Hu, X., Bell, R. A., Kravitz, R. L. and Orrange, S. (2012). The Prepared Patient: Information Seeking of Online Support Group Members Before Their Medical Appointments, *Journal of Health Communication* **17**(8): 960–978.

Hu, X. and Liu, H. (2012). Text analytics in social media, *in* C. C. Aggarwal and C. Zhai (eds), *Mining Text Data*, Springer US, Boston, MA, pp. 385–414.

Hu, X., Sun, N., Zhang, C. and Chua, T.-S. (2009). Exploiting internal and external semantics for the clustering of short texts using world knowledge, *Proceedings of the 18th ACM conference on Information and knowledge management*, ACM, pp. 919–928.

Huber, J., Ihrig, A., Peters, T., Huber, C. G., Kessler, A., Hadaschik, B., Pahernik, S. and Hohenfellner, M. (2011). Decision-making in localized prostate cancer: Lessons learned from an online support group, *BJU International* **107**(10): 1570–1575.

Huber, J., Maatz, P., Muck, T., Keck, B., Friederich, H. C., Herzog, W. and Ihrig, A. (2017). The effect of an online support group on patients treatment decisions for localized prostate cancer: An online survey, *Urologic Oncology: Seminars and Original Investigations* **35**(2): 37.e19–37.e28.

Hutto, C. J. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text, *Eighth International AAAI Conference on Weblogs and ...*, San Francisco, CA, USA, pp. 216–225.

Ihrig, A., Keller, M., Hartmann, M., Debus, J., Pfitzenmaier, J., Hadaschik, B., Hohenfellner, M., Herzog, W. and Huber, J. (2011a). Treatment decision-making in localized prostate cancer: why patients chose either radical prostatectomy or external beam radiation therapy, *BJU international* **108**(8): 1274–1278.

Ihrig, A., Keller, M., Hartmann, M., Debus, J., Pfitzenmaier, J., Hadaschik, B., Hohenfellner, M., Herzog, W. and Huber, J. (2011b). Treatment decision-making in localized prostate cancer: Why patients chose either radical prostatectomy or external beam radiation therapy, *BJU International* **108**(8): 1274–1278.

Izard, C. E. (1993). Four systems for emotion activation: Cognitive and noncognitive processes., *Psychological review* **100**(1): 68.

Jansen, B. J., Zhang, M., Sobel, K. and Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth, *Journal of the American Society for Information Science and Technology* **60**(11): 2169–2188.

Java, A., Song, X., Finin, T. and Tseng, B. (2007). Why We Twitter: Understanding Microblogging Usage and Communities, *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis - WebKDD/SNA-KDD '07*, ACM Press, New York, New York, USA, pp. 56–65.

Jayles, B., Kim, H.-r., Escobedo, R., Cezera, S., Blanchet, A., Kameda, T., Sire, C. and Theraulaz, G. (2017). How social information can improve estimation accuracy in human groups, *Proceedings of the National Academy of Sciences* **114**(47): 201703695.

Jensen, P. B., Jensen, L. J. and Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care., *Nature Reviews Genetics* **13**(6): 395–405.

Joachims, T. (1996). A probabilistic analysis of the rocchio algorithm with tfidf for text categorization., *Technical report*, Carnegie-mellon univ pittsburgh pa dept of computer science.

Jourard, S. M. and Lasakow, P. (1958). Some factors in self-disclosure., *The Journal of Abnormal and Social Psychology* **56**(1): 91.

Jurafsky, D. (2000). *Speech & language processing*, Pearson Education India.

Kaas, J. H., Nelson, R. J., Sur, M., Lin, C.-S. and Merzenich, M. M. (1979). Multiple representations of the body within the primary somatosensory cortex of primates, *Science* **204**(4392): 521–523.

Kahai, S. S. and Cooper, R. B. (2003). Exploring the core concepts of media richness theory: The impact of cue multiplicity and feedback immediacy on decision quality, *Journal of Management Information Systems* **20**(1): 263–299.

Kahneman, D. and Riis, J. (2012). Living, and thinking about it: Two perspectives on life, *The Science of Well-Being*.

Kahneman, D. et al. (1999). Objective happiness, *Well-being: The foundations of hedonic psychology* **3**: 25.

Kaplan, A. M. and Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media, *Business Horizons* **53**(1): 59–68.

Kemp, S. (2019). 2019 Q4 Global Digital Statshot Report.

Kertesz, A. (1983). *Localization in neuropsychology*, Academic Pr.

Kietzmann, J. H., Hermkens, K., McCarthy, I. P. and Silvestre, B. S. (2011). Social media? Get serious! Understanding the functional building blocks of social media, *Business Horizons* **54**(3): 241–251.

Kim, M. Y., Xu, Y., Zaiane, O. and Goebel, R. (2013). Patient information extraction in noisy tele-health texts, *Proceedings - 2013 IEEE International Conference on Bioinformatics and Biomedicine, IEEE BIBM 2013*, IEEE, pp. 326–329.

Kim, Y. and Schulz, R. (2008). Family caregivers' strains: comparative analysis of cancer caregiving with dementia, diabetes, and frail elderly caregiving, *Journal of Aging and Health* **20**(5): 483–503.

Kleinberg, J. (2003). Bursty and hierarchical structure in streams, Vol. 7, Springer, pp. 373–397.

Knox, S., Hess, S. A., Petersen, D. A. and Hill, C. E. (1997). A qualitative analysis of client perceptions of the effects of helpful therapist self-disclosure in long-term therapy., *Journal of counseling psychology* **44**(3): 274.

Kock, N. (2004). The Psychobiological Model: Towards a New Theory of Computer-Mediated Communication Based on Darwinian Evolution, *Organization Science* **15**(3): 327–348.

Kock, N. (2005). Media richness or media naturalness? The evolution of our biological communication apparatus and its influence on our behavior toward e-communication tools, *IEEE Transactions on Professional Communication* **48**(2): 117–130.

Kohonen, T. (1990). The self-organizing map, *Proceedings of the IEEE* **78**(9): 1464–1480.

Kohonen, T. (1997). Exploration of very large databases by self-organizing maps, *Proceedings of International Conference on Neural Networks (ICNN'97)*, Vol. 1, IEEE, pp. PL1–PL6.

Kohonen, T. (2013). Essentials of the self-organizing map, *Neural networks* **37**: 52–65.

Kohonen, T., Kaski, S., Lagus, K., Salojarvi, J., Honkela, J., Paatero, V. and Saarela, A. (2000). Self organization of a massive document collection, *IEEE transactions on neural networks* **11**(3): 574–585.

Kononenko, O., Baysal, O., Holmes, R. and Godfrey, M. W. (2014). Mining modern repositories with elasticsearch, *Proceedings of the 11th Working Conference on Mining Software Repositories - MSR 2014*, pp. 328–331.

Kraut, R., Patterson, M., Lundmark, V., Kiesler, S., Mukophadhyay, T. and Scherlis, W. (1998). Internet paradox: A social technology that reduces social involvement and psychological well-being?, *American psychologist* **53**(9): 1017.

Kring, A. M. and Gordon, A. H. (1998). Sex differences in emotion: Expression, experience, and physiology., *Journal of Personality and Social Psychology* **74**(3): 686–703.

Kripke, S. A. (1972). Naming and necessity, *Semantics of natural language*, Springer, pp. 253–355.

Kummervold, P. E., Gammon, D., Bergvik, S., Johnsen, J. A. K., Hasvold, T. and Rosenvinge, J. H. (2002). Social support in a wired world: Use of online mental health forums in Norway, *Nordic Journal of Psychiatry* **56**(1): 59–65.

Kurtz, L. F. (1990). The self-help movement: Review of the past decade of research, *Social Work with Groups* **13**(3): 101–115.

Kvåle, R., Auvinen, A., Adami, H.-O., Klint, Å., Hernes, E., Møller, B., Pukkala, E., Storm, H. H., Tryggvadottir, L., Tretli, S. et al. (2007). Interpreting trends in prostate cancer incidence and mortality in the five nordic countries, *Journal of the National Cancer Institute* **99**(24): 1881–1887.

Kwak, H., Lee, C., Park, H. and Moon, S. (2010). What is Twitter, a social network or a news media?, *Proceedings of the 19th international conference on World wide web*, pp. 591–600.

Lafferty, J. D., McCallum, A. and Pereira, F. C. N. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 282–289.

Lapidot-Lefler, N. and Barak, A. (2012). Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition, *Computers in human behavior* **28**(2): 434–443.

Lappin, S. and Leass, H. J. (1994). An Algorithm for Pronominal Anaphora Resolution, *Computational Linguistics* **20**(4): 535–561.

Larsen, B. and Aone, C. (1999). Fast and effective text mining using linear-time document clustering, *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 16–22.

Lau, J. H., Collier, N. and Baldwin, T. (2012a). On-line trend analysis with topic models: #twitter trends detection topic model online, *Proceedings of COLING 2012*, The COLING 2012 Organizing Committee, pp. 1519–1534.

Lau, J. H., Collier, N. and Baldwin, T. (2012b). On-line trend analysis with topic models:\# twitter trends detection topic model online, *Proceedings of COLING 2012* pp. 1519–1534.

Lau, R. Y., Liao, S. S., Wong, K.-F. and Chiu, D. K. (2012). Web 2.0 environmental scanning and adaptive decision support for business mergers and acquisitions, *MIS quarterly* **36**(4): 1239–1268.

Lazarus, R. S. and Lazarus, R. S. (1991). *Emotion and adaptation*, Oxford University Press on Demand.

Li, Y., Qian, M., Jin, D., Hui, P. and Vasilakos, A. V. (2015). Revealing the efficiency of information diffusion in online social networks of microblog, *Information Sciences* **293**: 383–389.

Lin, N. (2002). *Social capital: A theory of social structure and action*, Vol. 19, Cambridge university press.

Lintz, K., Moynihan, C., Steginga, S., Norman, A., Eeles, R., Huddart, R., Dearnaley, D. and Watson, M. (2003). Prostate cancer patients' support and psychological care needs: survey from a non-surgical oncology clinic, *Psycho-Oncology: Journal of the Psychological, Social and Behavioral Dimensions of Cancer* **12**(8): 769–783.

Liu, B. and Zhang, L. (2012). A survey of opinion mining and sentiment analysis, *Mining text data*, Springer, pp. 415–463.

Liu, L., Wu, J., Li, P. and Li, Q. (2015). A social-media-based approach to predicting stock comovement, *Expert Systems with Applications* **42**(8): 3893–3901.

Liu, Y., Xu, S., Yoon, H.-J. and Tourassi, G. (2014). Extracting patient demographics and personal medical information from online health forums., *AMIA Annual Symposium Proceedings, AMIA Symposium.*, Vol. 2014, pp. 1825–1834.

Lloyd, S. (1982). Least squares quantization in pcm, *IEEE transactions on information theory* **28**(2): 129–137.

Lotan, G., Graeff, E., Ananny, M., Gaffney, D., Pearce, I. and Boyd, D. (2011). The Arab Spring— The Revolutions Were Tweeted: Information Flows during the 2011 Tunisian and Egyptian Revolutions.
URL: *http://ijoc.org/index.php/ijoc/article/view/1246*

Lovejoy, K., Waters, R. D. and Saxton, G. D. (2012). Engaging stakeholders through twitter: How nonprofit organizations are getting more out of 140 characters or less, *Public Relations Review* **38**(2): 313–318.

Luxford, K., Safran, D. G. and Delbanco, T. (2011). Promoting patient-centered care: a qualitative study of facilitators and barriers in healthcare organizations with a reputation for improving the patient experience, *International Journal for Quality in Health Care* **23**(5): 510–515.

Ma, Z., Pant, G. and Sheng, O. R. (2011). Mining competitor relationships from online news: A network-based approach, *Electronic Commerce Research and Applications* **10**(4): 418–427.

Mahdisoltani, F., Biega, J. and Suchanek, F. M. (2013). YAGO3: A knowledge base from multilingual wikipedias.

Mahendiran, A., Wang, W., Arredondo, J., Lira, S., Huang, B., Getoor, L., Mares, D. and Ramakrishnan, N. (2014). Discovering Evolving Political Vocabulary in Social Media,

Behavior, Economic and Social Computing (BESC), 2014 International Conference on, IEEE, pp. 1–7.

Manne, S., Ostroff, J., Winkel, G., Goldstein, L., Fox, K. and Grana, G. (2004). Posttraumatic growth after breast cancer: patient, partner, and couple perspectives., *Psychosomatic medicine* **66**(17): 442–454.

Manning, C. D., Raghavan, P. and Schütze, H. (2008). *Introduction to Information Retrieval*, Cambridge University Press, New York, NY, USA.

Manovich, L. (2011). Trending: The promises and the challenges of big social data, *Debates in the digital humanities* **2**: 460–475.

Mao, Y., Wei, W., Wang, B. and Liu, B. (2012). Correlating s&p 500 stocks with twitter data, *Proceedings of the first ACM international workshop on hot topics on interdisciplinary social networks research*, ACM, pp. 69–72.

Marr, D. and Thach, W. T. (1991). A theory of cerebellar cortex, *From the Retina to the Neocortex*, Springer, pp. 11–50.

Maslow, A. (1962). *Toward a psychology of being*, Vol. 214, J. Wiley & Sons.

Maslow, A. H. (1943). A theory of human motivation, *Psychological Review* **50**(4): 370–396.

Mathioudakis, M. and Koudas, N. (2010). Twittermonitor: trend detection over the twitter stream, *SIGMOD '10 Proceedings of the 2010 ACM SIGMOD International Conference on Management of data* pp. 1155–1158.

Mazanderani, F., Locock, L. and Powell, J. (2012). Being differently the same: The mediation of identity tensions in the sharing of illness experiences, *Social Science and Medicine* **74**(4): 546–553.

McClelland, D. C. (1967). *Achieving society*, Vol. 92051, Simon and Schuster.

McClelland, D. C. (1987). *Human motivation*, CUP Archive.

McFarland, L. A. and Ployhart, R. E. (2015). Social media: A contextual framework to guide research and practice, *Journal of Applied Psychology* **100**(6): 1653–1677.

Medhat, W., Hassan, A. and Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey, *Ain Shams engineering journal* **5**(4): 1093–1113.

Mehrotra, R., Sanner, S., Buntine, W. and Xie, L. (2013). Improving LDA topic models for microblogs via tweet pooling and automatic labeling, *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval-SIGIR '13*, ACM Press, New York, New York, USA, pp. 889—-892.

Melville, P., Gryc, W. and Lawrence, R. D. (2009). Sentiment analysis of blogs by combining lexical knowledge with text classification, *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 1275–1284.

Meshi, D., Tamir, D. I. and Heekeren, H. R. (2015). The Emerging Neuroscience of Social Media.
URL: *https://www.sciencedirect.com/science/article/pii/S1364661315002284#bib0375*

Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C. and Hurdle, J. F. (2008). Extracting information from textual documents in the electronic health record: a review of recent research, *Yearbook of medical informatics* **17**(01): 128–144.

Middleton, S. E., Middleton, L. and Modafferi, S. (2014). Real-time crisis mapping of natural disasters using social media, *IEEE Intelligent Systems* **29**(2): 9–17.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J. (2013). Distributed representations of words and phrases and their compositionality, *in* C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger (eds), *Advances in Neural Information Processing Systems 26*, Curran Associates, Inc., pp. 3111–3119.

Mikolov, T., Yih, W.-t. and Zweig, G. (2013). Linguistic regularities in continuous space word representations, *Proceedings of NAACL-HLT*, number June, pp. 746–751.

Miller, G. A. (1995). WordNet: a lexical database for English, *Communications of the ACM* **38**(11): 39–41.

Miller, K., Siegel, R. and Jemal, A. (2016). Cancer treatment & survivorship facts & figures 2016-2017, *Atlanta, GA: American Cancer Society* .

Mimno, D., Wallach, H. M., Talley, E., Leenders, M. and McCallum, A. (2011). Optimizing semantic coherence in topic models, *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, number 2, Association for Computational Linguistics, pp. 262–272.

Miotto, R., Li, L., Kidd, B. A. and Dudley, J. T. (2016). Deep patient: an unsupervised representation to predict the future of patients from the electronic health records, *Scientific reports* **6**: 26094.

Mishne, G., Glance, N. S. et al. (2006). Predicting movie sales from blogger sentiment., *AAAI spring symposium: computational approaches to analyzing weblogs*, pp. 155–158.

Mishra, M. V., Bennett, M., Vincent, A., Lee, O. T., Lallas, C. D., Trabulsi, E. J., Gomella, L. G., Dicker, A. P. and Showalter, T. N. (2013). Identifying barriers to patient acceptance of active surveillance: content analysis of online patient communications., *PloS one* **8**(9): e68563.

Mitkov, R. (2014). *Anaphora Resolution.*, Taylor and Francis.

Mnih, A. and Teh, Y. W. (2012). A Fast and Simple Algorithm for Training Neural Probabilistic Language Models, *Proceedings of the 29th International Conference on Machine Learning (ICML'12)*, pp. 1751–1758.

Mohammad, S. M. (2016). Sentiment Analysis: Detecting Valence, Emotions, and Other Affected States from Text, *Emotion Measurement*, pp. 201–237.

Mohammad, S. M. and Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon, *Computational Intelligence* **29**(3): 436–465.

Mosher, C. E. and Danoff-Burg, S. (2005). A review of age differences in psychological adjustment to breast cancer., *Journal of psychosocial oncology* **23**(2-3): 101–14.

Munezero, M. D., Montero, C. S., Sutinen, E. and Pajunen, J. (2014). Are they different? affect, feeling, emotion, sentiment, and opinion detection in text, *IEEE transactions on affective computing* **5**(2): 101–111.

Murdoch, T. B. and Detsky, A. S. (2013). The inevitable application of big data to health care, *JAMA* **309**(13): 1351–1352.

Murray, N. and Perronnin, F. (2014). Generalized Max Pooling, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2473–2480.

Naaman, M., Boase, J. and Lai, C.-H. (2010). Is it really about me?: message content in social awareness streams, *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, ACM, pp. 189–192.

Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification, *Lingvisticae Investigationes* **30**(1): 3–26.

Nahapiet, J. and Ghoshal, S. (2000). Social capital, intellectual capital, and the organizational advantage, *Knowledge and social capital*, Elsevier, pp. 119–157.

Naik, A., Bogart, C. and Rose, C. (2017). Extracting personal medical events for user timeline construction using minimal supervision, *BioNLP 2017* pp. 356–364.

Nanton, V., Appleton, R., Loew, J., Ahmed, N., Ahmedzai, S. and Dale, J. (2018). Men don't talk about their health, but will they chat? the potential of online holistic needs assessment in prostate cancer, *BJU international* **121**(4): 494–496.

Nass, C. and Brave, S. (2005). *Wired for speech: How voice activates and advances the human-computer relationship.*, MIT press.

National Alliance for Caregiving (2016). Cancer caregiving in the us: An intense, episodic, and challenging care experience.

Nelson, S. J., Zeng, K., Kilbourne, J., Powell, T. and Moore, R. (2011). Normalized names for clinical drugs: RxNorm at 6 years., *Journal of the American Medical Informatics Association : JAMIA* **18**(4): 441–8.

Nesi, J. and Prinstein, M. J. (2015). Using social media for social comparison and feedback-seeking: Gender and popularity moderate associations with depressive symptoms, *Journal of abnormal child psychology* **43**(8): 1427–1438.

Neuman, W. L. (2013). *Social research methods: Qualitative and quantitative approaches*, Pearson education.

Nguyen, Q. (2012). *Detecting Experience Revealing Sentences in Product Reviews*, PhD thesis, University of Amsterdam.
URL: *https://esc.fnwi.uva.nl/thesis/centraal/files/f1006996408.pdf*

Nie, N. H., Hillygus, D. S. and Erbring, L. (2002). Internet use, interpersonal relations, and sociability, *The Internet in everyday life* pp. 215–243.

Niolon, R. (n.d.). List of Feeling Words.
URL: *http://www.psychpage.com/learning/library/assess/feelings.html*

Northouse, L. L., Katapodi, M. C., Song, L., Zhang, L. and Mood, D. W. (2010). Interventions with family caregivers of cancer patients: meta-analysis of randomized trials, *CA: a cancer journal for clinicians* **60**(5): 317–339.

Northouse, L. L., Mood, D. W., Montie, J. E., Sandler, H. M., Forman, J. D., Hussain, M., Pienta, K. J., Smith, D. C., Sanda, M. G. and Kershaw, T. (2007). Living with prostate cancer: Patients' and spouses' psychosocial status and quality of life, *Journal of Clinical Oncology* **25**(27): 4171–4177.

O 'connor, B., Balasubramanyan, R., Routledge, B. R. and Smith, N. A. (2010). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series, *ICWSM*, pp. 122–129.

Oh, S. (2012). The characteristics and motivations of health answerers for sharing information, knowledge, and experiences in online environments, *Journal of the American Society for Information Science and Technology* **63**(3): 543–557.

Oja, M., Kaski, S. and Kohonen, T. (2003). Bibliography of self-organizing map (som) papers: 1998-2001 addendum, *Neural computing surveys* **3**(1): 1–156.

Osborne, M., Petrovic, S., McCreadie, R., Macdonald, C. and Ounis, I. (2012). Bieber no more: First story detection using twitter and wikipedia, *Sigir 2012 workshop on time-aware information access*.

Pak, A. and Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining., *Lrec* pp. 1320–1326.

Paltoglou, G. (2016). Sentiment-based event detection in Twitter, *Journal of the Association for Information Science and Technology* **67**(7): 1576–1587.

Pan, B., Zheng, Y., Wilkie, D. and Shahabi, C. (2013). Crowd sensing of traffic anomalies based on human mobility and social media, *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, ACM, pp. 344–353.

Pang, B. and Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales, *Proceedings of the 43rd annual meeting on association for computational linguistics*, Association for Computational Linguistics, pp. 115–124.

Pang, B., Lee, L. and Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques, *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP '02*, Vol. 10, Association for Computational Linguistics, Morristown, NJ, USA, pp. 79–86.

Pang, B., Lee, L. et al. (2008). Opinion mining and sentiment analysis, *Foundations and Trends® in Information Retrieval* **2**(1–2): 1–135.

Park, K. C., Jeong, Y. and Myaeng, S.-H. H. (2010). Detecting Experiences from Weblogs, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (July): 1464–1472.

Paul, M. J. and Dredze, M. (2011). You are what you tweet: Analyzing twitter for public health, *Fifth International AAAI Conference on Weblogs and Social Media*.

Peng, X., Wang, L., Wang, X. and Qiao, Y. (2015). Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice, *Computer Vision and Image Understanding* **150**: 109–125.

Pennebaker, J. W., Francis, M. E. and Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001, *Mahway: Lawrence Erlbaum Associates* **71**(2001): 2001.

Petrović, S., Osborne, M. and Lavrenko, V. (2012). Using paraphrases for improving first story detection in news and twitter, *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies*, Association for Computational Linguistics, pp. 338–346.

Pew Research Center (2019). Social media fact sheet.
URL: *https://www.pewresearch.org/internet/fact-sheet/social-media*

Phuvipadawat, S. and Murata, T. (2010). Breaking news detection and tracking in Twitter, *Proceedings - 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Workshops, WI-IAT 2010* pp. 120–123.

Plutchik, R. (1980a). *Emotion: A Psychoevolutionary Synthesis*, Harpercollins College Division.

Plutchik, R. (1980b). *Emotion: A psychoevolutionary synthesis*, Harpercollins College Division.

Plutchik, R. (1991). *The emotions*, University Press of America.

Plutchik, R. (2001). The nature of emotions, *American scientist* **89**(4): 344–350.

Polikar, R., Udpa, L., Udpa, S. S. and Honavar, V. (2001). Learn++: An incremental learning algorithm for supervised neural networks, *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews* **31**(4): 497–508.

Popescu, A.-M. and Pennacchiotti, M. (2010). Detecting controversial events from twitter, *Proceedings of the 19th ACM international conference on Information and knowledge management*, ACM, pp. 1873–1876.

Posner, J., Russell, J. A. and Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology, *Development and psychopathology* **17**(3): 715–734.

Preece, J. (1999). Empathic communities: Balancing emotional and factual communication, *Interacting with computers* **12**(1): 63–77.

Pudrovska, T. (2010). What makes you stronger: age and cohort differences in personal growth after cancer., *Journal of health and social behavior* **51**(3): 260–273.

Putnam, R. D. (2001). *Bowling alone: The collapse and revival of American community*, Simon and Schuster.

Putnam, R. D. and Feldstein, L. (2009). *Better together: Restoring the American community*, Simon and Schuster.

Qiu, L., Kan, M.-Y. and Chua, T.-S. (2004). A Public Reference Implementation of the RAP Anaphora Resolution Algorithm, *Proceedings of the 4th Language Resources and Evaluation Conference (LREC 2004)* **2**(Lrec): 291–294.

Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M. et al. (2018). Scalable and accurate deep learning with electronic health records, *NPJ Digital Medicine* **1**(1): 18.

Ramakrishnan, N., Korkmaz, G., Kuhlman, C., Marathe, A., Zhao, L., Hua, T., Chen, F., Lu, C. T., Huang, B., Srinivasan, A., Trinh, K., Butler, P., Getoor, L., Katz, G., Doyle, A., Ackermann, C., Zavorin, I., Ford, J., Summers, K., Fayed, Y., Arredondo, J., Gupta, D., Muthiah, S., Mares, D., Self, N., Khandpur, R., Saraf, P., Wang, W., Cadena, J. and Vullikanti, A. (2014). 'Beating the news' with EMBERS, *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14*, number February, pp. 1799–1808.

Ranasinghe, W., de Silva, D., Bandaragoda, T., Adikari, A., Alahakoon, D., Persad, R., Lawrentschuk, N. and Bolton, D. (2018). Robotic-assisted vs. open radical prostatectomy: A machine learning framework for intelligent analysis of patient-reported outcomes from online cancer support groups, *Urologic Oncology: Seminars and Original Investigations* **36**(12): 529–e1.

Ratliff, F. (1965). *Mach bands: quantitative studies on neural networks*, Vol. 2, Holden-Day, San Francisco London Amsterdam.

Reale, R. A. and Imig, T. J. (1980). Tonotopic organization in auditory cortex of the cat, *Journal of Comparative Neurology* **192**(2): 265–291.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora, *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, ELRA, Valletta, Malta, pp. 45–50. `http://is.muni.cz/publication/884893/en`.

Resnick, P. et al. (2001). Beyond bowling together: Sociotechnical capital, *HCI in the New Millennium* **77**: 247–272.

Riloff, E., Patwardhan, S. and Wiebe, J. (2006). Feature subsumption for opinion analysis, *Proceedings of the 2006 conference on empirical methods in natural language processing*, Association for Computational Linguistics, pp. 440–448.

Ritter, A., Clark, S., Etzioni, O. et al. (2011). Named entity recognition in tweets: an experimental study, *Proceedings of the conference on empirical methods in natural language processing*, Association for Computational Linguistics, pp. 1524–1534.

Robinson, K. M. (2001). Unsolicited narratives from the internet: A rich source of qualitative data, *Qualitative Health Research* **11**(5): 706–714.

Robinson, M. D. and Clore, G. L. (2002). Belief and feeling: evidence for an accessibility model of emotional self-report., *Psychological bulletin* **128**(6): 934.

Röder, M., Both, A. and Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures, *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15*, ACM Press, New York, New York, USA, pp. 399–408.

Rosen-Zvi, M., Chemudugunta, C., Griffiths, T., Smyth, P. and Steyvers, M. (2010). Learning author-topic models from text corpora, *ACM Transactions on Information Systems (TOIS)* **28**(1): 4.

Rosen-Zvi, M., Griffiths, T., Steyvers, M. and Smyth, P. (2004). The author-topic model for authors and documents, *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, AUAI Press, pp. 487–494.

Rosenfeld, P., Culbertson, A. L. and Magnusson, P. (1992). Human needs: A literature review and cognitive life span model, *Technical report*, NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER SAN DIEGO CA.

Roth, A. J., Kornblith, A. B., Batel-Copel, L., Peabody, E., Scher, H. I. and Holland, J. C. (1998). Rapid screening for psychologic distress in men with prostate carcinoma: a pilot study, *Cancer: Interdisciplinary International Journal of the American Cancer Society* **82**(10): 1904–1908.

Rubin, Z. (1975). Disclosing oneself to a stranger: Reciprocity and its limits, *Journal of Experimental Social Psychology* **11**(3): 233–260.

Rui, H., Liu, Y. and Whinston, A. (2013). Whose and what chatter matters? the effect of tweets on movie sales, *Decision Support Systems* **55**(4): 863–870.

Russell, J. A. (1980). A circumplex model of affect., *Journal of personality and social psychology* **39**(6): 1161.

Sadovykh, V., Sundaram, D. and Piramuthu, S. (2015). Do decision-making structure and sequence exist in health online social networks?, *Decision Support Systems* **74**: 102–120.

Sakaki, T., Okazaki, M. and Matsuo, Y. (2010a). Earthquake shakes Twitter users: real-time event detection by social sensors, *WWW '10: Proceedings of the 19th international conference on World wide web* p. 851.

Sakaki, T., Okazaki, M. and Matsuo, Y. (2010b). Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors, *Proceedings of the 19th international conference on World wide web (WWW '10)*, pp. 851–860.

Salton, G. and Buckley, C. (1988a). Term-weighting approaches in automatic text retrieval, *Information processing & management* **24**(5): 513–523.

Salton, G. and Buckley, C. (1988b). Term-weighting approaches in automatic text retrieval, *Information Processing and Management* **24**(5): 513–523.

Salton, G., Wong, A. and Yang, C.-S. (1975). A vector space model for automatic indexing, *Communications of the ACM* **18**(11): 613–620.

Sankaranarayanan, J., Teitler, B. E., Samet, H., Lieberman, M. D. and Sperling, J. (2009). TwitterStand : News in Tweets, *Information Storage and Retrieval* **156**: 42–51.

Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C. and Chute, C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications, *Journal of the American Medical Informatics Association* **17**(5): 507–513.

Schachter, S. and Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state., *Psychological review* **69**(5): 379.

Schacter, D. L., Norman, K. A. and Koutstaal, W. (1998). The cognitive neuroscience of constructive memory, *Annual review of psychology* **49**(1): 289–318.

Schram, W. E. (1954). The process and effects of mass communication.

Schramm, W. (1954). How communication works, *The process and effects of mass communication* pp. 3–26.

Schütze, H. and Silverstein, C. (1997). Projections for efficient document clustering, *ACM SIGIR Forum*, Vol. 31, ACM, pp. 74–81.

Seale, C. (2006). Gender accommodation in online cancer support groups., *Health (London, England : 1997)* **10**(3): 345–360.

Seale, C., Charteris-Black, J., MacFarlane, A. and McPherson, A. (2010). Interviews and Internet Forums: A Comparison of Two Sources of Qualitative Data, *Qualitative Health Research* **20**(5): 595–606.

Shannon, C. E. (1948). A mathematical theory of communication, *The Bell System Technical Journal* **27**(3): 379–423.

Shepherd, M. D., Schoenberg, M., Slavich, S., Wituk, S., Warren, M. and Meissen, G. (1999). Continuum of professional involvement in self-help groups, *Journal of Community Psychology* **27**(1): 39–53.

Shouse, E. (2005). Feeling, emotion, affect, *M/c journal* **8**(6): 26.

Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H. and Deng, X. (2013). Exploiting topic based twitter sentiment for stock prediction, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 24–29.

Siegel, R. L., Miller, K. D. and Jemal, A. (2018). Cancer statistics, 2018, *CA: A Cancer Journal for Clinicians* **68**(1): 7–30.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank, *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval, *Journal of documentation* **28**(1): 11–21.

Späth, H. (1980). *Cluster analysis algorithms for data reduction and classification of objects*, Horwood.

Sprecher, S. and Hendrick, S. S. (2004). Self-disclosure in intimate relationships: Associations with individual and relationship characteristics over time, *Journal of Social and Clinical Psychology* **23**(6): 857–877.

Stacey, D., Cl, B., Mj, B., Nf, C., Kb, E., Lyddiatt, A., Légaré, F., Thomson, R., Stacey, D., Bennett, C. L., Barry, M. J., Col, N. F., Eden, K. B., Holmes-rovner, M. and Llewellyn, H. (2011). Decision aids for people facing health treatment or screening decisions., *Cochrane Database of Systematic Reviews* **10**(10).

Standage, T. (2013). *Writing on the wall: Social media-The first 2,000 years*, Bloomsbury Publishing USA.

Steinbach, M., Karypis, G., Kumar, V. et al. (2000). A comparison of document clustering techniques, *KDD workshop on text mining*, Vol. 400, Boston, pp. 525–526.

Stone, P. J., Dunphy, D. C. and Smith, M. S. (1966). The General Inquirer: A Computer Approach to Content Analysis.

Strapparava, C. and Mihalcea, R. (2007). Semeval-2007 task 14: Affective text, *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pp. 70–74.

Strötgen, J. and Gertz, M. (2010). Heideltime: High quality rule-based extraction and normalization of temporal expressions, *Proceedings of the 5th International Workshop on ...* pp. 321–324.

Suler, J. (2004). The online disinhibition effect, *Cyberpsychology & behavior* **7**(3): 321–326.

Suls, J., Martin, R. and Wheeler, L. (2002). Social comparison: Why, with whom, and with what effect?, *Current directions in psychological science* **11**(5): 159–163.

Sun, W., Rumshisky, A. and Uzuner, O. (2013). Evaluating temporal relations in clinical text: 2012 i2b2 Challenge.
URL: *https://academic.oup.com/jamia/article-lookup/doi/10.1136/amiajnl-2013-001628*

Szreter, S. and Woolcock, M. (2004). Health by association? Social capital, social theory, and the political economy of public health.
URL: *https://watermark.silverchair.com/dyh013.pdf?token=AQECAHi208BE49Ooan9kkhW_Ercy7Dm32DyVSAERJaTRAgEQgIIBcNMIDJ2AJmTXIRrRteDw-bTee6smqZHVePY7MnsXJa5F8esU*

Tamir, D. I. and Mitchell, J. P. (2012). Disclosing information about the self is intrinsically rewarding, *Proceedings of the National Academy of Sciences* **109**(21): 8038–8043.

Tan, S., Li, Y., Sun, H., Guan, Z., Yan, X., Bu, J., Chen, C. and He, X. (2014). Interpreting the public sentiment variations on twitter, *IEEE transactions on knowledge and data engineering* **26**(5): 1158–1170.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L. and Su, Z. (2008). Arnetminer: extraction and mining of academic social networks, *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 990–998.

Tanis, M. (2008). Health-Related On-Line Forums: What's the Big Attraction?, *Journal of Health Communication* **13**(7): 698–714.

Tao, C., Wei, W.-Q., Solbrig, H. R., Savova, G. and Chute, C. G. (2010a). A Semantic Web Ontology for Temporal Relation Inferencing in Clinical Narratives., *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* **2010**: 787–791.

Tao, C., Wei, W.-Q., Solbrig, H. R., Savova, G. and Chute, C. G. (2010b). A Semantic Web Ontology for Temporal Relation Inferencing in Clinical Narratives., *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* **2010**: 787–791.

Taylor, S. E. and Lobel, M. (1989). Social comparison activity under threat: Downward evaluation and upward contacts., *Psychological review* **96**(4): 569.

Thelwall, M., Buckley, K. and Paltoglou, G. (2011). Sentiment in Twitter Events, *Journal of the American Society for Information Science and Technology* **62**(2): 406–418.

Thelwall, M., Buckley, K. and Paltoglou, G. (2012). Sentiment Strength Detection for the Social Web, *Journal of the American Society for Information Science and Technology* **63**(1): 163–173.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D. and Kappas, A. (2010). Sentiment strength detection in short informal text, *Journal of the American Society for Information Science and Technology* **61**(12): 2544–2558.

Tidwell, L. C. and Walther, J. B. (2002). Computer-Mediated Communication effects on self-disclosure, impressions, and interpersonal evaluations, *Human Communication Research* **28**(3): 317–348.

Tomkins, S. S. (1984). Affect theory, *Approaches to emotion* **163**(163-195).

Trevino, L. K., Lengel, R. H. and Daft, R. L. (1987). Media symbolism, media richness, and media choice in organizations: A symbolic interactionist perspective, *Communication research* **14**(5): 553–574.

Tumasjan, A., Sprenger, T., Sandner, P. and Welpe, I. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment, *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media* pp. 178–185.

Valenzuela, S., Park, N. and Kee, K. F. (2009). Is there social capital in a social network site?: Facebook use and college students' life satisfaction, trust, and participation, *Journal of computer-mediated communication* **14**(4): 875–901.

Van Essen, D. C. (1985). Functional organization of primate visual cortex., *The cerebral cortex* .

Veatch, R. M. (1972). Models for ethical medicine in a revolutionary age: What physician-patient roles foster the most ethical relationship?, *Hastings Center Report* **2**(3): 5–7.

Vogel, E. A., Rose, J. P., Roberts, L. R. and Eckles, K. (2014). Social comparison, social media, and self-esteem., *Psychology of Popular Media Culture* **3**(4): 206.

Von der Malsburg, C. (1973). Self-organization of orientation sensitive cells in the striate cortex, *Kybernetik* **14**(2): 85–100.

Vydiswaran, V. G. V., Mei, Q., Hanauer, D. A. and Zheng, K. (2014). Mining Consumer Health Vocabulary from Community-Generated Text, *AMIA Annual Symposium Proceedings* pp. 1150–1159.

Wallsten, S. (2013). What are we not doing when we're online, *Technical report*, National Bureau of Economic Research.

Walther, J. B. (1992). Interpersonal Effects in Computer-Mediated Interaction: A Relational Perspective, *Communication Research* **19**(1): 52–90.

Walther, J. B. (2007). Selective self-presentation in computer-mediated communication: Hyperpersonal dimensions of technology, language, and cognition, *Computers in Human Behavior* **23**(5): 2538–2557.

Wang, A. H. (2010). Don't follow me: Spam detection in Twitter.

Wang, X., Hripcsak, G., Markatou, M. and Friedman, C. (2009). Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study, *Journal of the American Medical Informatics Association* **16**(3): 328–337.

Wang, X. and McCallum, A. (2006). Topics over time: a non-markov continuous-time model of topical trends, *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 424–433.

Wang, Y., Hunt, K., Nazareth, I., Freemantle, N. and Petersen, I. (2013). Do men consult less than women? an analysis of routinely collected uk general practice data, *BMJ open* **3**(8): e003320.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function, *Journal of the American statistical association* **58**(301): 236–244.

Watson, D. and Tellegen, A. (1985). Toward a consensual structure of mood., *Psychological bulletin* **98**(2): 219.

Weber, B. A., Roberts, B. L. and McDougall Jr, G. J. (2000). Exploring the efficacy of support groups for men with prostate cancer, *Geriatric Nursing* **21**(5): 250–253.

Weis, J. (2003). Support groups for cancer patients.
URL: *https://link.springer.com/content/pdf/10.1007%2Fs00520-003-0536-7.pdf*

Wellman, B., Haase, A. Q., Witte, J. and Hampton, K. (2001). Does the internet increase, decrease, or supplement social capital? social networks, participation, and community commitment, *American behavioral scientist* **45**(3): 436–455.

Wen, M. and Rose, C. P. (2012). Understanding participant behavior trajectories in online health support groups using automatic extraction methods, *Proceedings of the 17th ACM international conference on Supporting group work - GROUP '12*, p. 179.

Weng, J., Yao, Y., Leonardi, E., Lee, F. and Lee, B.-s. (2011). Event Detection in Twitter, *ICWSM* **11**(98): 401–408.

Wiessner, P. W. (2014). Embers of society: Firelight talk among the ju/hoansi bushmen, *Proceedings of the National Academy of Sciences* **111**(39): 14027–14035.

Williams, G. (1984). The genesis of chronic illness: narrative reconstruction, *Sociology of Health & Illness* **6**(2): 175–200.

Wills, T. A. (1981). Downward comparison principles in social psychology., *Psychological bulletin* **90**(2): 245.

Willshaw, D. J. and Von Der Malsburg, C. (1976). How patterned neural connections can be set up by self-organization, *Proceedings of the Royal Society of London. Series B. Biological Sciences* **194**(1117): 431–445.

Winzelberg, A. J., Classen, C., Alpers, G. W., Roberts, H., Koopman, C., Adams, R. E., Ernst, H., Dev, P. and Taylor, C. B. (2003). Evaluation of an internet support group for women with primary breast cancer, *Cancer* **97**(5): 1164–1173.

Woolcock, M. and Narayan, D. (2000). Social capital: Implications for development theory, research, and policy, *The world bank research observer* **15**(2): 225–249.

Woolcock, M. et al. (2001). The place of social capital in understanding social and economic outcomes, *Canadian journal of policy research* **2**(1): 11–17.

Wright, K. B. (2006). Researching Internet-Based Populations: Advantages and Disadvantages of Online Survey Research, Online Questionnaire Authoring Software Packages, and Web Survey Services, *Journal of Computer-Mediated Communication* **10**(3): 00–00.

Xu, J., Dailey, R. K., Eggly, S., Neale, A. V. and Schwartz, K. L. (2011). Mens perspectives on selecting their prostate cancer treatment, *Journal of the National Medical Association* **103**(6): 468.

Xu, Z., Ru, L., Xiang, L. and Yang, Q. (2011). Discovering User Interest on Twitter with a Modified Author-Topic Model, *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, Vol. 1, IEEE, pp. 422–429.

Yan, X., Guo, J., Lan, Y. and Cheng, X. (2013). A biterm topic model for short texts, *Proceedings of the 22nd international conference on World Wide Web*, ACM, pp. 1445–1456.

Yan Zhang (2016). Understanding the Sustained Use of Online Health Communities From a Self-Determination Perspective, *Journal of the Association for Information Science and Technology* **67**(12): 2842–2857.

Yang, Y., Pierce, T. and Carbonell, J. (1998). A study of retrospective and on-line event detection, *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 28–36.

Zamir, O., Etzioni, O., Madani, O. and Karp, R. M. (1997). Fast and intuitive clustering of web documents., *KDD*, Vol. 97, pp. 287–290.

Zappavigna, M. (2015). Searchable talk: the linguistic functions of hashtags, *Social Semiotics* **25**(3): 274–291.

Zeliadt, S. B., Ramsey, S. D., Penson, D. F., Hall, I. J., Ekwueme, D. U., Stroud, L. and Lee, J. W. (2006). Why do men choose one treatment over another? A review of patient

decision making for localized prostate cancer.
URL: *http://doi.wiley.com/10.1002/cncr.21822*

Zhang, S., Bantum, E. O., Owen, J., Bakken, S. and Elhadad, N. (2017). Online cancer communities as informatics intervention for social support: Conceptualization, characterization, and impact, *Journal of the American Medical Informatics Association* **24**(2): 451–459.

Zhang, S., Kang, T., Qiu, L., Zhang, W., Yu, Y. and Elhadad, N. (2017). Cataloguing Treatments Discussed and Used in Online Autism Communities, *Proceedings of the 26th International Conference on World Wide Web - WWW '17*, Vol. 2017, ACM Press, New York, New York, USA, pp. 123–131.

Zhang, T., Cho, J. H. D. and Zhai, C. (2015). Understanding User Intents in Online Health Forums, *IEEE Journal of Biomedical and Health Informatics* **19**(4): 1392–1398.

Zhang, Z., Iria, J., Brewster, C. and Ciravegna, F. (2008). A comparative evaluation of term recognition algorithms, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, pp. 2108–2113.

Zhao, D. and Rosson, M. B. (2009). How and Why People Twitter: The Role that Micro-blogging Plays in Informal Communication at Work, *Proceedings of the ACM 2009 international conference on Supporting group work*, ACM Press, New York, New York, USA, pp. 243—-252.

Zhou, X. and Chen, L. (2014). Event detection over twitter social media streams, *VLDB Journal* **23**(3): 381–400.

Zhu, H., Ni, Y., Peng, C., Qiu, Z. and Cao, F. (2012). Automatic extracting of patient-related attributes: Disease, age, gender and race, *Studies in Health Technology and Informatics*, Vol. 180, pp. 589–593.

Ziebland, S. and Wyke, S. (2012). Health and Illness in a Connected World: How might sharing experienes on the Internet affect people's health, *The Millbank Quarterly* **90**(2): 219–49.

Zigmond, A. S. and Snaith, R. P. (1983). The hospital anxiety and depression scale, *Acta psychiatrica scandinavica* **67**(6): 361–370.