Western Graduate&PostdoctoralStudies

Western University
Scholarship@Western

Electronic Thesis and Dissertation Repository

11-7-2016 12:00 AM

On the Promotion of the Social Web Intelligence

Taraneh Khazaei, *The University of Western Ontario*

Supervisor: Dr. Lu Xiao, *The University of Western Ontario*
Joint Supervisor: Dr. Robert Mercer, *The University of Western Ontario*
A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Computer Science
© Taraneh Khazaei 2016

Follow this and additional works at: https://ir.lib.uwo.ca/etd

 Part of the Artificial Intelligence and Robotics Commons, Mass Communication Commons, and the Social Media Commons

Recommended Citation

Khazaei, Taraneh, "On the Promotion of the Social Web Intelligence" (2016). *Electronic Thesis and Dissertation Repository*. 4308.
https://ir.lib.uwo.ca/etd/4308

This Dissertation/Thesis is brought to you for free and open access by Scholarship@Western. It has been accepted for inclusion in Electronic Thesis and Dissertation Repository by an authorized administrator of Scholarship@Western. For more information, please contact wlswadmin@uwo.ca.

Abstract

Given the ever-growing information generated through various online social outlets, analytical research on social media has intensified in the past few years from all walks of life. In particular, works on *social Web intelligence* foster and benefit from the wisdom of the crowds and attempt to derive actionable information from such data. In the form of *collective intelligence*, crowds gather together and contribute to solving problems that may be difficult or impossible to solve by individuals and single computers. In addition, the consumer insight revealed from social footprints can be leveraged to build powerful *business intelligence* tools, enabling efficient and effective decision-making processes. This dissertation is broadly concerned with the intelligence that can emerge from the social Web platforms. In particular, the two phenomena of *social privacy* and *online persuasion* are identified as the two pillars of the social Web intelligence, studying which is essential in the promotion and advancement of both collective and business intelligence.

The first part of the dissertation is focused on the phenomenon of social privacy. This work is mainly motivated by the privacy dichotomy problem. Users often face difficulties specifying privacy policies that are consistent with their actual privacy concerns and attitudes. As such, before making use of social data, it is imperative to employ multiple safeguards beyond the current privacy settings of users. As a possible solution, we utilize user social footprints to detect their privacy preferences automatically. An unsupervised collaborative filtering approach is proposed to characterize the attributes of publicly available accounts that are intended to be private. Unlike the majority of earlier studies, a variety of social data types is taken into account, including the social context, the published content, as well as the profile attributes of users. Our approach can provide support in making an informed decision whether to exploit one's publicly available data to draw intelligence.

With the aim of gaining insight into the strategies behind online reasoning, the second part of the dissertation studies written comments in online deliberations. Specifically, we explore different dimensions of the language, the temporal aspects of the communication, as well as the attributes of the participating users to understand what makes people change their beliefs. In addition, we investigate the factors that are perceived to be the reasons behind persuasion by the users. We link our findings to traditional persuasion research, hoping to uncover when and how they apply to online persuasion. A set of rhetorical relations is known to be of importance in persuasive discourse. We further study the automatic identification and disambiguation of such rhetorical relations, aiming to take a step closer towards automatic analysis of reasoning traces in online platforms. Finally, a small proof of concept tool is presented, showing the value of our persuasion and rhetoric studies.

Keywords: Social Web, Intelligence, Social Privacy, Persuasive Discourse

Acknowledgement

First and foremost, I start by thanking my supervisors Dr. Xiao and Dr. Mercer. I want to show the depth of my gratitude for their continuous and insightful support throughout all these years, which made my doctoral experience a very joyful and a productive one. This research would not have been possible without their support and patient encouragement.

I would also like to acknowledge the financial, academic, and technical assistance from the Department of Computer Science and Faculty of Information and Media Studies at the University of Western Ontario. Additionally, I would like to thank Mitacs Accelerate for supporting our successful collaboration with an industry partner for a year. I would also like to thank our partner, InfoTrellis Inc., for their support during the internship. I would particularly like to acknowledge my supervisor at InfoTrellis, Mr. Atif Khan, for the excellent cooperation and for all of the opportunities I was given to conduct my research and further my dissertation.

Finally, I would like to express appreciation to my best friend and my beloved husband who gave me unconditional support and encouragement throughout this process. Last but not the least, I would like to thank my parents and my brother for their continuous support throughout the duration of my Ph.D. studies.

Co-Authorship Statement

This dissertation is written in an integrated article format. Chapter 1 is the original work of the dissertation author in introducing the dissertation and providing a brief background. Chapter 2 to Chapter 5 are all related to the project conducted in collaboration with our industry partner, InfoTrellis. Consequently, they are all collaborative efforts with the academic supervisors, Dr. Xiao and Dr. Mercer, and the industry supervisor, Mr. Atif Khan. The author of this dissertation is the primary author for all of these articles and carried out the literature review, collected data, developed the approach, conceived the design, performed the statistical analysis, and drafted the manuscripts to be published. The supervisors of the project have participated in the design of the study, interpretation of the results, and critical revisions of the manuscripts. It should be noted that one of the datasets utilized in the project was provided by the company. In addition, the company provided support in publishing online experiment through Amazon Mechanical Turk and collecting data from it.

Chapters 6 and 8 are co-authored with the principal supervisor, Dr. Xiao, while Chapters 7 and 9 are collaborative efforts with both supervisors, Dr. Xiao and Dr. Mercer. The author of the dissertation is the primary author of all these articles as well and surveyed the literature, designed and developed the approach, performed the analysis, and drafted the manuscripts for publication. In the capacity of supervisor, Dr. Xiao and Dr. Mercer have participated in the design of the study, interpretation of the results, and preparation of the articles. Chapter 10 is a co-authored article with Dr. Xiao, where the author of the dissertation is the second author. For this article, Dr. Xiao provided the main conceptual idea and drafted parts of the paper. The dissertation author designed and developed the system and crafted parts of the paper accordingly. Finally, Chapter 11 is the original work of the author in drawing conclusions and summarizing the dissertation.

Contents

Abstract i

Acknowledgement ii

Co-Authorship Statement iii

List of Figures ix

List of Tables xi

1 Introduction 1

1.1 Motivation 1
1.2 Objectives 4
1.3 Contributions 5
1.4 Thesis Organization 6

2 Social Privacy: A Review 10

2.1 Introduction 10
2.2 Privacy in Social Networks 12
2.3 Literature Review 13
2.3.1 Personal and Profile Attributes 14
2.3.2 Social Context 15
Social Circle Management and Labeling 16
Collaborative Filtering for Privacy Inference 17
2.3.3 Published Content 18
2.4 Discussion 20
2.5 Summary and Conclusion 21

3 Privacy and Profile Attributes in Twitter 29

3.1 Introduction 29
3.2 Related Work 30

3.3 Data Source . . . 32
3.3.1 User Selection . . . 32
3.3.2 Profile Features . . . 33
3.4 Analysis of Profile Attributes . . . 34
3.4.1 Surface-based Profile Features . . . 35
3.4.2 Profile Descriptions: A Closer Look . . . 35
3.5 User Classification . . . 38
3.6 Discussion . . . 39
3.7 Conclusions . . . 40

4 Addressing Privacy Dichotomy in Twitter **45**
4.1 Introduction . . . 45
4.2 Related Work . . . 48
4.3 Dataset . . . 52
4.3.1 Data Collection . . . 52
4.3.2 Descriptive Analysis . . . 53
4.4 Amazon Mechanical Turk Experiment . . . 54
4.5 User Analysis . . . 56
4.5.1 Observable Attributes . . . 56
Profile Attributes . . . 56
Language Use of the Content . . . 59
Tweet Sentiment . . . 62
Communication Behaviour . . . 63
4.5.2 Latent Attributes . . . 63
Method . . . 63
Results . . . 64
4.6 Discussion . . . 67
4.7 Conclusion . . . 69

5 Privacy Preference Inference via Collaborative Filtering **76**
5.1 Introduction . . . 76
5.2 Related Work . . . 77
5.3 Neighbourhood-based Latent Factor Model . . . 78
5.4 Profile Attributes for Privacy Prediction . . . 79
5.5 Privacy Graph and Latent Attributes . . . 80
5.5.1 Graph Construction and Properties . . . 80
5.5.2 Latent Attribute Detection . . . 82

5.6 Conclusion . . . 83

6 Computational Analysis of Collective Intelligence: A Review 86
6.1 Introduction . . . 86
6.2 Methodology . . . 89
6.2.1 Framework for the Systematic Review . . . 89
6.2.2 Article Selection Process . . . 90
6.3 Review of the Literature . . . 91
6.3.1 Communication Pattern Analysis . . . 91
6.3.2 Content Analysis . . . 93
Sentiment Analysis . . . 93
Task/Purpose Performance Analysis . . . 95
6.3.3 Impact Analysis . . . 98
6.4 Discussion . . . 99
6.5 Conclusion . . . 101

7 Determinants of Online Persuasion 109
7.1 Introduction . . . 109
7.2 Related Work . . . 111
7.3 Reddit Dataset . . . 113
7.4 Persuasive Comments: How are they Different? . . . 116
7.4.1 Relevance to the Original Post . . . 116
7.4.2 Timing and Order . . . 118
7.4.3 Psychological Attributes of the Language . . . 119
7.4.4 Writing Sophistication and Comprehensibility . . . 124
7.4.5 User Delta and Karma Score . . . 128
7.4.6 Prediction Evaluation . . . 128
7.5 Persuaded Users: Why are they Persuaded? . . . 131
7.6 Discussion . . . 133
7.7 Conclusion . . . 134

8 RST Cue Extraction and Analysis 141
8.1 Introduction . . . 141
8.2 Related Work . . . 142
8.3 Methodology . . . 143
8.3.1 Underlying Corpora . . . 143
8.3.2 Lexical Cue Extraction . . . 144

8.4 Experiment Results . . . 145
8.5 Discussion . . . 149
8.6 Conclusion . . . 155

9 RST Cue Disambiguation 160
9.1 Introduction . . . 160
9.2 Related Work . . . 162
9.3 Approach . . . 164
9.3.1 Corpora . . . 164
9.3.2 Lexical Cue Selection . . . 165
9.3.3 Lexical Cue Disambiguation . . . 166
Data Collection . . . 166
Syntactic Representations . . . 167
Graph Modeling . . . 167
Cue Disambiguation Model . . . 168
9.4 Evaluation . . . 168
9.5 Discussion . . . 170
9.6 Conclusion . . . 173

10 ProjectTales: A Proof of Concept 177
10.1 Introduction . . . 177
10.2 Related Work . . . 178
10.3 ProjectTales: An interactive visualization tool . . . 179
10.3.1 Database and Its Design Rationale . . . 179
10.3.2 Interface Design . . . 180
History Overview Component . . . 180
Project Detailed View Component . . . 181
10.4 Evaluation . . . 182
10.5 Conclusion . . . 182

11 Conclusions 185
11.1 Summary of Research . . . 185
11.2 Contributions . . . 188
11.3 Future Directions . . . 189

A Amazon Mechanical Turk Documentation 192

B Amazon Mechanical Turk Experiment and Results 197

Curriculum Vitae **203**

List of Figures

1.1 An overview of the social Web intelligence cycle. 3

2.1 An example of a social network with focal user U. 13
2.2 An overview of the automated privacy prediction approaches. 21

4.1 An overview of the Twitter data retrieval procedure. 52
4.2 An overview of the degree distribution across all users in the network on a log-log scale. 53
4.3 Correlation of the number of *public* contacts and the number of *protected* contacts. 54
4.4 Correlation of users' privacy ratio and the average privacy ratio for all of their contacts. 54
4.5 An overview of the network transformation process. 65

5.1 A snapshot of the privacy graph. 81
5.2 The distribution of protected contacts. 81

6.1 Categorization of collective intelligence based on the types of interaction. . 88
6.2 A simplified version of McGrath's framework 90

7.1 Example of a CMV post. 115
7.2 Density of relevance grade distribution for persuasive comments. 118
7.3 Density of temporal grade distribution for persuasive comments. 119
7.4 Analysis of comments across different temporal quarters. 120
7.5 The result of the feature analysis conducted via Boruta. 130
7.6 An example explanation parsed by a dependency parser. 131
7.7 The grammatical analysis of the user explanations. 132

8.1 An overview of the cue extraction, filtering, and classification processes. . . 146
8.2 The tf-idf metric for the top cues. 148
8.3 The precision metric for the top cues. 150
8.4 The recall metric for the top cues. 151

8.5 The F metric for the top cues. 152

9.1 An example sentence parsed in the form of RST 161
9.2 A high-level overview of the cue extraction and disambiguation approach. . 169

10.1 A screenshot of ProjectTales 180

B.1 The distrbution of responses for the desired privacy 198
B.2 The distrbution of responses for the desired privacy when benefits are given 199
B.3 The distrbution of responses for the first AMT annotation task 200
B.4 The distrbution of responses for the second AMT annotation task 200
B.5 The distribution of the number of private categories across different user timelines 201

List of Tables

2.1 A privacy specification example in a social network. 13

3.1 A set of popular accounts in Twitter and the statistics of their collected follower sets. 33
3.2 Analysis of binary profile attributes of protected and public accounts. 36
3.3 Analysis of numeric profile attributes of protected and public accounts. . . . 36
3.4 LIWC categories and their corresponding percentage for protected and public descriptions. 37
3.5 LIWC summary variables and their corresponding values for protected and public descriptions. 38
3.6 Evaluation of classification results for protected and public accounts. 39

4.1 Analysis of binary profile attributes and privacy-related features. 58
4.2 Analysis of numeric profile attributes and privacy-related features. 59
4.3 Correlation analysis of the LIWC categories and the privacy ratios. 61
4.4 Analysis of a set of lingustic indicators and the privacy-related features. . . 62
4.5 Descriptive statistics of the features that represent user timelines 66
4.6 Example features extracted from user timelines along with their corresponding privacy ratios 67
4.7 Topics extracted from user timelines along with their corresponding privacy ratios. 67

5.1 Profile features to detect protected accounts. 80
5.2 Latent factors extracted from the privacy graph. 83

6.1 A summary of earlier research according to the McGraths framework. . . . 100

7.1 persuasion corpus statistics 116
7.2 Welch T-Test for the persuasive and non-persuasive comment groups in terms of the presence of psychological indicators. 122

7.3 Welch T-Test for the persuasive and non-persuasive comment groups in terms of writing quality. . . . 126
7.4 The evaluation results of supervised machine learning algorithms for the detection of persuasive comments. . . . 129

8.1 The percentage of the rhetorical relations of focus in the underlying corpora. 144
8.2 Sample lexical cues extracted from the two corpora. . . . 145
8.3 Cue labels and their corresponding lexical cues for the circumstance relation 147
8.4 The precision, recall, and F score calculated for cue sets of different sizes. . 154
8.5 Distribution of atomic and embedded rhetorical relations. . . . 154

9.1 Different granularity levels of POS tags. . . . 169
9.2 Evaluation results for the classification of the circumstance relation across different cues for the SFU corpus. . . . 170
9.3 Evaluation results for the classification of the circumstance relation across different cues for the RST corpus. . . . 171
9.4 Evaluation results for the classification of the circumstance relation on the SFU corpus with the direct usage of features. . . . 171
9.5 Evaluation results for the classification of the circumstance relation on the RST corpus with the direct usage of features. . . . 172

Chapter 1

Introduction

1.1 Motivation

The advent and ongoing advancements of computer-mediated communication have revolutionized our social lives, affecting how we perceive time, distance, and global boundaries. The emergence of Web 2.0 technologies, in particular, have enabled an increasing number of individuals to engage in online social network activities, leading to unprecedented amounts of daily content generation and communication taking place on the Web. Wikipedia, Q&A forums, social networking and multimedia content sharing websites are all examples of such social platforms. This dissertation is mainly concerned with the emergence of intelligence from such large social footprints and studies two social phenomena that can profoundly affect the intelligence power of the social Web.

In spite of the long history of research from a variety of disciplines, no standard and well-established definition of intelligence exists [6]. In fact, as quoted in [4]: "Viewed narrowly, there seem to be almost as many definitions of intelligence as there were experts asked to define it". For this research, we adapted the well-known definition provided by the psychologist Howard Gardner. He defines intelligence as "the ability to solve problems, or to create products, that are valued within one or more cultural settings" [3]. The Social Web can leverage two types of intelligence that are different primarily in terms of the cultural setting for which they are of value, namely collective and business intelligence.

In its traditional form, collective intelligence refers to the intelligence that emerges from local interactions among individual people [5]. In the last few decades, social Web technologies have enabled new forms of collective intelligence that allow massive numbers of loosely organized individuals to interact with one another, solve problems, and create high-quality intellectual artifacts. By enabling collaboration and deliberation on a massive scale, such social platforms move beyond problem-solving capabilities of a small group

of authorities and can effectively harness the collective power of a large set of individuals with diverse backgrounds and expertise. Currently, many social platforms are contributing towards the emergence and development of collective intelligence. In Wikipedia, for instance, users' collective efforts to generate and edit content has resulted in the creation of such a massive encyclopedia. Users contributions are built upon each other to create high-quality open-source projects such as Linux. Prediction markets, which let people bet on the outcomes of events like presidential elections, often result in surprisingly accurate results.

In addition to the phenomenon of collective intelligence, businesses are tapping into social media as an essential component of their next-generation business intelligence, which can assist them in product design, consumer relation management, and targeted marketing and advertising. Social media data can be of great value for businesses industry-wide, ranging from food truck owners updating their truck locations to non-profit organizations sharing stories that can touch the hearts of millions. In addition, billion-dollar brands are actively investing in social media analytics platforms to constantly communicate with existing loyal customers and to recommend and advertise relevant products or services to potential customers. For instance, Apple Inc. acquired the Twitter analytics platform Topsy Labs Inc. for more than $200 million dollars in 2012 [12]. A giant social media analytics platform, called Networked Insight, is helping Samsung, Revlon, and Disney make effective marketing decisions [9]. The Ford Motor Company has been utilizing social media to build anticipation and emotional attachment to their vehicles even months before they are released [11].

As mentioned earlier, these two phenomena are mainly different in terms of the cultural setting that they are of value for. The main purpose of business intelligence tools and methods are to increase corporate revenue and profit, while collective intelligence that arises from interactions between individuals is primarily intended to serve the community itself. However, these two kinds of intelligence are not isolated and can affect one another. Marketers can first use social media data to target potential customers and advertise relevant products and services to them. The selected crowd may then get involved in a collective intelligence process and might collectively propose refinements and improvements to the products, forming an iterative intelligence process. The following event provides an example of this process.

Soon after the launch of iPhone 6 Plus, social media users started to complain about the phone being prone to bending by posting videos, tweets, blogs, etc. In particular, *#BendGate*, which was used to report this issue, went viral across social networking websites. These social media activities created a storm of discussions among users: some users

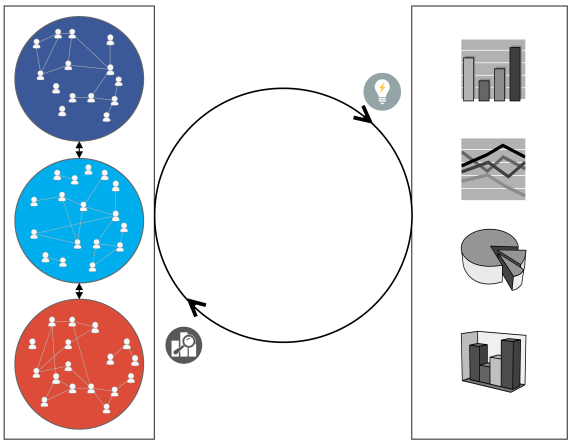


Figure 1.1: Business Intelligence techniques can be used to target communities and even individuals across multiple social media platforms (e.g., Facebook, Twitter, and Google+). The advertised products and services may then become the topic of discussion among social media users, wherein possible flaws may be identified, solutions may be proposed and compared, and enhancements and improvements may be suggested. The intelligence that emerges from such potentially massive interactions can then be exploited by the firms for their future business development and marketing plans.

blamed the aluminum exterior, while some attributed the problem to the extremely thin size of the phone; other users, argued that other phones would deform under the same pressure and even posted videos and photos of their experiments. Many companies are now taking this type of social feedback into account to finesse their next generation products and services accordingly [2]. Continuing with the iPhone 6 story, the new iPhone is known to use a new alloy for the phone body to prevent bending even under extreme force. In comparison, this iterative process may emerge from a collective intelligence. For instance, social media users may collectively express their need for a new service or product and may even propose solutions to resolve these needs. This iterative process forms a cycle, wherein collective and business intelligence phenomena influence one another. Figure 1.1 shows a simplified overview of this process.

Facilitating this cycle requires intensive computational resources and sophisticated techniques to collect, analyze, and mine data and to learn and model user preferences. However, the entire cycle of the social Web intelligence hinges heavily on two social phenomena: *social privacy* and *justification and reasoning*. Any systematic use of user personal and social data requires privacy measures that if disregarded will result in privacy violations. Therefore, for this cycle to even begin, privacy concerns need to be adequately addressed. In addition, reasoning strategies play a crucial role, affecting every single relation in the cycle. Businesses need to be able to influence and persuade their targeted customers. In

addition, users may need to influence each other and possibly be able to change beliefs in online deliberations and discussions for the collective intelligence to arise. In this dissertation, we explore these two important aspects of intelligence in the social Web. In particular, we discuss the current research gaps in studying these two phenomena, our research efforts to address these gaps, as well as suggested directions in social media mining and analytics to support collective and business intelligence.

1.2 Objectives

The primary objective of the dissertation is to study and explore two core social elements of privacy and reasoning, both influencing the whole paradigm of collective and business intelligence. With this high-level idea in mind, we developed the following objectives:

Identification of research gaps through multiple comprehensive and systematic literature reviews: A systematic review of prior relevant research is an integral part of any academic project [13]. The advent and an ever increasing role of online social platforms in our lives have ushered in a new research direction and community that have been quite active in the past few decades. However, similar to any other new and evolving topic, we expected the related work to be more focused on a specific set of issues, while leaving certain areas underexplored and even unnoticed. Hence, we aim to gain insight into the current realm of research on social intelligence by conducting multiple exhaustive and systematic literature reviews.

Understanding the behaviours and characteristics of users with different privacy behaviours and attributes: Social networking websites are known to suffer from the privacy dichotomy problem [1, 7]. Privacy dichotomy occurs when users' privacy behaviours are not consistent with their actual privacy attitudes and concerns. Such a disparity can be mainly attributed to the complexities associated with making privacy decisions and users' lack of awareness about the default privacy setting. Therefore, the current privacy settings of users cannot accurately reflect their privacy attitudes and preferences. Motivated by this widely observed issue, we plan to make use of online social activities to gain insight into the behaviours and attributes of users with different levels of privacy concern. As a result of this analysis, we hope to take a step closer to building a model for the identification of the users who are concerned about their privacy, yet are following open and permissive privacy configurations.

Understanding the mechanisms behind online persuasion: Studying online persuasion and influence can greatly contribute towards our understanding of online reasoning traces and strategies. In addition, achieving social intelligence through deliberation and discus-

sion is heavily reliant on the ability to persuade and influence others, yet this capacity is known to be one of the most challenging social skills to develop and possess. With social networking websites becoming a crucial platform for routine social activities, gaining insight into the strategies and mechanisms behind online persuasion can be of great interest and value to a variety of disciplines. Hence, we aim to study written comments in online deliberations and understand what dimensions of language makes people change their beliefs.

Facilitating identification and disambiguation of linguistic relations that are of importance in persuasion: Rhetorical relations, also known as discourse relations, are paratactic or hypotactic relations that hold between spans of text, explaining the construction of coherence in discourse [10]. According to earlier research, a subset of such rhetorical relations is known to appear commonly in rationales [14]. Rationales are the pieces of text that users provide to back up their claims. Thus, their analysis can greatly contribute towards the study of online argumentation and reasoning. As such, we attempt to study this subset of rhetorical relations and to build a model for their automatic identification in a given discourse text.

Development of a proof of concept: Studying online reasoning strategies from different perspectives can be vastly beneficial to a variety of disciplines including philosophy, sociology, and artificial intelligence. In addition to these benefits, the resulting knowledge and models can be incorporated into the design of novel interfaces and visualization tools to further promote and foster intelligence, thus contributing to the field of human-computer interaction as well. Hence, we build such a proof of concept system that takes advantage of the models to detect rationales and persuasion to reinforce the intelligence power of the social Web.

1.3 Contributions

In general, works on user modeling and analysis in the context of privacy can improve our knowledge of users' privacy behaviours and attitudes, contributing to the field of sociology. The work on online persuasion and influence can also contribute towards philosophy and sociology research as the resulting models can be utilized to understand peoples' reasoning and persuasion strategies on the social Web. For the areas of human-computer interaction and information visualization, the findings can provide valuable insights for the design and development of novel systems. Our main contributions in the context of artificial intelligence are highlighted below:

Our exhaustive reviews of the literature generate new knowledge about the topic of so-

cial Web intelligence and computing. In particular, our literature surveys on social privacy and discourse-based collective intelligence contributes towards creating a firm foundation for the current knowledge of the field and closing the areas where a plethora of research already exists [13]. Also, these reviews contribute towards uncovering areas where further research is required [13].

With the ultimate goal of predicting one's privacy preferences in social media, we ran a series of experiments that shed light on the relationships of privacy attitudes and other social attributes of the user. In particular, we found differences in how the profile attributes of users with varying privacy settings are configured. We identified a set of clues, showing that users privacy preferences are similar to the privacy behaviour of their social contacts, signaling that privacy preferences may be localized in social networks. Finally, we found differences in the textual content shared by users with different privacy features.

Online reasoning traces can be studied by exploring various related social processes including online persuasion and influence, opinion mining, and rationale detection. Despite the long history of persuasion theories in traditional settings, online persuasion, and belief change has received relatively little attention. One of the contributions of this dissertation is the detection of a set of attributes that are influential in the persuasion process and the analysis of their predictive power. The majority of the attributes are associated with the persuasive impact of various components of the language, including readability, cohesion, and the presence of psychological indicators. As another effort to understand reasoning strategies, this dissertation contributes toward the identification and disambiguation of rhetorical relations that are commonly observed in rationales. The capacity of our approach to work across different text genres is also examined.

Natural language is known to reveal important aspects of people's social and psychological worlds [8]. In addition, the power of language in impacting people's emotions, beliefs, and social and psychological states is inescapable. This dissertation, once again, validates the essential role of both linguistic style and linguistic content in understanding user behaviour in online social networks, regardless of the length and the informality of the language.

1.4 Thesis Organization

In this research, a structured approach has been employed to study and analyze the key elements related to the phenomenon of social Web intelligence. The first step in this approach has been to examine the two forms of intelligence on the social Web, to develop a social Web intelligence paradigm, and finally to identify the central social factors affecting the

overall process (i.e., social privacy and online reasoning). This part of the study has been highlighted earlier in this Chapter (**Chapter 1**).

As an initial step toward addressing the privacy dichotomy issue, we review and examine the literature that makes use of users' online social footprints to discover desired privacy settings. Throughout the analysis, a set of gaps is identified, requiring further research attention. This literature review is presented in **Chapter 2**. Next, we primarily focus on Twitter and study whether profile attributes of Twitter users with varying privacy settings are configured differently. Our efforts to address this research question are presented in **Chapter 3**. In addition to the profile attributes, we study the value of users' social context and published content in characterizing their privacy attitudes. A set of attributes are found that are expected to characterize publicly available accounts that are intended to be private. These findings are discussed in **Chapter 4**. Finally, **Chapter 5** proposes a high-level hybrid collaborative approach to detect privacy preferences.

The second part of the thesis starts with a broad literature review on the concept of discourse-centric collective intelligence. This literature review, presented in **Chapter 6**, shows the current research status of the field, identifies a set of gaps, and suggests new directions. In addition, techniques and methods reviewed in this Chapter informed some of our decisions in studying and understanding online persuasion. **Chapter 7** studies different dimensions of the language, the temporal aspects of the communication, as well as the attributes of the participating users and their relations to the persuasion process. In addition to the linguistic elements analyzed in **Chapter 7**, the presence of a set of rhetorical relations can signal the presence of the users' rationales in their arguments, thus can be valuable in online reasoning studies. The only datasets annotated based on rhetorical relations belong to genres other than online social communications. Therefore, **Chapter 8** studies how this subset of rhetorical relations are signaled by lexical cues across different text genres. Also, **Chapter 9** provides an extended approach to disambiguate such lexical cues. **Chapter 10** demonstrates a proof-of-concept design, showing the potential value of rationale identification models.

Finally, as part of our privacy study, we utilized a crowdsourcing platform to collect human-annotated data. The documentation of this experiment is provided in **Appendix A**, while the description of the study design and the preliminary analysis of the results are provided in **Appendix B**.

Bibliography

[1] Alessandro Acquisti and Ralph Gross. Imagined communities: Awareness, information sharing, and privacy on the Facebook. In *Proceedings of the Conference on Privacy Enhancing Technologies*, pages 36–58, 2006.

[2] Josh Bernoff and Charlene Li. Harnessing the power of the oh-so-social web. *MIT Sloan Management Review*, 49(3):36, 2008.

[3] Howard Gardner. *Frames of mind: The theory of multiple intelligences*. Basic Books, 2011.

[4] R.L. Gregory and O.L. Zangwill. *The Oxford Companion to the Mind*. Oxford University Press, 1987.

[5] J.F. Kennedy, J. Kennedy, R.C. Eberhart, and Y. Shi. *Swarm Intelligence*. Evolutionary Computation Series. Morgan Kaufmann Publishers, 2001.

[6] Shane Legg and Marcus Hutter. A collection of definitions of intelligence. In *Proceedings of the Conference on Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms*, pages 17–24, 2007.

[7] Heather Richter Lipford, Andrew Besmer, and Jason Watson. Understanding privacy settings in Facebook with an audience view. In *Proceedings of the Conference on Usability, Psychology, and Security*, pages 2:1–2:8, 2008.

[8] James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577, 2003.

[9] Bruce Rogers. Dan neely's networked insights wants to be the oracle of marketing. `http://www.forbes.com/sites/brucerogers/2013/09/16/dan-neelys-networked-insights-wants-to-be-the-oracle-of-marketing/#30bde04053bd`. Accessed: 2016-08-03.

[10] Maite Taboada. Discourse markers as signals (or not) of rhetorical relations. *Journal of Pragmatics*, 38(4):567 – 592, 2006.

[11] David Vinjamuri. Ford remixes the fiesta movement for 2014. http://www.forbes.com/sites/davidvinjamuri/2013/02/19/ford-revives-the-fiesta-movement-to-launch-the-2014-fiesta/#10451eed72b5. Accessed: 2016-08-03.

[12] Daisuke Wakabayashi and Douglas Macmillan. Apple taps into Twitter, buying social analytics firm Topsy. http://www.wsj.com/articles/SB10001424052702304854804579234450633315742. Accessed: 2016-08-03.

[13] Jane Webster and Richard T. Watson. Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly*, 26(2):xiii–xxiii, 2002.

[14] Lu Xiao. Do members converge to similar reasoning styles in teamwork? A study of shared rationales in small team activities. In *Proceedings of the iConference*, pages 524–530, 2013.

Chapter 2

Social Privacy: A Review

2.1 Introduction

With the growing interest in regular communication and information sharing over online social media, privacy has emerged as a serious concern. Prior studies on users' online behaviour indicate that there is a disparity between the privacy-related attitudes of social media users and their actual behaviour in specifying their privacy policies [1, 19, 28, 49]. Even though social media users may be highly concerned about their privacy, they face difficulties managing their privacy policies, so only a small percentage of users change their default privacy settings. This issue may occur due to their misconceptions regarding the visibility of their data [49] and the complex and unusable interfaces [13]. In addition, psychology research has shown that defaults are often perceived as the recommended course of action [31, 21]. Additionally, users are rarely reminded to reconsider their privacy policies after their initial profile creation; hence, they often overlook the visibility of their social networking data [49].

Meanwhile, as more people engage on social media, they are providing businesses with unprecedented amounts of data that give insight into various facets of customer behaviour. Current social networking platforms allow users to publish their activities, opinions, locations, as well as their social interactions through different forms of communication (e.g., text, image, and video), leading to large *social footprints* [20]. The insight into customer behaviour provided by such social footprints affords immense opportunities for businesses to engage audiences with compelling and personalized content and experiences. By har-

A version of this chapter has been published in the *proceedings of the iConference.*

nessing this additional information about individuals, traditional customer databases can be transformed from historical artifacts into powerful business intelligence tools, enabling efficient and effective business decision-making processes. For example, social media data can be used to detect users' upcoming life events and provide them with relevant offers, or to gain insight into users' psychographics and send marketing messages uniquely tailored to them.

In addition to the tremendous potential that social media data can provide for businesses, prior customer studies have shown that customers and clients value personalized content [24]. In addition, as customers increase their digital footprints, they expect more personalization [42]. However, the effectiveness of this win-win opportunity relies on addressing users' privacy concerns and reconciling the tension between personalization and privacy. The Facebook "Beacon" feature is an alarming example of disregarding users' privacy preferences. Launched in November 2007, "Beacon" allowed third-party websites, such as Coca-Cola, Sony Pictures, and Verizon, to access Facebook profiles and to provide personalized content and services to them and their friends. This feature immediately encountered mass protests and was retracted from Facebook several weeks later.

Users' privacy is violated when information intended for a particular audience (such as one's family and friends) unintentionally becomes available to a broader audience (such as companies and organizations) [41]. Given that users often fail to specify privacy policies that match their actual concerns, it is vital for businesses to take extra precautions when dealing with customer data, even when the underlying data is voluntarily disclosed and is publicly available. Following supplementary privacy-preserving methods provide organizations with a competitive advantage [39] and allows them to build and maintain customer trust to avoid the negative consequences that may arise from neglecting customers' privacy preferences and to build effective personalization while preserving privacy.

The solutions proposed to address social network privacy issues include studies that present a set of privacy-enhancing principles and guidelines to design personalization systems [24, 52], the works that suggest usable interfaces and visualization tools for specifying privacy policies [30, 28, 5, 15], as well as automated policy prediction models and frameworks [47, 13]. In this review, we focus on the latter direction. Specifically, we review the approaches that propose automated methods and utilize large social footprints available in online social networks to predict desired privacy settings. We conceptualize that an online social network is a virtual place in which individuals are allowed to create profiles and share their personal attributes, preferences, and opinions. In addition, they can connect to each other through different types of relationships and establish and maintain rich interactions with their peers on the network.

The remainder of this document is as follows: Section 2.2 describes the concept of privacy in the context of social networks. Then the major approaches on privacy preference inference in the context of social networks are provided in Section 2.3. Section 2.4 discusses a set of gaps based on the literature review. Finally, the manuscript concludes in Section 2.5

2.2 Privacy in Social Networks

Social networks are typically represented as a graph $G = \langle V, E \rangle$, where each user corresponds to a node $i \in V$. An edge $(i, j) \in E$ in the graph indicates some sort of social connection between the two users i and j. The labeling function F can be defined as $F : V \rightarrow R$, where V is the set registered users and $R = \{r_1, r_2, ..., r_n\}$ is the finite set of the possible relationships connecting the users. A relationship r_k can be either bidirectional (e.g., friend relation in Facebook) or unidirectional (e.g., follower relation in Twitter). In addition, each user can have a set of properties and profile items $P = \{p_1, p_2, ..., p_n\}$ that indicates who a user is in the social network, such as their identity and personal information. Users may also be associated with a set of contents $C = \{c_1, c_2, ..., c_n\}$ that describes what a user has exposed in the social network, such as uploaded text, images, videos, and other data items created through various activities in the network.

In the majority of the current social networks, privacy settings are described in terms of access control for the shared profile and content items. Most of the popular social networking sites such as Facebook and Google+ allow users to specify their privacy settings by controlling "who sees what" of their data. Varying granularity degrees of privacy specification are typically supported for both "who" and "what" variables. For example, users can set the information visibility as either public or private; or they can assign various specifications for different groups of their social contacts, or even different social contacts individually. Likewise, users may be allowed to specify privacy attributes for all of their published items at once, different categories of items, or each piece of shared items individually. In addition, the information access control can be either specified as a binary value (e.g., *allow* and *deny*) or on a nominal scale (e.g., *view*, *comment*, and *re-share*).

For instance, consider the partial network of a user presented in Figure 2.1, where the focal user U is connected with three social contacts u_1, u_2, u_3 through a similar relation type (i.e., friendship). Suppose this user has one published profile item $p = \{p_1\}$ and two shared content items $c = \{c_1, c_2\}$. Table 1 represents a possible privacy specification for user U in a social network, where access control is specified at a binary level for each of the social contacts and each of the published items separately. As discussed earlier, managing such

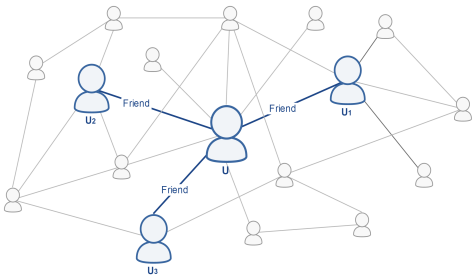


Figure 2.1: An example of a social network with focal user U.

	u_1	u_2	u_3
p_1	allow	deny	deny
c_1	allow	allow	allow
c_2	deny	allow	deny

Table 2.1: An example privacy specification in a social network.

privacy settings can be a cumbersome and a tedious task for hundreds of social contacts and shared items and so is often overlooked by users. As such, some researchers have utilized online social footprints and have proposed semi-automatic or automatic techniques to derive privacy policies that are similar to what current social networks provide as their privacy settings (e.g., [11]). Instead of focusing on privacy configurations that are specific to a particular social networking site, a set of studies attempted to use such footprints to characterize users' general privacy preferences. In these works, privacy preferences of users can be determined by mapping users to a binary, numeric, or an ordinal scale of desired privacy (e.g., [9]). Here, we attempt to provide a comprehensive review of the studies that take either of the approaches to infer privacy attributes.

2.3 Literature Review

Due to the growing interest in regular communication and information sharing over online social media, research on these platforms gained great attention in the last few decades. The rich information available in social media can greatly benefit individuals, communities and societies, businesses, politicians and governments, as well as scholars. The prior works on automatic detection of privacy preferences are categorized based on the type of the data used to make the prediction. Therefore, Section 2.3.1 explains the studies that have relied on the potential links between personal characteristics of the users and their privacy preferences to infer the privacy attributes. Section 2.3.2 presents the algorithms that are primarily focused on the users' social context and ties. In addition, some researchers have

used the content published by users to derive their desired privacy features. These content-based approaches are reviewed in Section 2.3.3.

2.3.1 Personal and Profile Attributes

There is a large body of research linking demographic information as well as personal traits to privacy preferences. For example, various surveys have established a positive relation of age, education, and income linked with privacy concerns [24]. In the context of social media, various studies have been conducted to find possible connections between gender and age and privacy behaviour [14, 25, 8, 12]. For gender, the results are inconclusive as some of the studies found no gender difference to privacy settings, while some found female users to be more private. Similarity, even though Dey et al. [12] have shown that adults tend to be more private in social media, the research conducted by Christofides et al. [10] has shown that adults and adolescents exhibit similar privacy behaviour. Similar attempts have been also made to study the possible connections between location [12, 10] as well as ethnicity [33] with privacy attributes.

In addition, personality attributes of people have been shown to be associated with their privacy attitudes and behaviour. For instance, in the context of location-based services, Junglas et al. [23] have researched the possible connections of the so-called Big Five personality traits (i.e., agreeableness, extraversion, emotional stability, openness to experience, and conscientiousness) and privacy concerns. They found that people who scored high on agreeableness, conscientiousness, and openness expressed lower levels of privacy concerns.

Despite the large body of works linking demographics and personality features to privacy concerns, to the best of our knowledge, there exist only a few works that have utilized this type of information to recommend privacy features. By using an online survey, Minkus and Memon [33] examined the privacy settings of users on Facebook and related their choices to demographic and personality features. The survey results are later used to build and deploy an online application, called MyPrivacy, that automatically recommends privacy settings. Their survey results provide evidence for earlier studies, indicating that personality traits and demographics are linked with privacy behaviour. In particular, they found that neuroticism, age, ethnicity, and the self-reported concern for privacy are related to the customized privacy settings of users on Facebook. Therefore, MyPrivacy first asks multiple questions from users to determine these attributes and then uses a supervised learning algorithm to recommended privacy settings. The evaluation of MyPrivacy showed positive subjective opinions of real Facebook users toward the tool. To recommend privacy

settings for a particular shared item, [36] proposed a supervised method as well. Their algorithm is built on a set of demographic features including age, gender, and location; along with a set of metadata associated with the shared item.

The lack of approaches focused on personal attributes can be due to the inconsistent and inconclusive results obtained from the studies that analyze the connections between such attributes and privacy preferences. These conflicting empirical differences may stem from the differences in what they measure as privacy preference. While some researchers may measure privacy behaviours to indicate privacy preference, some may be focused on privacy attitudes. Besides, in these works, the data collection process is often limited to specific and often rather small participant pools, such as people living in New York City [12] or college students [25]. Further studies with large and diverse participant sets may lead to consistent results that can be used reliably in automatic prediction tools. In addition, demographics and personal attributes may not be directly accessible through social media profiles, leading to the availability of a sparse set of attributes. Even though successful attempts have been made to extract this information from users' activities in their social network [2, 18, 40], these approaches are often complex and require extra computational resources.

2.3.2 Social Context

Compared to the use of personal attributes, a large set of studies have focused on the social context of the focal user to analyze and predict privacy-related features. These studies can be categorized into two primary groups. The first set of works mainly focuses on privacy in terms of information visibility to different groups of social contacts, often referred to as social circles. Hence, they propose approaches to assist users in creating and maintaining such social circles and their corresponding privacy policies.

While the aforementioned group of works are focused on partitioning and clustering users' social contacts, they do not make use of the valuable information hidden in their social context. Hence, another set of researchers has adapted techniques from the area of collaborative filtering to assign privacy policies to a user based on the preferences of other users. One approach to determine this set of users is to select them from within the social contacts of the focal user (e.g., friends in Facebook or followers in Twitter). This method follows the principle of homophily, which refers to the tendency of people to associate with similar individuals and has been observered in the context of online social networks [32]. As an alternative to the use of social contacts, a set of researchers has developed and used a set of similarity measures to select users with similar backgrounds and characteristics with the focal user.

Social Circle Management and Labeling

Given that users have on average hundreds of friends[1,2], specifying a policy that manages access to various information items is a difficult and a tedious task even for privacy-conscious users. As a result, with the aim of easing the process of privacy policy management, there have been attempts to automatically categorize users' social contacts into meaningful social circles. Some studies have moved beyond clustering and proposed techniques to infer user's preferred privacy settings for the created circles of contacts.

Adu-Oppong et al. [3] proposed that the clustering algorithm presented in [34] can be used to effectively create social circles of densely and closely connected contacts in unidirectional networks. Following this approach, (α,β)-clusters will be formed so that any node in a cluster is adjacent to at least a β-fraction of the cluster and any node outside of a cluster is adjacent to at most an α-fraction of the cluster. In a somewhat similar approach, Danezis [11] proposed an algorithm to cluster one's social contacts into circles that are closely related to each other and have many links within themselves, while having fewer links with those who are not in the circle. In [47], a large number of unique characteristics such as educational background, hobbies, and age are taken into account for clustering social contacts. A modified version of the apriori algorithms [4] is used to dynamically select clustering features based on the attributes of the social contacts of the focal user.

Jones and O'Neill [22] conducted user studies and interviews to understand user rationales when grouping their social contacts for the purpose of privacy management. As a result of this experiment, a set of six criteria commonly considered by users was identified. Since these criteria are related to the relationships between users, a network clustering algorithm, called SCAN [53] is used to group one's social network into various circles.

Some researchers have proposed supplementary techniques to clustering to recommend privacy settings for the created clusters. In [43], for instance, after the clusters of contacts are formed, the user is asked to label a number of randomly selected contacts from each cluster. Through the labeling process, the user indicates his/her willingness to share a specific item with them. A classifier is trained on the profile attributes as well as the network attributes of the labeled contacts to predict the privacy preferences of the user for unlabeled contacts relative to a specific object information item. They achieved an accuracy of 83% with 20% training.

Fang and LeFevre [13] built a privacy wizard that iteratively asks the user to label carefully-selected informative contacts. In these questions, the user specifies his/her willingness to share a specific piece of profile information with a social contact. To auto-

[1] http://www.theguardian.com/news/datablog/2014/feb/04/facebook-in-numbers-statistics
[2] http://news.yahoo.com/twitter-statistics-by-the-numbers-153151584.html

matically label other contacts, these labeled information is utilized in a classifier, wherein contacts are represented as feature vectors that encompass users' community structure and profile features such as age, gender, and education.

A classifier is developed in [26] to decide whether a data item should be visible to a contact of a user. Based on the assumption that the privacy labels that have been explicitly assigned to friends are correct, users' current privacy settings is used as the labeled data. Similar to [13], friends are then represented as a feature vector that includes their community attributes and their personal features. A classifier is then built on this feature set to assign privacy labels to unlabeled contacts.

Collaborative Filtering for Privacy Inference

Instead of categorization and labeling of the social context, some researchers have proposed methods to identify privacy preferences based on the privacy characteristics of the social context. Users' information sharing behaviour has shown to be extensively influenced by an inner circle of close friends [9]. For instance, the amount of private information shared by a user has shown to be correlated with the amount of private information shared by friends. Similarily, people with similar backgrounds tend to have similar privacy concerns [46]. These findings has motivated researchers to adapt collaborative filtering methods and determine one's privacy preferences from attributes of his/her network. Collaborative filtering uses the known preferences of a group of users to make recommendations of the unknown preferences of other users and is mainly utilized in the context of recommendation systems [50].

Squicciarini et al. [47, 45] provide an algorithm to form social circles based on users' characteristics such as their gender, hobbies, and occupation. These circles are further utilized to recommend privacy policies for newly added objects (i.e., added contacts or uploaded data items). When a new object is uploaded, the system first seeks the social circles that is most likely to deal with the object in a similar way as the user. Then the privacy policies used by the selected circle is the basis for predicting the privacy policy for the newly added object. Similar idea is applied to user-uploaded images [46], in which a policy prediction algorithm assigns a policy to a newly uploaded image based on the information captured from social circles.

In [44], active learning and the properties of the social graph are first used to detect a set of the most informative contacts to be labeled as training samples. In the labeling process, the user specifies whether he/she is willing to share a specific data item with the selected contact. Then an iterative semi-supervised approach is followed to label the other contacts of the user, where labels are propagated from labeled instances to unlabeled instances in

the social graph. This propagation is guided by the user similarity metric that is represented through edge weights. The similarity computation is based on profile information of contacts, their networks metrics, as well as the community structure. The evaluation results of this approach provided higher accuracy and precision compared to a supervised learning and a random walk based approach.

In the context of a location-based social network, Toch et al. [51] provide users with recommended privacy policies that similar users have previously selected. A large set of privacy policies is first clustered based on their location, time, and social group properties. Policies within each cluster are then ranked according to the number of policies they are similar to and their similarity degree. To recommend privacy policies, clusters that are relevant to the current user are selected based on the policies chosen by similar users (e,.g., users that are within the same Facebook network). Finally, top-ranked policies from the selected clusters are presented as recommendations.

Collaborative filtering is also followed in [16], where the authors take advantage of a set of profile features, user's interests implied in their social media, and their privacy configurations to find a set of users similar to the user of focus. In their approach, users are first characterized according to their privacy preference as either privacy fundamentalist, privacy pragmatist, or privacy unconcerned. Users' privacy decisions and settings regarding their photo albums is considered as an indication of their privacy preference. In particular, users are assigned to these three categories based on the number of their public, customized, and private photo albums. Then K-nearest neighbour algorithm is used to determine which privacy categorization the focus user belongs to. Based on the features of the assigned category, the system then recommends privacy settings.

2.3.3 Published Content

A frequent user activity on social networks is to publish and share content such as status messages, comments, images, and videos. All instances of shared data types can be used to draw inferences about the users' personality and preferences. In particular, natural language has been shown to be a reflection and a mediator of internal states [38]. Our words can reveal personality, emotional states and feelings, attention patterns, thought, and social situations [38, 17]. Therefore, a variety of automated content analysis techniques have been developed to measure such psychometric metrics from natural language. These methods range from the use of predefined dictionaries and taxonomies such as Language Inquiry and Word Count (LIWC) to more sophisticated computational algorithms that utilize complex data mining and machine learning based techniques.

In the context of privacy, Gill et al. [17] provide a set of privacy-related categories, each of which is associated with a number of words that are relevant in the semantic analysis of the privacy domain. The dictionary consists of 388 privacy related words that are grouped into eight high-level theoretically sound categories based on their semantic similarity. LIWC contains a large number of semantic categories with possible relevance to privacy features. Therefore, LIWC is used as the baseline for the evaluation of the privacy dictionary. The evaluation results indicate that the privacy dictionary is capable of capturing unique linguistic features in privacy language and is more reliable in detecting privacy-oriented content.

Caliskan-Islam et al. [9] used the privacy dictionary, along with a variety of methods and tools including topic modeling, named entity recognition, and sentiment analysis to automatically deem if a tweet contains private information. Annotated data were collected from Amazon Mechanical Turk (AMT), wherein AMT workers were asked to label collected tweets according to privacy categories. Then users are given privacy scores based on the amount of private information they published in their Twitter timeline. The timelines of these labeled users are then utilized in a supervised machine learning technique to assign privacy scores to unlabeled users based on their shared textual content.

The prediction model proposed in [36] follows a supervised machine learning approach to recommend privacy settings for a given post in Facebook. Besides the demographic features (as explained in Section 2.3.1), they used a set of content-based features associated with the post. The sentiment score of the post is included, along with some topical attributes. In addition, the entire bag-of-word representation of the content is taken into account, where only a set of words with a high tf-idf score is considered. Some contextual metadata elements are also used, such as the time of the day the post is shared.

Given an unstructured linguistic content published by a user, [48] first detects sensitive information such as phone number, address, and location from the text. Then the model proposed in [29] is adopted to quantify the privacy risk of the user, wherein the identified sensitive parts are treated the same way as information items in [29]. In [29], a mathematically sound model is developed, taking into account the sensitivity and visibility of the shared items. The proposed model provides users with a privacy score that quantifies the potential privacy risk of the user. The premise behind their model is that the more sensitive information the user discloses, the higher his or her privacy risk. In addition, the more visible the shared information becomes in the network, the higher the privacy risk.

In the context of image sharing, [46] uses the previous images by users and their corresponding privacy policies to assign a privacy policy to a new image. Image clustering and policy association mining are used for privacy generation. However, if the user is new

or there have been significant changes to the user's privacy trends, users' social context is used to predict the policy as explained in Section 2.3.2. To detect images with private content, Zerr et al. [54] used a variety of visual features, such as the occurrence of faces, in a supervised learning algorithm. They also utilized the textual metadata associated with images and found correlations between topics and the content of private images. For instance, topics used to describe personal concepts, emotions and sentiment, and human body were shown to be mostly used for private images. On the other hand, topics related to nature, architecture, and inanimate objects have been mostly found in non-private images.

2.4 Discussion

Automatic privacy inference has received relatively little attention. In addition, our analysis of the related literature suggests that the research efforts are mainly focused on the prediction and recommendation of privacy settings that are specific to the underlying social network. On the other hand, attempts have been made to quantify users' privacy risk [29, 6] and users' current privacy level [9]. Ghazinour et al. [16] (discussed in 2.3.2) characterized individuals by classifying them into different levels of privacy concerns based on their privacy settings on their photo albums. This lack of work may be attributed to the fact that privacy is context-dependent issue [37], making it challenging to develop generic methods. However, a recent study in the context of mobile applications has revealed that despite the diversity of privacy preferences, users can be clustered into a set of meaningful privacy profiles that effectively captures their desired privacy [27]. These studies imply the potential of characterizing users according to their platform-independent privacy preferences.

Another research gap we identified is the limited set of data types that prior studies have focused on. These data types are often the users' profile features, social and network attributes, and the content of communications. Many other data types are left unexplored in the literature [41, 7]. For instance, *ratings/interests* of users can be of value in gaining insight into one's privacy preferences and latent attributes, and are often readily accessible in users' profiles. However, to the best of our knowledge, only [16] has used it to infer privacy. Another example is *contextual data*. This data type refers to the property of an item that is made explicit and is provided with semantics, such as the tags provided in Facebook images and status messages or mentions in tweets. Although users' privacy preferences may be revealed from these contextual data, researchers have not used them in detecting privacy features.

By analyzing the approaches used, it can be seen that the existing literature lacks a study of hybrid techniques and of mixed data types. However, in similar areas such as

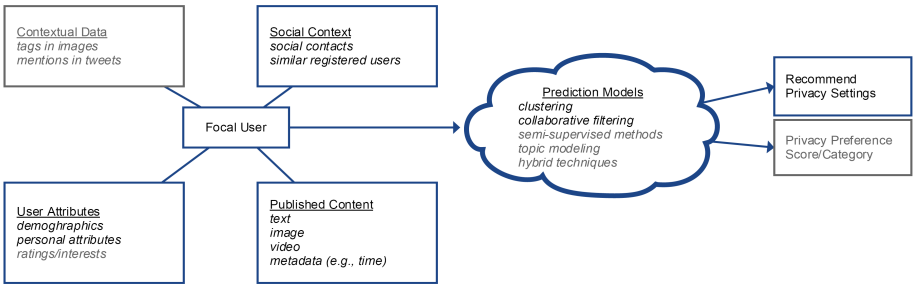


Figure 2.2: An overview of privacy prediction approaches in terms of their input, techniques, and goals.

recommendation systems, hybrid approaches have been shown to be very effective and can offset limitations of either approach and improve the prediction performance [50].

Many of the reviewed studies have utilized supervised methods to classify and predict privacy attributes. However, they require labeled input, and may not seem feasible in the context of social media, where the labeled information normally constitutes a very small portion of the available data. In such a context, unsupervised and semi-supervised techniques can be of great interest. In particular, semi-supervised techniques, which have the advantage of utilizing fewer labeled data to achieve better predictions, can be a potential research avenue to explore further. A graph-based semi-supervised method has been proposed to effectively capture privacy preference [44], and other methods such as Expectation and Maximization (EM), topic modeling, and co-training need to be investigated further. For instance, co-training has been successfully used to detect users' latent personal attributes in social networks [35].

Figure 2.2 provides an overview of our reviewed studies in terms of their input, proposed techniques, as well as their goals and purposes. Examples of each of these elements are also provided. The figure also indicates the research gaps we discussed above with gray boxes. The discussed limitations of prior studies call for further attempts to deeply analyze how different facets of large online social footprints can be utilized to effectively characterize users' privacy preferences.

2.5 Summary and Conclusion

Mining the treasure trove that exists in social media has tremendous potential for companies to improve the customer experience through personalization and targeted marketing. However, customers may not be willing to be profiled online due to their privacy concerns.

On the other hand, users' privacy concerns are often not well translated into their social network privacy configuration, resulting in generating data that is accessible to the public. It is a dilemma for companies whether to use one's publicly available data to provide valuable personalized content or not to use such data to avoid disregarding privacy preferences. One potential direction to manage this dilemma is to develop novel algorithms and techniques that take advantage of users' social footprints and characterize their privacy behaviour and attitude. In this document, we reviewed the existing literature on automatic privacy preference inference in the context of social networks.

We categorized and reviewed the existing studies on privacy preference inference according to the data type of focus, namely demographics and profile features, social context and network features, as well as the shared content. The potential and limitations of the approaches are further discussed, where a set of gaps are identified in the literature. Based on our study of the literature, we call for more studies of general user modeling and characterization according to their privacy preference. In addition, researchers studying privacy detection are encouraged to use a wider range of data types available in social media as well as hybrid techniques to make the predictions.

Bibliography

[1] Alessandro Acquisti and Ralph Gross. Imagined communities: Awareness, information sharing, and privacy on the Facebook. In *Proceedings of the Conference on Privacy Enhancing Technologies*, pages 36–58, 2006.

[2] S. Adali and J. Golbeck. Predicting personality with social behavior. In *Proceedings of the IEEE/ACM Conference on Advances in Social Networks Analysis and Mining*, pages 302–309, 2012.

[3] Fabeah Adu-Oppong, Casey Gardine, Apu Kapadia, and Patrick Tsang. Social circles: Tracking privacy in social networks. In *Proceedings of the Symposium on Usable Privacy and Security*, 2008.

[4] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. In *Proceedings of the Conference on Very Large Databases*, pages 487–499, 1994.

[5] Mohd Anwar, Philip W.L. Fong, Xue-Dong Yang, and Howard Hamilton. Visualizing privacy implications of access control policies in social network systems. In Joaquin Garcia-Alfaro, Guillermo Navarro-Arribas, Nora Cuppens-Boulahia, and Yves Roudier, editors, *Data Privacy Management and Autonomous Spontaneous Security*, volume 5939 of *Lecture Notes in Computer Science*, pages 106–120. Springer Berlin Heidelberg, 2010.

[6] Justin Lee Becker and Hao Chen. Measuring privacy risk in online social networks. In *Web 2.0 Security and Privacy Workshop*, 2009.

[7] Michael Beye, ArjanJ.P. Jeckmans, Zekeriya Erkin, Pieter Hartel, ReginaldL. Lagendijk, and Qiang Tang. Privacy in online social networks. In *Computational social networks: security and privacy*, pages 87–113. 2012.

[8] Danah Boyd and Eszter Hargittai. Facebook privacy settings: Who cares? *First Monday*, 15(8):1–24, 2010.

[9] Aylin Caliskan Islam, Jonathan Walsh, and Rachel Greenstadt. Privacy detective: Detecting private information and collective privacy behavior in a large social network. In *Proceedings of the Workshop on Privacy in the Electronic Society*, pages 35–46, 2014.

[10] Emily Christofides, Amy Muise, and Serge Desmarais. Hey mom, what's on your facebook? Comparing facebook disclosure and privacy in adolescents and adults. *Social Psychological and Personality Science*, 3, 2011.

[11] George Danezis. Inferring privacy policies for social networking services. In *Proceedings of the ACM Workshop on Security and Artificial Intelligence*, pages 5–10, 2009.

[12] R. Dey, Z. Jelveh, and K. Ross. Facebook users have become much more private: A large-scale study. In *Proceedings of the Conference on Pervasive Computing and Communications Workshops*, pages 346–352, 2012.

[13] Lujun Fang and Kristen LeFevre. Privacy wizards for social networking sites. In *Proceedings of the International Conference on World Wide Web*, pages 351–360, 2010.

[14] Joshua Fogel and Elham Nehmad. Internet social network communities: Risk taking, trust, and privacy concerns. *Computers in Human Behavior*, 25(1):153–160, 2009.

[15] Bo Gao and Bettina Berendt. Circles, posts and privacy in egocentric social networks: An exploratory visualization approach. In *Proceedings of the IEEE/ACM Conference on Advances in Social Networks Analysis and Mining*, pages 792–796, 2013.

[16] Kambiz Ghazinour, Stan Matwin, and Marina Sokolova. Monitoring and recommending privacy settings in social networks. In *Proceedings of the Joint EDBT/ICDT Workshops*, pages 164–168, 2013.

[17] Alastair J. Gill, Asimina Vasalou, Chrysanthi Papoutsi, and Adam N. Joinson. Privacy dictionary: A linguistic taxonomy of privacy for content analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3227–3236, 2011.

[18] Jennifer Golbeck, Cristina Robles, and Karen Turner. Predicting personality with social media. In *CHI Extended Abstracts on Human Factors in Computing Systems*, pages 253–262, 2011.

[19] Ralph Gross and Alessandro Acquisti. Information revelation and privacy in online social networks. In *Proceedings of the ACM Workshop on Privacy in the Electronic Society*, pages 71–80, 2005.

[20] D. Irani, S. Webb, Kang Li, and C. Pu. Large online social footprints–An emerging threat. In *Proceedings of the Conference on Computational Science and Engineering*, volume 3, pages 271–276, 2009.

[21] EricJ. Johnson, Steven Bellman, and GeraldL. Lohse. Defaults, framing and privacy: Why opting in-opting out. *Marketing Letters*, 13(1):5–15, 2002.

[22] Simon Jones and Eamonn O'Neill. Feasibility of structural network clustering for group-based privacy control in social networks. In *Proceedings of the Symposium on Usable Privacy and Security*, pages 9:1–9:13, 2010.

[23] Iris A. Junglas, Norman A. Johnson, and Christiane Spitzmüller. Personality traits and concern for privacy: An empirical study in the context of location-based services. *European Journal of Information Systems*, 17(4):387–402, 2008.

[24] Alfred Kobsa. Privacy-enhanced personalization. *Communications of ACM*, 50(8):24–33, 2007.

[25] Kevin Lewis, Jason Kaufman, and Nicholas Christakis. The taste for privacy: An analysis of college student privacy settings in an online social network. *Journal of Computer-Mediated Communication*, 14(1):79–100, 2008.

[26] Qingrui Li, Juan Li, Hui Wang, and A. Ginjala. Semantics-enhanced privacy recommendation for social networking sites. In *Proceedings of the Conference on Trust, Security and Privacy in Computing and Communications*, pages 226–233, 2011.

[27] Jialiu Lin, Bin Liu, Norman Sadeh, and Jason I Hong. Modeling users' mobile app privacy preferences: Restoring usability in a sea of permission settings. In *Symposium on Usable Privacy and Security*, 2014.

[28] Heather Richter Lipford, Andrew Besmer, and Jason Watson. Understanding privacy settings in Facebook with an audience view. In *Proceedings of the Conference on Usability, Psychology, and Security*, pages 2:1–2:8, 2008.

[29] Kun Liu and Evimaria Terzi. A framework for computing the privacy scores of users in online social networks. *ACM Transactions on Knowledge Discovery from Data*, 5(1):6:1–6:30, 2010.

[30] Alessandra Mazzia, Kristen LeFevre, and Eytan Adar. The PViz comprehension tool for social network privacy settings. In *Proceedings of the Symposium on Usable Privacy and Security*, pages 13:1–13:12, 2012.

[31] Craig R.M. McKenzie, Michael J. Liersch, and Stacey R. Finkelstein. Recommendations implicit in policy defaults. *Psychological Science*, 17(5):414–420, 2006.

[32] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001.

[33] T. Minkus and N. Memon. Leveraging Personalization To Facilitate Privacy. *ArXiv e-prints*, 2014.

[34] Nina Mishra, Robert Schreiber, Isabelle Stanton, and RobertE. Tarjan. Clustering social networks. In Anthony Bonato and Fan R.K. Chung, editors, *Algorithms and Models for the Web-Graph*, volume 4863 of *Lecture Notes in Computer Science*, pages 56–67. Springer Berlin Heidelberg, 2007.

[35] Mingzhen Mo, Dingyan Wang, Baichuan Li, Dan Hong, and I. King. Exploit of online social networks with semi-supervised learning. In *Proceedings of the Joint Conference on Neural Networks*, pages 1–8, 2010.

[36] Kaweh Naini Djafari, IsmailSengor Altingovde, Ricardo Kawase, Eelco Herder, and Claudia Niederée. Analyzing and predicting privacy settings in the social web. In Francesco Ricci, Kalina Bontcheva, Owen Conlan, and Samus Lawless, editors, *User Modeling, Adaptation and Personalization*, volume 9146 of *Lecture Notes in Computer Science*, pages 104–117. Springer International Publishing, 2015.

[37] Helen Nissenbaum. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press, 2009.

[38] James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577, 2003.

[39] Sőren Preibusch, Dorothea Kübler, and AlastairR. Beresford. Price versus privacy: An experiment into the competitive advantage of collecting less personal information. *Electronic Commerce Research*, 13(4):423–455, 2013.

[40] Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. Classifying latent user attributes in twitter. In *Proceedings of the International Workshop on Search and Mining User-generated Contents*, pages 37–44, 2010.

[41] Christian Richthammer, Michael Netter, Moritz Riesner, Johannes Sänger, and Gnther Pernul. Taxonomy of social network data types. *EURASIP Journal on Information Security*, 2014(1), 2014.

[42] SAS. Finding the right balance between personalization and privacy. *SAS Report*, 2015.

[43] M. Shehab, G. Cheek, H. Touati, A.C. Squicciarini, and Pau Cheng. User centric policy management in online social networks. In *Proceedings of the IEEE Symposium on Policies for Distributed Systems and Networks*, pages 9–13, 2010.

[44] Mohamed Shehab and Hakim Touati. Semi-supervised policy recommendation for online social networks. In *Proceedings of the Conference on Advances in Social Networks Analysis and Mining*, pages 360–367, 2012.

[45] A. Squicciarini, S. Karumanchi, D. Lin, and N. DeSisto. Automatic social group organization and privacy management. In *Proceedings of the Conference on Collaborative Computing: Networking, Applications and Worksharing*, pages 89–96, 2012.

[46] A.C. Squicciarini, Dan Lin, S. Sundareswaran, and J. Wede. Privacy policy inference of user-uploaded images on content sharing sites. *IEEE Transactions on Knowledge and Data Engineering*, 27(1):193–206, 2015.

[47] Anna Squicciarini, Sushama Karumanchi, Dan Lin, and Nicole DeSisto. Identifying hidden social circles for advanced privacy configuration. *Computers & Security*, 41:40 – 51, 2014.

[48] A. Srivastava and G. Geethakumari. Measuring privacy leaks in online social networks. In *Proceedings of the Conference on Advances in Computing, Communications and Informatics*, pages 2095–2100, 2013.

[49] Katherine Strater and Heather Richter Lipford. Strategies and struggles with privacy in an online social networking community. In *Proceedings of the British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction*, pages 111–119, 2008.

[50] Xiaoyuan Su and Taghi M. Khoshgoftaar. A survey of collaborative filtering techniques. *Advances in Artificial Intelligence*, pages 1–19, 2009.

[51] Eran Toch, Norman M. Sadeh, and Jason Hong. Generating default privacy policies for online social networks. In *Extended Abstracts on Human Factors in Computing Systems*, pages 4243–4248, 2010.

[52] Eran Toch, Yang Wang, and LorrieFaith Cranor. Personalization and privacy: A survey of privacy risks and remedies in personalization-based systems. *User Modeling and User-Adapted Interaction*, 22(1-2):203–220, 2012.

[53] Xiaowei Xu, Nurcan Yuruk, Zhidan Feng, and Thomas AJ Schweiger. SCAN: A structural clustering algorithm for networks. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 824–833, 2007.

[54] Sergej Zerr, Stefan Siersdorfer, Jonathon Hare, and Elena Demidova. Privacy-aware image classification and search. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 35–44, 2012.

Chapter 3

Privacy and Profile Attributes in Twitter

3.1 Introduction

The increasing levels of engagement in online social media have led to the accumulation of large social footprints left by millions of users on a daily basis. This massive source of information can deliver relevant information in the right context, leading to tremendous opportunities for both businesses and individuals. However, to effectively harness this treasure trove of data, it is imperative to address possible privacy complications. Such privacy issues are especially concerning due to the disparity between users' privacy behaviours and their attitudes in social media [6, 20], making their current privacy settings unreliable. As such, methods that can detect users' privacy preferences are desired so that data related to the privacy-concerned users can be discarded.

This privacy dichotomy may be due to users' misconceptions regarding the visibility of their data [20], the complex privacy specification interfaces [4], as well as their false, yet common, perception of the default setting as the recommended privacy policy [12]. In addition, even privacy-aware users may decide to choose public settings for the anticipated social gain, while they may not be willing to be profiled online by business and companies.

We argue that user *social footprints* [7] in social media environments can characterize their privacy preferences, offering an alternative and reliable source for the detection of privacy preferences. The social footprints are available in three types of social media data: users' profile attributes, their social context and ties, and their published content. In this study, we focus on the analysis of the profile attributes to explore their potential links to the

A version of this chapter has been published in the *companion volume of the World Wide Web conference proceedings (WWW MSM'6).*

user privacy preferences. In particular, we analyzed profile attributes of Twitter accounts to examine whether people with different levels of privacy setting configure these attributes differently.

Privacy configuration in Twitter is relatively simple and follows a binary specification. The Twitter users can follow the default *public* setting, which indicates that their tweets and follower/friend lists are accessible by the public. Alternatively, they can change the setting to *protected*, which makes their tweets and follower/friend lists accessible only by their approved followers. It is noteworthy that the users' profile attributes are visible to and accessible by the public and the Twitter API regardless of their privacy settings.

We analyzed a set of users' profile features and descriptions that are readily available from their Twitter accounts. We also developed and analyzed three additional features based on the existing profile attributes. Based on the analysis results, a feature set is developed and utilized in multiple classifiers to automatically detect the users with the protected privacy setting. Compared to the users' social network structure and their content-related features, their Twitter profiles contain very limited information. Despite this limitation, our classifier has obtained an F-score of 0.71, which improves a random and a naive baseline by over 20%. This finding can have implications for designing privacy-preserving personalization tools and indicates the value of profile attributes in the detection of privacy-concerned users.

The remainder of this paper is as follows: Section 3.2 reviews the earlier studies on privacy prediction in social media and the user attribute classification in Twitter. Section 3.3 describes our data collection process and the profile features in our dataset. Section 3.4 explains the analysis of the profile attributes for the users with the protected and default privacy settings. Section 3.5 presents the selected feature set and the evaluation of our classification. Finally, the paper concludes with our contributions and research plans.

3.2 Related Work

Detecting privacy attributes in social media. To address users' privacy concerns, three main approaches have been reported in the literature: the use of privacy-enhancing principles for designing personalization systems [9, 21], the design of usable interfaces and visualizations that enable users to specify their privacy policies [11, 1, 5], and the computational methods that automatically predict users' privacy preferences [18, 4]. Regardless of its potential value, the prediction approach has received much less attention compared to the other two [8]. Additionally, the majority of the prediction models are structured on the users' social networks and their generated content, while only a few have utilized profile

attributes and personal characteristics [8].

Minkus and Memon [13]conducted an online questionnaire study to examine the privacy settings of Facebook users and related the settings to their demographic and personality features. Their survey results were later used to build and deploy an online application, called MyPrivacy, that automatically recommends privacy settings. Specifically, MyPrivacy first asks multiple questions from users to determine the demographic and personality attributes and then uses a supervised machine learning algorithm to make recommendations based on the users' privacy settings. The evaluation of MyPrivacy showed that real Facebook users had positive subjective opinions toward the tool. However, this tool is semi-automated and requires direct input from the users.

Similarily, in [14] a supervised learning algorithm is proposed and is built on a large set of features to recommend privacy settings. These features include metadata elements regarding a shared item as well as users' demographic and profile features, such as the number of users' Facebook posts and their friends. It is worth noting that the algorithm is developed to recommend privacy settings for a particular shared item, such as individual Facebook posts, as opposed to predicting users' privacy preferences in general.

Dong et al. [3] proposed a privacy prediction model that takes into account social media behavioral analogs to psychological variables that are known to affect users' disclosure behavior. Some of the identified analogs are based the profile features of the users. For instance, user's trustworthiness is calculated based on the ratio of their followers to the total number of their social contacts.

Twitter user attribute detection. A variety of techniques and social data types have been utilized to detect Twitter users' latent attributes. For instance, [16] uses the profile fields, tweeting behaviour, tweet content as well as the network structure to understand political affiliation, ethnicity, as well as users' affinities to a specific business.

Another example is the work of Nguyen et al. [15], in which the connection between the language use and the age is studied in the Twitter context. The authors built a classifier based on the tweet unigrams and classified users according to their age category, life stage, and their age. In their study, Rao et al. [17] found distinctive variations in the language use of Twitter users across different gender, age, regional origin, and political orientation.

These studies on automatic detection of privacy preferences and Twitter user attributes provide valuable insights into the potentialities and limitations of the available features and techniques in the detection of users' latent attributes. Our purpose for using the profile attributes to detect users' general privacy preferences has received very little attention in the literature [3]. Although it is argued that the users' profile fields may not include enough good quality information for user classification purposes [16], our results suggest that they

can be promising in the detection of social privacy.

3.3 Data Source

3.3.1 User Selection

We have built a directory of Twitter users by crawling a number of famous Twitter accounts and collecting their followers. Table 3.1 presents these Twitter accounts, the number of collected followers each account has, as well as the percentage of the number of protected followers to the number of collected followers of the account. Please note that in the table the numbers are rounded (e.g., the number of CNN followers is 12,246,514 and we rounded it to 12.2M), and the percentages are calculated based on the exact numbers and then rounded.

It is known that a large number of Twitter accounts are inactive thus are more likely not to follow any account [10]. In addition, such accounts are more likely to follow the default public privacy setting compared to the active accounts. Our underlying set of Twitter accounts follow at least one account (e.g., follow "Bill Gates" as shown in Table 3.1). Therefore, the percentage of protected accounts is anticipated to be higher than that of Twitter accounts in general. On average, 4.8% of Twitter users have protected accounts [10]. As can be seen in Table 3.1, the percentage of the protected accounts is similar or above the average for all our follower sets, which confirms our expectation.

As shown in the table, the percentage of the protected accounts for "CNN Breaking News" is 11%, considerably higher than the other follower sets and the average percentage in Twitter. One possibility is that CNN tweets may cover more topics that attract private users in Twitter, such as privacy-related news, compared to the other accounts in the table. This might also be true for other Twitter news accounts as we expect the news accounts to cover a wider variety of the topics that are related to and have a potential impact on the users' lives and societies. Further analysis of the percentage of the protected Twitter accounts that follow other Twitter news accounts and the analysis of the news content are expected to offer more insights on this issue.

As mentioned in the introduction section, the users' privacy settings may not reflect their actual privacy preferences [6, 20]. In other words, Twitter users who have the default public privacy setting may, in fact, be private. Therefore, these users' profile attributes reflect the private users' configurations. To study the possible differences in how protected and public profiles are configured, we need a set of accounts wherein such a situation is minimized. We thus chose the CNN follower set given that it has a higher percentage of

Table 3.1: A set of popular accounts in Twitter and the statistics of their collected follower sets.

Account	#Followers	#Protected	%Protected
Facebook	4.8M	261K	5%
CNN Breaking News	12.2M	1.5M	11%
Youtube	14.8M	788K	5%
Bill Gates	5M	374K	7%
Obama	23.8M	52M	7%
Katy Perry	75.3M	52M	7%

the protected accounts.

To ensure that inactive users are not taken into account, we filtered the CNN follower set to include only those who have published at least ten tweets. In Twitter, the accounts that belong to key individuals and brands are marked as *verified*. To focus on the general public, we removed these *verified* accounts from the set. Finally, we filtered the set to include only those accounts whose language is set to English. By applying these three criteria, we were able to select roughly 850K protected accounts from the original 1.5M accounts. Our public accounts also dropped from 12.2M to almost 10M. To have a relatively balanced set of accounts, we randomly pulled 1M public accounts from this set.

3.3.2 Profile Features

Each Twitter account is associated with a set of profile attributes. A set of profile features is configured by the account holder, often when the account is created, and is intended to represent who the user is in the network. Examples of such profile attributes include the username, the profile image, and the location information. Another set of attributes, which are also specified by the user, is related to the settings of their account. For instance, they can specify whether or not they want their tweets to be geo-tagged by setting the value of the geo-enabled attribute. Other examples of such attributes include their preferred interface language and whether or not their account should be withheld from certain countries. Finally, a set of contextual attributes is specified by Twitter. For instance, the time of the account creation, the number of tweets published by each user, and the number of followers/friends.

A subset of the available profile attributes is deemed to be relevant for our purpose and selected in our analysis. This list is shown below, along with a brief description of the attributes. Please note that the descriptions are adapted from Twitter API specification[1].

[1]https://dev.twitter.com/overview/api/users

- Name: The name of the user.
- Username: The alias that users identify themselves with.
- Description: A piece of text users provide to describe their account.
- URL: A URL provided by the user in association with their profile.
- Location: The user-defined location for this account's profile.
- Geo-enabled: When true, indicates that the user has enabled the possibility of geo-tagging their Tweets.
- Default Image: When true, indicates that the user has not uploaded their picture and a default avatar is used instead.
- Default Profile: When true, indicates that the user has not altered the theme or background of their user profile.
- Favorite Count: The number of tweets this user has favorited in the account's lifetime
- Tweet Count: The number of tweets issued by the user.
- Follower Count: The number of followers this account currently has.
- Friend Count: The number of users this account is following.
- List Count: The number of public lists that this user is a member of.

3.4 Analysis of Profile Attributes

In our analysis, the geo-enabled attribute, default profile, and default image are all binary attributes and are studied as binary variables. Similarly, the numeric attributes of the favorite count, the tweet count, the follower/friend count, as well as the list count are analyzed as is.

Based on the declared name in the Twitter account, we created a binary and a numeric attribute: we matched the account name against a directory of English names to check whether any part of their declared name is indeed a person's name in the dictionary. We also counted the number of parts in the account name that are available in the dictionary. For example, an account name that has only the first name matched has the value of 1, whereas an account name that has both the first and the last name appearing in the dictionary has the value of 2. For the Twitter account's username, we checked to see if it contains the declared name of the user. For description, URL, and location attributes, we simply checked whether the corresponding piece of information is provided by the user.

Finally, we used a linguistic analysis tool to study the account's profile descriptions to understand how the users of different privacy settings describe themselves in Twitter. The analysis results of the surface-based profile features are provided in Section 3.4.1, while Section 3.4.2 explains the results of the linguistic analysis of the profile descriptions.

3.4.1 Surface-based Profile Features

Table 3.2 presents the selected binary features, along with the percentage of the protected and public accounts for which these binary attributes hold. Although the Chi-Square test results suggest statistical significance for all the features, the effect size values suggest that only three features have practically different values in the public versus the protected accounts: *has location*, *is geo-enabled*, and *is default profile*. We calculated the effect size using Cramer's V and followed the convention to interpret the value [2]. A Cramer's V needs to be at least .1 to show a practically significant effect in reality. As shown in the table, a larger percentage of protected accounts has enabled their geo-tagging feature and has provided information for the location attribute. Besides, more protected accounts have changed their default profile settings compared to the public accounts.

Table 3.3 provides an average value of our numeric features in the two types of accounts. We calculated the effect size using Cohen's d, and followed the convention to interpret the value [2]. Specifically in our study context, a feature's Cohen's d value needs to be at least .2 to be considered as a practically useful feature that distinguishes the protected and public accounts. Although the t-test results suggest statistical significance for all the features, the effect size values suggest that only the Tweet count feature has a practically different value in the public versus the protected accounts. The results show that on average, protected accounts tweet more often and this feature's effect is close to medium (d = .29) (see Table 3.3). The protected account seems to have a larger number of favorite tweets although the effect is still quite small (d = .09).

In general, the results are interesting and contrary to what we expected before the analysis. For example, we anticipated that because the protected accounts represent a more private or more privacy aware population, they would be less likely to enable the location tracking feature or change the default profile theme, or even tweet often. These findings, however, indicate otherwise.

3.4.2 Profile Descriptions: A Closer Look

As explained earlier, the Twitter users can provide up to 160 characters in the description field. In our set of the CNN followers, there are almost 500K of the protected accounts and

Table 3.2: Analysis of binary profile attributes of protected and public accounts.

Binary Attributes	%Protected	%Public	Effect Size
Has Name	71.17	68.86	0.02
Username Has Name	3.01	3.45	0.01
Has Description	56.79	51.69	0.05
Has URL	15.01	16.78	0.02
Has Location	64.70	49.05	0.15
Is Geo-enabled	39.78	25.59	0.15
Is Default Profile	33.26	71.17	0.38
Is Default Image	6.43	8.80	0.04

Table 3.3: Analysis of numeric profile attributes of protected and public accounts.

Numeric Attributes	Protected	Public	Effect Size
Favirote Count	189.32	115.43	0.09
Tweet Count	1389.16	384.55	0.29
Follower Count	80.71	166.78	0.03
Friend Count	255.78	242.76	0.03
List Count	1.01	0.93	0.0006
Name Count	1.10	1.07	0.04

roughly 500K of the public accounts that have descriptions. We used Language Inquiry and Word Count (LIWC 2015) to analyze the language categories in these descriptions. The LIWC program processes each text file word by word and compares them against a pre-built dictionary to detect the LIWC category that the word belongs to. After processing all the words in the text, LIWC calculates and outputs the percentage of each LIWC category. Before conducting the linguistic analysis by LIWC, we applied the following pre-processing steps on the descriptions:

- removed HTML characters
- replaced apostrophe elisions (e.g., I'm -¿ I am).
- replaced URLs with the word "url"
- replaced emoticons with their corresponding meanings (e.g., :) -¿ smile)
- removed punctuation marks
- replaced user handlers with the word "mention"

The LIWC dictionary is structured in a hierarchical format, wherein each category may encompass several sub-categories. Details about these categories can be found in the LIWC

Table 3.4: LIWC categories and their corresponding percentage for protected and public descriptions.

LIWC Category	Protected	Public	Effect Size
Function Words	37.51	33.68	0.15
Affect	8.68	7.50	0.10
Social Processes	11.08	10.75	0.02
Cognitive Processes	7.86	6.71	0.09
Drives and Needs	11.01	11.38	0.02
Relativity	10.03	10.23	0.0006

website [2]. Since the users' profile descriptions are usually very short (commonly between 8-10 words), the percentages provided by LIWC are often very small for the majority of the categories. Therefore, we only focused on the higher-level categories that are at the top of the LIWC hierarchy.

Table 3.4 provides these categories as well as their corresponding percentages for the protected and public accounts. Here, we dropped those LIWC categories that had less than 5% of matching words in the entire corpus of descriptions. In addition, LIWC outputs a set of summary dimensions along with the percentage of their matching words. Table 3.5 provides the summary variables deemed relevant and their corresponding percentages for the two sets of accounts. A t-test is performed for these categories, along with the effect size measured by Cohen's d. All the categories have statistically significant different values between the protected and the public accounts, but these differences are small based on the Cohen's d (see Table 3.4). It is still interesting to note that the protected account has a larger percentage of the *function words* and *affect words*, which being similar to our findings regarding the surface-based attributes is in contrast to our prior expectation.

In addition to the LIWC main categories, an analysis of the summary dimensions shows that protected accounts contain a smaller number of lengthy words (i.e., words with six or more letters). They use fewer words representing *analytical thinking* and *clout*. However, they have a higher percentage of words that bear *emotional tone* and *authenticity*. The differences are statistically significant based on the t-test results, but are not practically significant from the Cohen's d value (see Table 3.5).

[2] http://liwc.wpengine.com/

Table 3.5: LIWC summary variables and their corresponding values for protected and public descriptions.

Summary Dimension	Protected	Public	Effect Size
Six Letter Words	22.73	26.69	0.14
Analytical Thinking	75.64	84.04	0.15
Clout	66.83	72.99	0.10
Emotional Tone	98.58	96.55	0.05
Authentic	28.89	21.24	0.13

3.5 User Classification

We utilized the binary and numeric features introduced in Section 3.4.1, along with the LIWC features discussed in Section 3.4.2, in multiple classifiers to identify protected accounts. The classifications are conducted by the machine learning toolkit Weka [3] and the results are evaluated using stratified 10-fold cross validation.

For the classification, we modified some of the LIWC features. We changed the LIWC categories that matched less than 25% of the entire corpus to binary attributes (see Table 3.4 for category percentages), such as *affect*, *social processes*, and *drives and needs*. For instance, if a description contains any word that is categorized as an *affect word* in LIWC, the corresponding feature is set to 1; otherwise, it is set to 0. The remaining LIWC attributes are kept as is.

We also added the presence of four keywords as supplementary features. Throughout our keyword analysis of the descriptions, the two keywords of *follow* and *business* were found to be commonly present in the public descriptions compared to the protected ones. In addition, the public descriptions seem to *mention* other accounts more often than their protected counterparts. On the other hand, the word *smile* is frequently used in the protected descriptions, which can either be the use of the word directly or a smiley face replaced with the word *smile* in the cleaning phase.

Since our classification is conducted on the users who have a description, the feature that checks the presence of the description is of no value here and so is removed. In addition, Twitter users with protected accounts need to approve their followers, while any user can instantly follow the public accounts. Therefore, the differences in the follower counts may not necessarily stem from differences in privacy attitude; hence, the follower count is also removed from our feature set. In summary, our classifiers are built on a total of 26 profile features extracted from each account, consisting of 11 surface-based attributes, 11

[3] http://sourceforge.net/projects/weka/

Table 3.6: Evaluation of classification results.

Algorithm	Precision	Recall	F-score
Naive Bayes	0.66	0.67	0.66
Regression	0.71	0.70	0.71
Logistic	0.69	0.70	0.69
J48	0.68	0.66	0.67
KNN	0.67	0.59	0.63

attributes extracted by LIWC, as well as four keyword-based features.

Table 3.6 provides our evaluation results of multiple classifiers. Our best classification results are obtained using *ClassificationViaRegression* with a performance of 0.71. Since there is no comparable study in the literature, we used a random classifier as our baseline. Based on the exact numbers of protected and public accounts in our underlying set, 48% of the set is composed of protected accounts, while 52% are users with public accounts. Therefore, a random classifier will label protected instances with an F-score of 0.48. Our feature set outperforms this baseline across all algorithms, and improves the results by 23% in the best case.

In addition, we used a naive baseline to compare the results. This baseline decides to label users based on their geo-enabled feature. This rule is established based on how the Twitter *setting* configuration page interface is arranged. In this page, the section designed to change the geo-enabled attribute is placed right at the top of the one designed to modify the privacy setting. Therefore, this baseline naively assumes that the users who have changed their default geo-enabled field from *false* to *true* are aware of the default privacy setting and thus have changed their privacy setting to the *protected* mode as well. The naive baseline reaches an F-score of 50.77, which is roughly 21% worse that our best algortihm. These results suggest that even by only relying on the profile attributes, one may be able to automatically detect users' privacy preferences in social media. This finding is encouraging. We call for the analysis of other attributes of Twitter users and their potential relationships to their privacy behaviour.

3.6 Discussion

Characterizing user privacy preferences in social media is a difficult and challenging task that requires a careful examination of various aspects of the users' social footprints. One class of such footprints can be found in how users shape and build their account profiles. In this study, we identified possible connections between users' profile attributes in Twitter

and their privacy settings.

In particular, we found that protected accounts enable the geo-tracking feature more often compared to the public ones. As well, they tend to provide their location information and change their default profile theme. Besides, they tweet much more frequently. These differences, along with the common presence of emotion bearing and affect words in protected profile descriptions, can be associated with extraverted personality. Based on this interpretation, our finding is in contrast to earlier research on privacy in traditional settings, stating that introverts tend to be more privacy-concerned and are more likely to feel invaded when asked to reveal private information [19].

A possible speculation is that since users with protected accounts are aware that their tweets and accounts are private, they feel secure in this environment and are willing to voluntarily reveal more information about themselves and participate more actively in the network. On the other hand, users who are consciously following the public setting are utilizing a different strategy to protect their privacy, which is not including their location information, using a default theme, tweeting less, or using fewer function and affect words in their descriptions. If this is the case, then the users who are engaging in social media more actively (e.g., making changes to their profile attributes, sharing personal information, and revealing emotional states), tend to feel secure in the environment. Therefore, they are more likely to feel invaded by targeted advertising and marketing messages.

As discussed earlier, users' privacy behaviours may not necessarily match their privacy attitudes. Despite our efforts to choose a set of accounts with a minimal number of such false positives, our underlying set can still include public accounts that were meant to be protected. In spite of this issue, we obtained an F-score of 0.71, indicating the value and importance of profile attributes in the detection of privacy behaviour. Based on this finding, unsupervised or semi-supervised techniques can be developed to effectively identify the public accounts that belong to privacy-concerned people, taking into account user profile attributes.

3.7 Conclusions

Ongoing creation of large social footprints offers immense potentials for both business and the individual. However, as users' privacy concerns are often not well-translated into their privacy settings, their data may be unintentionally visible to the public and thus should not be used for user profiling purposes. Therefore, it is desirable to characterize users' privacy attitudes, allowing companies to make informed decisions whether to discard or use the publicly available data for business intelligence purposes.

In this study, we explored the benefits of using Twitter profile attributes to infer privacy settings. By building a feature set based on these attributes, we obtained an F-score of 0.71 for the detection of privacy-concerned accounts. The classifiers in our experiments consistently outperformed both a random and a naive baseline and proved to be of value for our task. To the best of our knowledge, this is the first study that attempts to find differences in how people with different privacy settings manage their profile attributes.

Our analysis of the profile features and our classifiers are based on CNN followers' accounts. To generalize these findings, we will conduct a similar analysis and classification process with other Twitter accounts. We also plan to explore other available data sources. For example, we may be able to offer more informed labeling of the users' privacy levels based on their network structures. Furthermore, the content of user tweets is expected to be of great potential toward user classification since natural language has been shown to be a reflection of internal states. We will also investigate the generalizability of our approach by analyzing similar feature sets across different social platforms.

Bibliography

[1] Mohd Anwar, Philip W.L. Fong, Xue-Dong Yang, and Howard Hamilton. Visualizing privacy implications of access control policies in social network systems. In Joaquin Garcia-Alfaro, Guillermo Navarro-Arribas, Nora Cuppens-Boulahia, and Yves Roudier, editors, *Data Privacy Management and Autonomous Spontaneous Security*, volume 5939 of *Lecture Notes in Computer Science*, pages 106–120. Springer Berlin Heidelberg, 2010.

[2] J. Cohen. *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Earlbaum Associates, 1988.

[3] Cailing Dong, Hongxia Jin, and Bart Knijnenburg. Predicting privacy behavior on online social networks. In *Proceedings of the AAAI Conference on Web and Social Media*, 2015.

[4] Lujun Fang and Kristen LeFevre. Privacy wizards for social networking sites. In *Proceedings of the International Conference on World Wide Web*, pages 351–360, 2010.

[5] Bo Gao and Bettina Berendt. Circles, posts and privacy in egocentric social networks: An exploratory visualization approach. In *Proceedings of the IEEE/ACM Conference on Advances in Social Networks Analysis and Mining*, pages 792–796, 2013.

[6] Ralph Gross and Alessandro Acquisti. Information revelation and privacy in online social networks. In *Proceedings of the ACM Workshop on Privacy in the Electronic Society*, pages 71–80, 2005.

[7] D. Irani, S. Webb, Kang Li, and C. Pu. Large online social footprints–An emerging threat. In *Proceedings of the Conference on Computational Science and Engineering*, volume 3, pages 271–276, 2009.

[8] Taraneh Khazaei, Lu Xiao, Rober Mercer, and Atif Khan. Detecting privacy preferences from online social footprint: A literature Review. In *Proceedings of the iConference*, 2016.

[9] Alfred Kobsa. Privacy-enhanced personalization. *Communications of ACM*, 50(8):24–33, 2007.

[10] Yabing Liu, Chloe Kliman-Silver, and Alan Mislove. The tweets they are a-changin': Evolution of Twitter users and behavior. In *Proceedings of the AAAI Conference on Weblogs and Social Media*, 2014.

[11] Alessandra Mazzia, Kristen LeFevre, and Eytan Adar. The PViz comprehension tool for social network privacy settings. In *Proceedings of the Symposium on Usable Privacy and Security*, pages 13:1–13:12, 2012.

[12] Craig R.M. McKenzie, Michael J. Liersch, and Stacey R. Finkelstein. Recommendations implicit in policy defaults. *Psychological Science*, 17(5):414–420, 2006.

[13] T. Minkus and N. Memon. Leveraging Personalization To Facilitate Privacy. *ArXiv e-prints*, 2014.

[14] Kaweh Naini Djafari, IsmailSengor Altingovde, Ricardo Kawase, Eelco Herder, and Claudia Niederée. Analyzing and predicting privacy settings in the social web. In Francesco Ricci, Kalina Bontcheva, Owen Conlan, and Samus Lawless, editors, *User Modeling, Adaptation and Personalization*, volume 9146 of *Lecture Notes in Computer Science*, pages 104–117. Springer International Publishing, 2015.

[15] Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. "how old do you think i am?" a study of language and age in twitter, 2013.

[16] Marco Pennacchiotti and Ana Popescu. A machine-learning approach to twitter user classification. In *Proceedings of the AAAI Conference on Weblogs and Social Media*, 2011.

[17] Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. Classifying latent user attributes in twitter. In *Proceedings of the International Workshop on Search and Mining User-generated Contents*, pages 37–44, 2010.

[18] Anna Squicciarini, Sushama Karumanchi, Dan Lin, and Nicole DeSisto. Identifying hidden social circles for advanced privacy configuration. *Computers & Security*, 41:40 – 51, 2014.

[19] Dianna L. Stone. Relationship between introversion/extraversion, values regarding control over information, and perceptions of invasion of privacy. *Perceptual and Motor Skills*, 62:371–376, 1986.

[20] Katherine Strater and Heather Richter Lipford. Strategies and struggles with privacy in an online social networking community. In *Proceedings of the British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction*, pages 111–119, 2008.

[21] Eran Toch, Yang Wang, and LorrieFaith Cranor. Personalization and privacy: A survey of privacy risks and remedies in personalization-based systems. *User Modeling and User-Adapted Interaction*, 22(1-2):203–220, 2012.

Chapter 4

Addressing Privacy Dichotomy in Twitter

4.1 Introduction

The notions of personalization and data privacy have always been intertwined since effective personalization heavily relies on collecting and processing personal and potentially sensitive user data. The tension between personalization and privacy is often referred to as the personalization-privacy paradox [35]. If privacy concerns are adequately addressed, personalization can bring customers and brands closer and can greatly increase firm revenues. Personalization at the expense of privacy violation, however, can easily drive customers away and may even arouse them to advocate against the firm. Therefore, finding the sweet spot between "individualized" and "invasive" communications is central to any successful personalized and customized marketing attempts [27].

Reconciling the personalization-privacy paradox can be particularly challenging in the context of social networking websites. In such platforms, users are able to protect their data by using privacy controls; however, studies on online behaviour indicate that there is a disparity between the privacy attitudes of social media users and their actual behaviour in specifying privacy policies [1, 18, 34]. These privacy decisions are often complex and require careful examination of the trade-offs between the potential social gain and possible privacy risks. Hence, many users avoid the hassle of privacy configuration and follow the default settings that are often inaccurately perceived as the recommended course of

A version of this chapter has been submitted for publication to *ACM Transactions on Social Computing*

action [34]. Nevertheless, users are normally unaware that the default privacy setting is commonly open and permissive (e.g., Twitter and Instagram). Even among the ones that make the effort to manage their privacy, many are still unaware of the implications of their decisions [20].

Meanwhile, individuals use social platforms to communicate with friends and family, to share personal details, beliefs, opinions, current activities, and even their location data. Such massive data sources accumulated on a daily basis can give insight into various facets of customer behaviour and provide unprecedented marketing opportunities for businesses. For instance, companies can leverage relevant information about each customer and provide valuable suggestions regarding the next best offer, marketing messages tailored to each individual, and right response for customer support services. However, due to the privacy dichotomy problem in social media, user data may be unintentionally visible to the public and thus should not be used for user profiling purposes. Effective use of social data, hence, requires privacy protection measures beyond the current privacy settings of the users.

The solutions proposed to address the personalization-privacy paradox in social media include the studies that present a set of privacy-enhancing principles for designing personalization systems [15, 37], the works that suggest usable interfaces and visualizations for privacy specification [22, 9], and the models to predict and infer privacy preferences from user activities [32, 7]. The latter direction has had modest success in predicting user's privacy preferences, yet its exploration has been very limited so far [13, 8]. Therefore, this approach is the focus of this study. In particular, we aim to a set of features to characterize the online behaviour of public user accounts that are meant to be private. Such a feature set can be used in a predictive model to raise a red flag, alerting businesses that the user of focus is probably not willing to be profiled online.

Our study is conducted on Twitter, wherein privacy control follows a simple binary specification. In Twitter, users can follow the default *public* setting, which indicates that their tweets and contact lists are available to the public and accessible by the Twitter API. Alternatively, they can change their privacy setting to *protected*, which makes their tweets and contacts only accessible by their approved followers. With the exception of profile attributes (e.g., name and username), data associated with *protected* accounts are not available through the Twitter API either, making it impossible to directly compare and analyze *protected* and *public* profiles. Instead, we primarily focused on privacy preferences as a localized attribute to gain insight into the attributes of users with different privacy preferences. Our analysis, experiments, and models are based on a social network of 23K Twitter users collected from a random user seed.

Homophily is a social phenomenon that is premised on the fact that "similarity breeds

connection" [23]. Earlier research on homophily in social networks suggests that the neighbourhood context carries substantial information about users and can be used to detect user preferences [16, 23]. In addition, preferences of social media may be influenced by their social contacts in the social network over time. Our primary analysis of adjacent neighbourhoods shows that privacy preferences are indeed localized in Twitter [14]. Besides, we published tweets from a subset of the collected accounts to a crowdsourcing platform and asked human annotators to judge the privacy preference of the tweet authors. We chose Amazon Mechanical Turk[1] (AMT) as the platform of focus. The AMT experiment results, as well, show that privacy preferences are localized on Twitter.

Therefore, we exploit the fact that privacy preferences are localized in the network and characterize ones privacy preferences by the degree of privacy concern expressed by his/her social contacts. In this work, the current privacy setting of the social contacts (i.e., *public* and *protected*) is considered as an indication of their privacy preference. In particular, we employed a privacy metric that quantifies the percentage of *protected* social contacts to all of the contacts of a user. Therefore, users with a higher privacy metric are the ones that are located in the more private neighbourhoods compared to those with a lower privacy metric. Based on our earlier analysis, we then assume that publicly available accounts located in private neighbourhoods are more likely to belong to a privacy-concerned user and to suffer from the privacy dichotomy problem. We further utilize this finding to characterize the behaviour of such users.

To find a set of attributes that are related to privacy preferences, we examine whether *public* users' Twittering behaviors are different if they are located in different neighbourhoods (with varying degrees of privacy concern). To probe these differences, we explore various behaviour indicators, including profile attributes that are readily available from Twitter accounts (e.g., tweet count and profile descriptions), linguistic aspects of tweets (e.g., use of function words, degree of authenticity, and the sentiment-related characteristics), and the attributes associated with their communication behaviour (e.g., reply and retweet count). According to the results, public accounts located within private neighbourhoods seem to share content that is more private according to the societal consensus [4]. This finding indicates that such users may be privacy-concerned, though they may be under the wrong impression that their data are protected in the network.

To discover latent features that are of interest to privacy-concerned users, we transformed each user in the Twitter network to a set of attributes representing the user (e.g., hashtags, unigrams, and topics used in their tweets). For each attribute node, we then calculate a privacy metric based on the number of *protected* and *public* accounts surrounding

[1] `https://www.mturk.com/`

the feature. Similar to our earlier analysis of the tweets, attributes that are more private are often seen within private neighbourhoods. For instance, the hashtags *#Annoyed* and *#SoTired* are more frequently used by *public* account with a large percentage of *protected* contacts. Since our work is based on preference locality and uses the known preferences of users to predict the unknown preferences of other users, it can be considered a collaborative filtering method [40].

The remainder of this paper is as follows: Section 4.2 reviews the earlier studies on automatic detection of privacy preferences. Section 4.3 describes the data collection process and provides a descriptive analysis of the collected Twitter data. Section 4.4 describes our AMT experiment and the results. Analysis of users placed in different neighbourhoods are provided in Section 4.5. Findings and limitations of the study are discussed in Section 4.6. Finally, Section 4.7 provides the concluding remarks and the future research plans.

4.2 Related Work

Prior studies on the automatic detection of social privacy preference have leveraged a variety of data types in the social networking websites. A set of studies has relied on the potential links between profile attributes of the users and their privacy preferences. Some researchers proposed algorithms that are primarily focused on the users' social context and ties. Some studies have used the content (e.g., text and images) published by users to derive their desired privacy features.

Profile Attributes. In the majority of social networking websites, each user account is associated with a set of profile attributes. A subset of such profile features is configured by the account holder, often when the account is created, and is intended to represent who the user is in the network. Examples of such profile attributes include the username, the profile image, and the location information of the user. Another subset of attributes, which are also specified by the user, is related to the settings of their account. For instance, they can specify their preferred interface language or whether they want their shared information to be geo-tagged. Finally, a set of contextual attributes is calculated by the social networking website itself based on user activities. For instance, the time of the account creation, the number of items published by each user, and the number of social contacts. A limited number of studies have focused on the potential relations of profile attributes and privacy behaviour and attitudes.

In the context of Facebook, a supervised learning algorithm is built on a large set of features to recommend privacy settings [25]. These features include metadata elements regarding a shared item as well as users' demographic and profile features. It is worth noting

that the algorithm is developed to recommend privacy settings for a particular shared item, such as individual Facebook posts, as opposed to predicting general privacy preferences. By focusing on Twitter, Dong et al. [6] proposed a privacy prediction model that takes into account social media behavioral analogs to psychological variables that are known to affect user disclosure behavior. Some of the identified analogs are based the profile features of the users. For instance, user trustworthiness is calculated based on the ratio of their followers to the total number of their social contacts.

Social Context. Relative to the number of studies focused on profile attributes, a large set of approaches have focused on the social context of the focal user to predict privacy features. These studies can be categorized into two primary groups. The first set of works mainly focuses on privacy in terms of information visibility to different groups of social contacts, often referred to as social circles. Hence, they propose approaches to assist users in creating and maintaining such social circles and to infer preferred privacy settings for the created circles of contacts. The second group of works has primarily focused on homophily and proposed collaborative filtering methods to identify the desired privacy features.

Various algorithms have been proposed to create circles of social contacts that are closely related to each other and have many links within themselves, while having fewer links with those who are not in the circle [5, 2]. In [28], after the clusters of contacts are formed, the user is asked to label a number of randomly selected contacts from each cluster in terms of his/her willingness to share a specific item with them. A classifier is trained on the profile attributes as well as the network attributes of the labeled contacts to predict the privacy preferences of the user for unlabeled contacts. In a somewhat relevant work, Fang and LeFevre [7] built a privacy wizard that iteratively asks the user to label carefully-selected informative contacts in terms of his/her willingness to share a specific piece of profile information with them. To automatically label other contacts, these labeled information is utilized in a classifier in which contacts are represented in terms of their community structure and profile attributes.

A closely related direction of research to this study has adapted techniques from the area of collaborative filtering to predict one's privacy preferences. These studies are motivated by homophily, which refers to the tendency of people to associate with similar individuals [23]. In addition, user information sharing behaviour has shown to be extensively influenced by an inner circle of close friends [4]. Therefore, one set of approaches is focused on predicting privacy preferences based on the preferences of social contacts of the focal user (e.g., friends on Facebook or followers on Twitter). Similarity, people with similar backgrounds tend to have similar privacy concerns [31]. As an alternative to the use of social contacts, a set of researchers has developed techniques to detect privacy preferences

based on the known preferences of users with similar backgrounds and characteristics with the focal user.

For instance, Squicciarini et al. [32, 30] provide an algorithm to form social circles based on users' characteristics such as their gender, hobbies, and occupation. When a new object (i.e., contacts or data items) is uploaded, the system first seeks the social circles that are most likely to deal with the object in a similar way as the user. Then the privacy policies used by the selected circle is the basis for predicting the privacy policy for the newly added object. In [29], the focal user first specifies whether he/she is willing to share a specific data item with a selected contact. Then an iterative semi-supervised approach is followed to label the other contacts of the user, where labels are propagated from labeled instances to unlabeled instances in the social graph. This propagation is guided by the user similarity metric that is based on profile information of contacts, their networks metrics, as well as the community structure.

In [10], users are first characterized according to their privacy preference as either privacy fundamentalist, privacy pragmatist, or privacy unconcerned. Users' privacy decisions and settings regarding their photo albums are considered as an indication of their privacy preference. In particular, users are assigned to these three categories based on the number of their public, customized, and private photo albums. Then K-nearest neighbour algorithm is used to determine which privacy categorization the focal user belongs to. Based on the features of the assigned category, the system then recommends privacy settings. In the context of a location-based social network, Toch et al. [36] provide users with recommended privacy policies that similar users have previously selected.

Published Content. A frequent user activity on social networks is to publish and share content such as text messages, images, and videos. These shared data types can be used to draw inferences about users' personality and preferences. In the context of textual context and privacy, Gill et al. [11] provide a set of privacy-related categories of words that are relevant in the semantic analysis of the privacy domain. LIWC[2] also contains a large number of semantic categories of words with possible relevance to privacy features. Caliskan-Islam et al. [4] used the privacy dictionary, along with a variety of methods and tools including topic modeling, named entity recognition, and sentiment analysis to automatically deem if a tweet contains private information. Then users are given privacy scores based on the amount of private information they published in their Twitter timelines. The timelines of the labeled users are then used in a supervised technique to assign privacy scores to unlabeled users.

The prediction model proposed in [25] also follows a supervised approach to recom-

[2]Language Inquiry and Word Count:`http://liwc.wpengine.com/`

mend privacy settings for a given post on Facebook. The sentiment score of the post is included in their feature set, along with some topical features that indicate whether the post is related to a set of pre-specified topics (e.g., family, work, travel, etc.) or not. In addition, the entire bag-of-word representation of the content is taken into account. Given an unstructured linguistic content published by a user, [33] first detects sensitive information such as phone number, address, and location from the text. Then the model proposed in [19] is adopted to quantify the potential privacy risk of the user in the network.

In the context of image sharing, [31] uses the previous images by users and their corresponding privacy policies to assign a privacy policy to a new image. Image clustering and policy association mining are used for privacy generation. However, if the user is new or there have been significant changes to the user's privacy trends, users' social context is used to predict the policy. In order to detect images with private content, Zerr et al. [42] used a variety of visual features, such as the occurrence of faces, in a supervised manner. They also utilized the textual metadata of images and found correlations of the topic and private image content. For instance, topic used to describe personal concepts, emotions and sentiment, and human body were shown to be mostly used for private images. On the other hand, topics related to nature, architecture, and inanimate objects have been mostly found in non-private images.

According to our earlier analysis of the literature [13], the majority of the studies are focused on predicting privacy settings that are specific to a particular social networking website, while less attention has been paid to general user modeling and characterization. Even though our platform of focus is Twitter, we aim to characterize general privacy preferences of publicly available user profiles. In addition, unlike many of the earlier attempts that are focused on a single or a few types of data, we incorporate a variety of data types (e.g., profile features, social context, and published content). Finally, many of the reviewed studies have utilized supervised methods to classify and predict privacy attributes. However, such methods require labeled input and may not seem feasible in the context of social media, where the labeled information normally constitutes a very small portion of the available data. Instead, we propose an unsupervised collaborative filtering approach that is motivated by the locality of preferences in social media.

4.3 Dataset

4.3.1 Data Collection

To collect and build a social network from Twitter, we first selected a random user by generating a random Twitter ID. We ensured that this initial user is publicly available because the social contacts of *protected* accounts are inaccessible through the Twitter API, which makes it impossible to expand the network from a *protected* user node. After this user was selected, we iteratively built a network of users in a Breadth First Search (BFS) manner. Given that our approach exploits preference locality, we focused only on reciprocated relations instead of the asymmetric follow or friend relation. Reciprocated relations are expected to indicate a stronger relationship between the two users, and they distinguish the social network section of the Twitter-sphere from its information network [24, 38]. As we are only focused on this mutual contacts, from now on, whenever we use the word *contact*, we refer to the social contacts of the focal user with reciprocated relations.

Before adding each *public* user to the network, we retrieve and calculate a set of metadata about the user. We first count the percentage of *protected* contacts to all of the contacts of the focal user. For instance, if a user has 100 social contacts among which 20 have protected their accounts, the user will be assigned the value of 20%. This percentage, called the *privacy ratio*, is a primary metric for our further analysis. In addition, we collect Twitter profile attributes (e.g., location and tweet count) and the latest 500 tweets published by the user as node metadata. Once the augmented user node is added to the network, we check if the new node has a reciprocated relationship with any of the existing nodes and add the corresponding edges. This process is repeated with a new *public* user pulled from the BFS queue. Figure 4.1 shows an overview of our data collection process.

It should be noted that users with less than 10 tweets or less than 30 followers/friends are considered inactive and thus are not included in the data collection process. In addition,

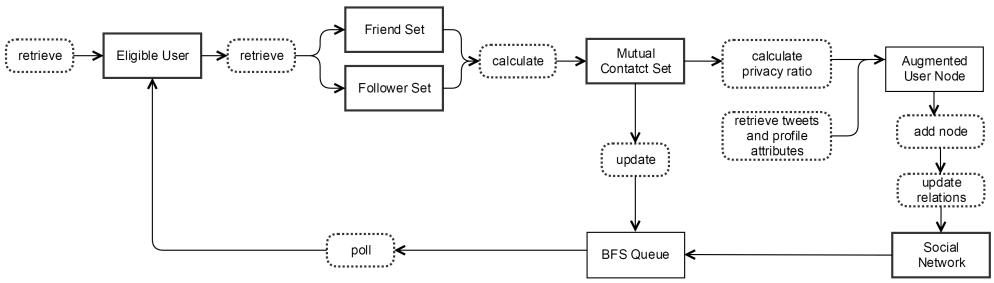
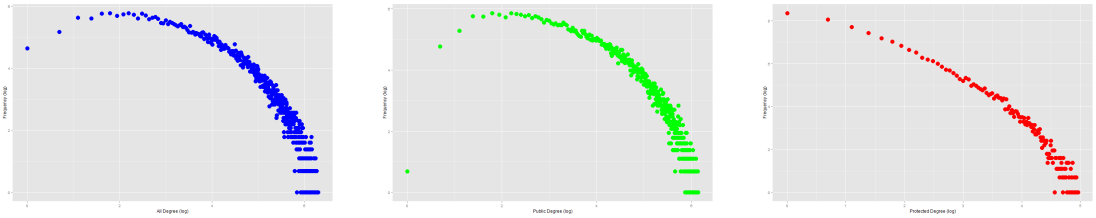


Figure 4.1: An overview of the data retrieval procedure.



(a) (b) (c)

Figure 4.2: The degree distribution across all users in the network on a log-log scale. The distribution is shown for all of the social contacts (a), the *public* contacts (b), as well as the *protected* ones (c).

verified users and users with more that 1K followers/friends are excluded since they often represent brands and celebrities and are not from the general public. By following this approach, we collected the total of 23,320 *public* user nodes and 6,489,419 tweets published by these users.

4.3.2 Descriptive Analysis

In this dataset, each Twitter account is mutually connected to an average of 86 contacts. Among these neighbours, an average of 76 are *public* and 10 are *protected*. In addition, each user is associated with an average of 339 tweets. We can obtain some insight into the network structure by examining the degree distributions. Figure 4.2 (a) shows the degree distribution for all of the mutual contacts across all users on a log-log scale. A heavy tail can be seen in the graph, resembling a power-law distribution. Similarly, the degree distribution for *public* contacts shown in Figure 4.2 (b) exhibits a heavy-tail. The same applies to the degree distribution for *protected* contacts, though to a larger extent compared to the other two (see Figure 4.2 (c)).

We also attempted to fit all of the three degree distributions to a power law distribution: $P(x) \sim x^{-\alpha}$. Throughout the fitting, we obtained the α values of 2.49, 2.98, and 1.79 for all, *public*, and *protected* accounts, respectively. For all of the three distributions, the Kolmogorov-Smirnov (KS) test indicates that the distribution is not refused ($P > 0.05$), and the power law can indeed be a good fit. Power law distribution is commonly observed in the context of social networks, though it is interesting to observe the same trend even after filtering the users with more than 1K friends/followers (as described in Section 4.3.1).

Figure 4.3 shows the relationship of *public* and *protected* contacts across all users. Not surprisingly, these two metrics are positively correlated, indicating that as the number of *public* contacts increases, so does the number of *protected* contacts. However, as the linear regression line represents, the number of *public* contacts grows at a larger scale compared to

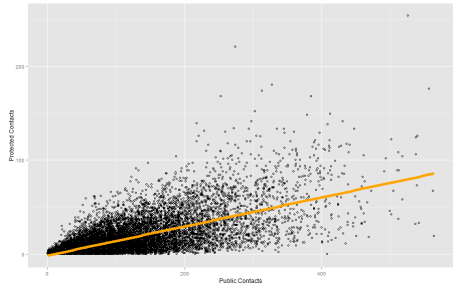


Figure 4.3: Correlation of the number of *public* contacts and the number of *protected* contacts.

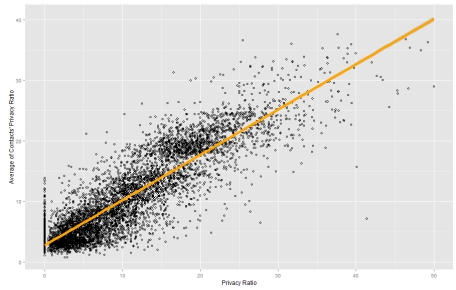


Figure 4.4: Correlation of users' privacy ratio and the average privacy ratio for all of their contacts.

the *protected* ones. Finally, as the first step to ensure that privacy preferences are localized in the context of privacy, we calculated the correlation between the privacy ratio of each node and the average privacy ratio of the contacts. The analysis of users who have at least 10 mutual contacts in the network (about 7000 users) showed a strong positive correlation between the two variables (Spearman $\rho = +0.89$). This correlation is also apparent in Figure 4.4, wherein the orange line is the linear regression line fitted to the data. This result indicates that users' privacy behaviour is either influenced by their close social contacts or individuals with similar privacy behaviour tend to cluster together in social networks. In either case, this finding implies the great potential of collaborative filtering approaches for privacy preference prediction.

4.4 Amazon Mechanical Turk Experiment

Even though the correlation analysis signals the existence of preference locality for the privacy attribute, we conducted another experiment based on human annotations to further validate this finding. We published 12K tweets to AMT and asked AMT workers to judge the privacy preference of the tweet author based on the tweet. These 12K tweets were

selected according to to the following procedure.

Our collected user set was first filtered only to include those users who tweet in English. Then, 1200 users with the lowest privacy ratio and 1200 users with the highest privacy ratio are extracted from the set. Next, we selected the five latest tweets published by these users that are not retweets or replies. We also ensured that these tweets include some text and are not generated automatically (e.g., is not labeled with *[autotweet]*). This process resulted in the selection of 12K tweets published by 2400 different users with extremely low or extremely high privacy ratios. Each tweet is then published to AMT to be labeled by a human annotator according to his/her judgment of the privacy preference of the tweet author. To do so, given a tweet, workers could choose one of the three given options:

- Privacy Unconcerned: He/she is not concerned if his/her data is used by companies
- Neutral/Objective: Neutral or objective tweets that reveal no information about the users' privacy preferences
- Privacy Concerned: He/she is not willing to share his/her data with companies

Each tweet is labeled with three different workers. Ideally, the majority label should be the subject of further analysis. However, the results are extremely unbalanced. Only about 1% of the tweets are labeled as a tweet posted by a privacy concerned user (i.e., the third option), while each of the other categories is selected more than 50% of the times. This problem might be due to the lack of expertise of the AMT workers on the subject, the features of the selected tweets, or the design of the annotation task. Further in-depth experiments can shed light on this matter. However, addressing this issue is out of the scope of this research and will be included in our future research activities. Here, to facilitate comparison and analysis, we employed an alternative approach.

A tweet is considered to be categorized as a tweet published by a privacy concerned user if it is annotated as such by at least one worker. Then, we counted the number of tweets that are categorized under this label (i.e., privacy concerned) for each user in the dataset. The comparison of the number of tweets of this kind among the two user groups shows a statistically significant difference (t-test P-value < 0.05). In particular, tweets published by the users with a high privacy ratio are associated with this category more often compared to the ones with a low privacy ratio. While we acknowledge the limitation of the analysis concerning the use of one worker's annotation, this finding implies that users located within private neighborhoods are judged to be privacy-concerned, providing another evidence for the locality of preferences in the context of privacy.

4.5 User Analysis

We analyzed different characteristics of *public* users and examined their potential relations to the privacy ratio metric. These characteristics are elicited from users' profile attributes and their published tweets, both of which are directly available from *public* user profiles. The related experiments and findings are discussed in Section 4.5.1. In addition, we transformed the network of users to a bipartite network of features and users, wherein the features are given privacy ratios. Our transformation procedure and the results are discussed in Section 4.5.2.

4.5.1 Observable Attributes

Profile Attributes

A subset of the available profile attributes is deemed relevant for our purpose and is selected to be the subject of further analysis. This list is shown below, along with a brief description of the attributes. Please note that the descriptions are adapted from Twitter API specification[3].

- Description: A piece of text users provide to describe their account.
- URL: A URL provided by the user in association with their profile.
- Location: The user-defined location for this account's profile.
- Geo-enabled: When true, indicates that the user has enabled the possibility of geotagging their Tweets.
- Default Image: When true, indicates that the user has not uploaded their picture and a default avatar is used instead.
- Default Profile: When true, indicates that the user has not altered the theme or background of their user profile.
- Favorite Count: The number of tweets this user has favorited in the account's lifetime
- Tweet Count: The number of tweets issued by the user.
- Follower Count: The number of followers this account currently has.
- Friend Count: The number of users this account is following.

[3] https://dev.twitter.com/overview/api/users

We employed a simplified binary attribute to study user descriptions, URLs, and locations. The value of this attribute is "1" if the corresponding piece of information is provided by the user and is "0" otherwise. The geo-enabled attribute, default profile, and default image are all binary attributes and are studied as binary variables. Similarly, the numeric attributes of the favorite count, tweet count, follower/friend count are analyzed as is. To study how profile attribute of users with different privacy ratios are configured, we calculated correlations between the profile attributes and the privacy ratio across all the users.

Twitter profile attributes are among a few data points that are visible to and accessible by the public and the Twitter API regardless of the user privacy settings. Therefore, we retrieved profile attributes of *protected* and *public* accounts and directly compared them. Such a supplementary experiment allows us to understand whether profile attributes of Twitter users with varying privacy settings are configured differently. In addition, by comparing these findings with the analysis of *public* users with different privacy ratios, we can investigate if *protected* accounts behave similarly to the *public* accounts located in more private neighbourhoods (regarding their profile attribute configuration). Finding similarities across the two sets can then contribute towards confirming the that the privacy attributes are localized.

To reliably compare the profile attributes of users with *protected* and *public* settings, we need a set of users in which the privacy attitude-behaviour dichotomy is minimized. Our earlier analysis of various follower sets associated with famous Twitter accounts [14] indicates that the followers of "CNN Breaking News" can be a good candidate set. This conclusion is made because the percentage of the *protected* followers to the total number of followers for "CNN Breaking News" has been shown to be considerably higher than the other follower sets and the average percentage in Twitter [14]. Therefore, we collected profile attributes of 1M *public* and 1M *protected* accounts from the CNN follower set and compared their profile features. In the collection process, we ensured that inactive accounts, brands, and celebrities are excluded. The details of the data collection process can be found in [14].

Table 4.1 presents the analysis of the selected binary features. The correlations between these binary attributes and the privacy ratio of users seem to be small for almost all the features. However, the differences between the *protected* and *public* accounts are statically significant for all of the variables and even practically significant for the three of them. Interestingly, with one exception (*Has Description*) the results show a similar behaviour for users with a high privacy ratio and those who have chosen to have *protected* accounts. In particular, the attribute of *geo-enabled* has a relatively higher positive correlation with

Binary Attributes	Privacy Ratio (Spearman ρ)	Privacy Setting (Chi-Square P-value)
Has Description	-0.07	+++
Has URL	-0.17	- - -
*Has Location	-0.06	+++
*Is Geo-enabled	+0.11	+++
*Is Default Profile	-0.07	- - -
Is Default Image	-0.03	- - -

Table 4.1: Analysis of binary profile attributes and privacy-related features. ++ for $P < .05$ and +++ for $P < .005$ when the values are greater for *protected* accounts, whereas - - and - - - are used when the values are smaller for *protected* accounts. The variables for which at least a small effect size (Cramer's V > 0.1) is observed are marked with an asterisk.

the privacy ratio and is more commonly used by users with *protected* accounts (both statistically and practically significant). In addition, *protected* accounts provide external URLs less often compared to the *public* ones. Similarly, the negative correlation of *Has URL* and the privacy ratio indicates that users located within more private neighbours tend not to provide URL information compared to their counterparts.

Table 4.2 provides the experiment findings for the numeric features. A positive correlation is observed across all the variables; however, the correlation coefficient is relatively higher for *tweet count* and *favorite count*. This positive correlation indicates that users with a high privacy ratio tend to tweet and favorite tweets more often. The results for the direct analysis of *protected* and *public* accounts are in line with the correlation coefficients since *protected* users have a significantly larger tweet, favorite, friend, and follower count. The difference for the *tweet count* feature is also practically significant.

Our direct analysis of *protected* and *public* accounts shows that *protected* accounts voluntarily reveal more information about themselves (e.g., geo-tags) and participate more actively in the network (e.g., tweet count). A possible speculation is that since users with *protected* accounts are aware that their data are private, they feel secure in this environment. On the other hand, users who are consciously following the *public* setting are utilizing a different strategy, such as self-censoring, to protect their privacy. Among the profile attributes, the existence of an external URLs shows a different pattern. *Public* accounts and users within public neighbourhoods seem to provide URLs more often. Despite our effort to exclude brands and celebrities, this difference can be attributed to professional Twitter accounts. For instance, artists who are on Twitter to promote their art often provide an external URL to their portfolio. Interestingly, our correlation findings reveal that publicly available users with a large percentage of *protected* contacts behave similarly to the

Binary Attributes	Privacy Ratio (Spearman ρ)	Privacy Setting (T-test P-value)
Favorite Count	+0.27	+++
*Tweet Count	+0.30	+++
Follower Count	+0.15	+++
Friend Count	+0.02	+++

Table 4.2: Analysis of numeric profile attributes and privacy-related features. ++ for $P <$ .05 and +++ for $P <$ .005 when the values are greater for *protected* accounts, whereas - - and - - - are used when the values are smaller for *protected* accounts. The variables for which at least a small effect size (Cohen's d > 0.2) is observed are marked with an asterisk.

protected contacts. The implications of this finding are two-folds. First, it indicates the existence of locality for privacy preferences. Second, this result implies that such users feel secure in the environment in terms of sharing their information, yet they are more privacy concerned than the other public accounts. It is thus expected that they are more likely to feel invaded by targeted advertising and marketing messages.

Language Use of the Content

Natural language has been shown to be a reflection and a mediator of internal states [26]. Our words can reveal personality, emotional states and feelings, attention patterns, thought, and social situations [26, 11]. Therefore, a variety of automated content analysis techniques has been developed to measure such psychometric metrics from natural language. These methods range from the use of predefined dictionaries and taxonomies such as LIWC to complex computational algorithms that often utilize data mining and machine learning methods.

LIWC dictionaries are capable of providing a broad range of social and psychological insights from the language. Hence, we used LIWC to analyze the language of tweets and to examine the links between a set of linguistic indicators and users' privacy behaviour. LIWC has a processing component that examines a text file word by word. Each word is then compared against the built-in dictionaries. Given that LIWC dictionaries are structured in a hierarchical format, the processing component then determines which LIWC categories or sub-categories the word belongs to. Once all the words are processed, LIWC outputs the percentage of words that belongs to a particular category to the total number of words in the text. In addition, a set of LIWC variables are measured independent of the dictionaries and are referred to as summary variables. These variables include four non-transparent language variables (analytical thinking, clout, authenticity, and emotional tone) and general

descriptive features of the text (words per sentence and percent of words that are longer than six letters). The definition and examples of each of these categories can be found at the LIWC website[4].

Tweet sets published by each user are first cleaned and pre-processed. For instance, elisions are handled (e.g., I'm –> I am), URLs and Twitter mentions are replaced with specific tokens, and emoticons are replaced by their corresponding meaning (e.g., ":)" –> smile). Then all of the collected tweets published by a user is treated as a single document and is given to LIWC for analysis. The percentages calculated by LIWC are then studied in in terms of their correlations with the privacy ratio. Table 4.3 summarizes the correlation results for LIWC categories and summary variables, which are ranked based on their correlation strength. The Table only includes those variables with their correlation coefficient beyond a certain threshold ($\rho > 0.20$ and $P < 0.005$).

The majority of the LIWC features that are positively correlated with the privacy ratio can be associated with private content according to societal consensus. Examples of these features include the use of swear words, expression of anger and anxiety, and sexual topics. In addition, a positive correlation is observed for the use of "I" and the privacy ratio. Personal pronouns mainly appear in narratives and tweets that describe personal events, feelings, opinions, etc. Similarly, the use of past tense is often observed in more private neighbourhoods.

In LIWC, the analytical thinking feature captures the degree to which a piece of text represents formal, logical, and hierarchical thinking. Analytical thinking is negatively correlated with the privacy ratio. This finding may be attributed to the professional accounts (e.g., belonging to artists, politicians, athletes, etc.) that are often located within public neighbourhoods and may normally use a formal and a logical language.

The following are two example tweets that are from the timelines scored high on analytical thinking:

- The staff on the Woodman ready to go our last cruise of the season.
- Up to 40,000 cardiac arrests occur each year in Canada. Without treatment, most of these cardiac arrests will result in death. Learn CPR.

A relevant category among the LIWC outputs is called authenticity, which captures the degree to which the language is more honest, personal, and disclosing. Even though the correlation is below our threshold and thus is not included in the Table ($\rho = +0.22$), the authenticity of the language is shown to be positively correlated with the privacy ratio.

[4] http://www.liwc.net/descriptiontable1.php

LIWC Feature	Privacy Ratio (Spearman ρ)
Swear Words	+0.40
Anger	+0.35
Negative Emotions	+0.34
Body	+0.32
Negations	+0.31
Adverbs	+0.30
Sexual	+0.29
Sad	+0.27
Analytical Thinking	-0.26
FocusPast	+0.26
Interrogative	+0.26
Feel	+0.26
I	+0.26
Pronoun	+0.25
Anxiety	+0.25

Table 4.3: Correlation analysis of the LIWC categories and the privacy ratios.

Again, this finding shows that people located within private neighbourhoods are likely to be privacy-concerned, but they are probably privacy-unaware and thus publish sensitive information about themselves. Below are two example tweets from the timelines that are scored high on authenticity:

- It's going to be one of those days where everything that can go wrong, will go wrong.
- First night with the new roomie. Watching The A-Team! #bradleycooper

As discussed earlier, tweets published by *protected* accounts are inaccessible, making it impossible to compare *protected* content with their *public* counterparts. However, there exists an accessible component that can represent linguistic characteristics of *protected* accounts: profile descriptions. When configuring their profile attributes, Twitter users can provide up to 160 characters in the description field. In our set of the CNN followers, there are almost 500K of the *protected* accounts and roughly 500K of the *public* accounts that have descriptions. We used LIWC to analyze the language categories in these descriptions. Since profile descriptions are often very short (commonly between 8-10 words), the percentages provided by LIWC are very small for the majority of the categories. Therefore, we only focused on the higher-level categories that are at the top of the LIWC hierarchy as well as the summary variables. Table 4.4 shows the result summary.

LIWC Feature	Privacy Ratio (Spearman ρ)	Privacy Setting (T-test P-value)
Analytical Thinking	-0.26	- - -
Authentic	+0.22	- - -
Clout	-0.15	+++
Function Words	+0.23	+++
Affect Words	+0.13	+++
Social Processes	+0.10	+++
Cognitive Processes	+0.23	+++
Drivers and Needs	-0.03	- - -

Table 4.4: Analysis of the lingustic indicators and the privacy-related features.

The two underlying datasets represent two different sets of users that are collected in different manners from Twitter. In addition, the language content of tweets and descriptions are provided for different purposes. However, we still see the same behaviour from *protected* accounts and the *public* accounts that are connected to a large percentage of *protected* neighbours in terms of their language use (see Tables 4.3 and 4.4). For instance, authenticity, the use of function words, affect words, social processes, and cognitive processes are positively correlated with the privacy ratio. Likewise, they are more observed in profile descriptions of the *protected* accounts. On the other hand, analytical thinking, clout, drivers and needs are negatively correlated with the privacy ratio and are observed less often in *protected* accounts. Again, such similarities signal the presence of locality for users' privacy behaviour. Therefore, features that are specific to *public* accounts that are connected a large number of private contacts can be considered the features that characterize privacy-concerned users.

Tweet Sentiment

LIWC captures the percentage of words that belong to different sentiment-related categories (e.g., positive and negative words, anger, and anxiety) in user tweet sets. In addition to these LIWC categories, we took advantage of a lexical resource, called SentiWordNet [3], to analyze tweet sentiments and privacy features from a different perspective. In SentiWordNet, each word is given three sentiment scores: positivity, negativity, and objectivity. We first cleaned and tokenized each tweet. The tokens are then POS tagged and stemmed. The resulting token-POS tag pair is then matched against SentiWordNet. Finally, the scores retrieved from SentiWordNet are then aggregated for all the tokens in the tweet to generate an overall tweet sentiment score. It should be noted that whenever a negation

is observed, the inverted score of the following token is taken into account. Once tweet sentiment scores were determined, we calculated the ratio of positive and negative tweets to the total number of tweets. Based on the analysis, the ratio of positive tweets has a small negative correlation with the privacy ratio ($\rho = -0.12$), while the ratio of negative tweets is positively correlated with the privacy ratio ($\rho = +0.26$). The number of negations observed in the tweets is also positively correlated with the privacy ratio ($\rho = 0.28$). Due to the inaccessibility of tweets published by *protected* account, the direct analysis of *protected* and *public* accounts in terms of the ratio of tweets with different sentiment labels is not possible.

Communication Behaviour

We employed a set of simple variables to characterize user's communication behaviour from their timelines. In Section 4.5.1, we observed a positive correlation between users' tweet count and their privacy ratios. In addition to the frequency of tweeting, we examined the average of tweet length and found a negative correlation with the privacy ratio ($\rho = -0.27$). Published tweets can either be retweets from other accounts or new tweets coming from the account of focus. Note that the retweets are excluded from the tweet length analysis. The correlation between the ratio of the retweets to the total number of tweets and the privacy ratio is very small ($\rho = +0.04$), and the ratio of the new tweets is also positively correlated with the privacy ratio ($\rho = +0.15$). In addition, users tweet to interact with each other and engage in conversations. The ratio of this conversational tweets also shows a positive correlation with the privacy ratio ($\rho = +0.22$). Another potentially interesting variable to investigate is the use of URLs when tweeting. We expect the professional and non-personal accounts to publish news and events more frequently, which are often linked with URLs. As expected, the ratio of the tweets with URLs is found to be negatively correlated with the privacy ratio ($\rho = -0.22$). We found negligible correlations between the hashtag usage patterns and the privacy ratio. Similar to tweet sentiment, direct analysis of *protected* and *public* accounts is impossible due to the privacy restrictions associated with *protected* accounts.

4.5.2 Latent Attributes

Method

To discover a set of latent attributes that are of interest to privacy-concerned users, we transformed each user node in the network into a set of attributes. For each attribute node, we then calculated the ratio of *protected* contacts to the total number of contacts. The resulting

network allows us to understand which features attract privacy-concerned users and which ones are more observed in public neighbourhoods. Figure 4.5 shows this transformation procedure for a sample network in which *public* nodes are encoded by blue and the *protected* ones are shown in red. Suppose that we have three users in the network: U_1, U_2, and U_3. The nodes with dotted borders are the social contacts of the users that are not yet added to the network but are counted in the metadata calculation process and added to the BFS queue (see Section 4.3.1). As Figure 4.5 (a) shows, U_1 is linked with three social contacts among which two are *public* and one is *protected*; therefore, it will be given the privacy ratio of $\theta_{u1} = 33\%$. Similarly, U_2 and U_3 are given the values of $\theta_{u2} = 50\%$ and $\theta_{u3} = 25\%$, respectively.

Now suppose that U_1 can be characterized with three features such that $U_1 = \{f_1, f_2, f_4\}$. These three features are then associated with all of the three social contacts of U_1. Then if $U_2 = \{f_2\}$, f_2 will also be associated with the neighbours of U_2. Finally, suppose that we have $U_3 = \{f_1, f_3, f_4\}$ in our example network. Figure 4.5 (b) shows the process of associating features with the contacts. The resulting network will then be a bipartite network of contacts and features, wherein privacy ratios can be calculated for features (see Figure 4.5 (c)). For instance, f_1 is a feature that characterizes users U_1 and U_3 and is not associated with U_2. Therefore, in Figures 4.5 (b) and 4.5 (c), it is linked with the neighbours of U_1 and U_3. Hence, f_1 is linked to the total of five *public* users and two *protected* ones, resulting in the privacy ratio of $\theta_{f1} = 40\%$.

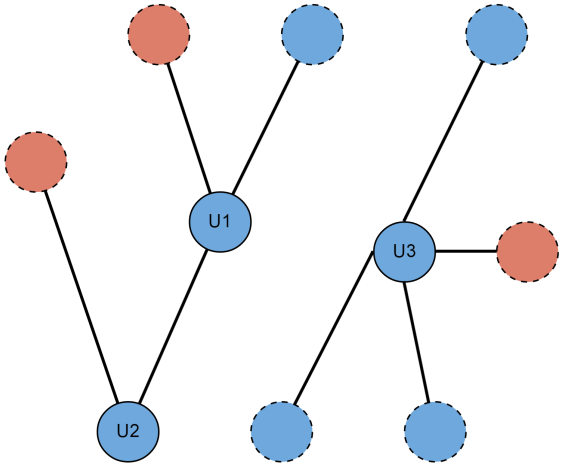
A variety of features can be extracted from Twitter timelines to describe the accounts. We conducted experiments with four primary features: 1) tweet unigrams, 2) hashtags, 3) Twitter accounts that users retweet from, and 4) topics. Each of these feature sets can capture and reveal a particular aspect of user behaviour and interest. Unigrams are extracted using Apache Lucene[5], which is a well-known text search engine library developed in Java. Hashtags and retweet sources are simply extracted using regular expressions. Finally, Latent Dirichlet Allocation (LDA) implementation in Mallet[6] is used to extract and analyze topics discussed in tweets.

Results

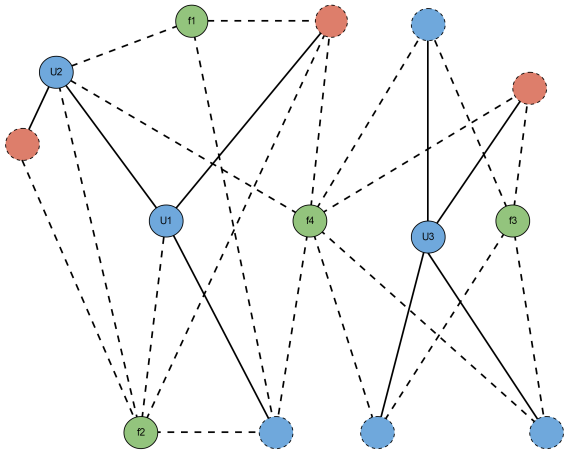
Table 4.5 provides an overview of the features' statistics in our dataset. To gain reliable privacy ratios for the features, they need to be used by a considerable number of users. Therefore, we filtered the lists only to include those features that are used by at least 100 users in the underlying set. This filtering resulted in 1228 hashtags, 906 retweet sources,

[5] `https://lucene.apache.org/core/`

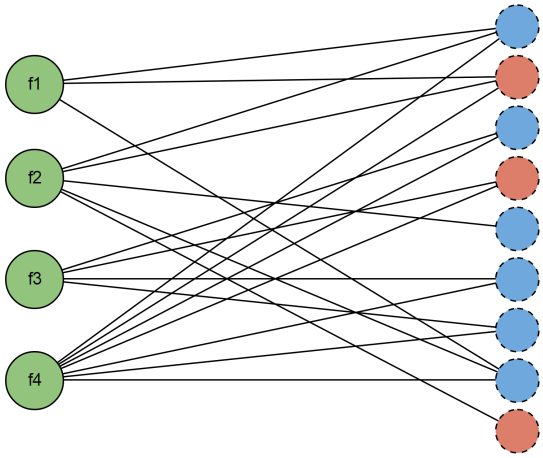
[6] `http://mallet.cs.umass.edu/topics.php`



(a)



(b)



(c)

Figure 4.5: An overview of the network transformation process.

Feature	Unique Instances	Filtered Instances	Privacy Ratio (Average)	Privacy Ratio (Variance)
Unigrams	1,094,642	23470	9.15	6.10
Hashtags	514,186	1228	9.43	10.29
Retweet Source	655,384	906	11.69	16.74
Topics	150 (pre-specified)	32	9.09	1.30

Table 4.5: Descriptive statistics of the features that represent user timelines.

more than 23K unigrams, and 32 topics. The last two columns of the table represent the mean and variance of the privacy ratios for these filtered lists.

Table 4.6 shows a set of example features that are located at the two ends of the privacy ratio spectrum. As can be seen, certain features in the list are specific to the dataset of focus. For instance, *@CityOfHamilton* and *@Kathleen_Wayne* show that a considerable number of users sampled in our data collection process happen to be located in Canada. This limitation, however, is a general problem that data-driven approaches suffer from and would diminish as more data is collected. In addition, employing ontologies to map these features to a higher-level space may address this issue and will be considered in our future research plans. Nevertheless, our focus here is to gain insight into the general patterns of these features and their privacy ratios. A glance at the example table shows that privacy-sensitive features are given high privacy ratios, while the examples at the other end can often be seen in the professional and non-personal content. These results are consistent with our earlier findings from the analysis of profile features and LIWC attributes. Given that this approach is only focused on user relations and their privacy settings, our analysis of unigrams and hashtags can be conducted independent of the language of the tweets. For instance, *usluge*, which means *service* in Bosnian, is a unigram with a very low privacy score ($\theta = 2.59$) in our list.

As explained earlier, we employed LDA to generate topics from user timelines. To build a topic model, LDA requires three input parameters that need to be specified: the number of topics to be generated (often denoted by k), the value of α, and the value of β. α represents document-topic density. As such, a higher α indicates that documents consist of more topics, while a lower α means that documents contain fewer topics. β represents topic-word density. Therefore, a high value of β implies that topics are made up of most of the words in the corpus, while a low β means they consist of few words. We built topic models with different number of topics (k = 50, 100, 150, and 200). For each model, we chose the value of α=50/k and β=0.01 as suggested in [12]. The topics generated by k=150 were more sensible according to our observation; hence, this model was chosen for

Feature Type	Top Features	Privacy Ratio	Bottom Features	Privacy Ratio
Unigrams	hooka	17.60	consultation	3.43
	shittiest	16.69	workshops	3.33
	hungover	16.63	catering	2.94
Hashtags	#TakeMeBack	19.76	#Adventure	3.88
	#SoTired	16.87	#ShopLocal	3.30
	#MissYou	16.64	#Entrepreneur	1.87
Retweet Source	@ColiegeStudent	17.5	@Kathleen_Wayne	4.60
	@FemalePains	17.10	@CityOfHamilton	4.05
	@MensHumour	16.91	@DanceMoms	4.02

Table 4.6: Example features extracted from user timelines along with their corresponding privacy ratios.

Topics	Label	Privacy Ratio
photo, gay, album, love, LGBT	sexual	10.67
health, dental, care, diet, smile	health and body	10.67
great, day, happy, weekend, tonight	positive experiences	10.55
food, wine, beer, lunch, delicious	dining	7.38
home, real estate, house, tips, mortgage	real estate	7.31
stats, followers, unfollowers, checked, automatically	follower control	6.35

Table 4.7: Topics extracted from user timelines along with their corresponding privacy ratios.

our further analysis. Table 4.7 lists the top three and bottom three topics ranked according to their privacy ratios. The table shows a set of sample words that represent each topic, the topic label that we crafted based on the given words, and their privacy ratios. Again, the results are consistent with our earlier findings since private topics are given high privacy ratios, while topics that can be associated with the professional and non-personal content are given lower privacy ratios.

4.6 Discussion

Our correlation analysis of the privacy ratio across different neighbourhoods shows that privacy preferences are localized. In addition, the human annotated data collected from AMT provided further support to justify the locality privacy preferences. Finally, the analysis of *public* and *protected* accounts and their comparison with users located in different neigh-

bourhoods of various privacy settings provided additional cues to confirm the locality in this context. This conclusion is primarily made based on the similarities observed between *protected* accounts and the *public* accounts which are connected with a larger percentage of *protected* neighbours. To the best of our knowledge, the study conducted by Caliskan-Islam et al. [4] is the only study that has examined the potential relations of users' privacy features and the privacy attributes of their neighbours in online social media. Consistent with our findings, their analysis of 45 Twitter users showed that the amount of private information shared by users is positively correlated with the amount of private information shared by their neighbours.

-revision(Reza and Anabel) The finding that privacy preferences are localized may indicate that privacy features are homophilous in the context of Twitter (see Section 4.3.2). Alternatively, Twitter user preferences may be influenced by their close neighbours. Further in-depth research to reveal the potential effects of each of these two processes on privacy preferences is warranted. Earlier sociology research distinguishes two types of homophily: status homophily and value homophily [23]. Status homophily is based on informal, formal, or ascribed statuses of peoples and includes sociodemographic dimensions such as race, ethnicity, age, gender, and location. Acquired characteristics such as religion, education, and occupation are categorized as status homophily as well. Status homophily has been widely observed in the context of online social media (e.g., [41] and [17]). In addition, there is a large body of research linking sociodemographic information to privacy preferences. For example, various surveys have established a positive relation of age, education, and income linked with privacy concerns [15]. Therefore, similarities between privacy preferences of social contacts may stem from their similarities on these homophilous characteristics. Value homophily, on the other hand, includes a wide variety of internal states that are presumed to shape our orientation toward future behavior. Therefore, privacy similarities can be directly the result of value homophily, which stems from users' values and concerns for privacy. As a result of our study, the following interesting research questions can be raised: Is it the value and concern for privacy that drives and affects connections? Or is the privacy similarity just a derivative of status homophily? Or maybe both? Further in-depth investigations of privacy are required to address these questions.

Our underlying datasets are liable to suffer from the privacy dichotomy problem, and we acknowledge that this issue will introduce some errors. However, we attempted to minimize this error by choosing the CNN dataset for direct analysis of *protected* and *public* accounts (see Section 4.5.1) and found statistically significant differences between the two groups. Our network-based approach is also reliant on the current privacy settings of the contacts, thus is prone to errors caused by the disparity between user's privacy attitudes

and settings. However, by aggregating multiple privacy labels from all the contacts, we hope to diminish the effects of this issue and capture overall privacy features effectively. In addition, our correlation studies are based on relative privacy ratios. Therefore, by assuming that the privacy dichotomy problem is homogeneously distributed across the network, our network-based approach should be able to capture general privacy preferences across different neighbourhoods. The differences found between the attributes of users in more private neighbourhoods can be an indication of the success of the approach. As well, despite the differences in the two user sets, sensible results obtained from the comparison of the *protected* accounts in the CNN data and users with larger privacy ratios can reaffirm our claim.

One of the main limitations of this study is the examination of privacy preference as a binary variable. However, prior studies on privacy have shown that privacy preferences and decision are far more complex. In fact, users may employ a variety of privacy protection strategies such as self-censoring or selective sharing in social networking websites [39]. Studying such privacy protection strategies and their relations to users' characteristics in the network is warranted. In addition, our study is only focused on Twitter, while the design of the social networking website is known to shape and influence user behavior [21]. Studying other platforms with larger sets of users can give insight into the generalizability of the approach and is included in our future research plans.

4.7 Conclusion

Despite the value of personalization and customization, facilitating such services heavily relies on the collection of personal data, thus raising serious privacy concerns that need to be addressed. An appealing approach is to characterize and predict one's privacy preferences based on their social footprints, allowing companies to make informed decisions whether to discard or use one's publicly available data for business intelligence purposes. Our study confirms that one's neighbourhood context can be used to detect privacy preferences. Then neighbourhood information is used to characterized the attributes of privacy-unaware people. In particular, we found privacy-unaware users to publish more personal and private information compared to those who intentionally follow the public setting.

Our current approach examines the existence of a relation between the two users and treats all of the social contacts the same. In our future studies, we aim to take into account the strength of these relations. For instance, users' degree of collaboration can be used to assign edge weights. Alternatively, similarity metrics can be employed to characterize the strength of these social relations. Additional human annotation studies will also be taken

into account. Analysis of other social media platforms such as Facebook and Instagram is included in our future research activities.

Bibliography

[1] Alessandro Acquisti and Ralph Gross. Imagined communities: Awareness, information sharing, and privacy on the Facebook. In *Proceedings of the Conference on Privacy Enhancing Technologies*, pages 36–58, 2006.

[2] Fabeah Adu-Oppong, Casey Gardine, Apu Kapadia, and Patrick Tsang. Social circles: Tracking privacy in social networks. In *Proceedings of the Symposium on Usable Privacy and Security*, 2008.

[3] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204, 2010.

[4] Aylin Caliskan Islam, Jonathan Walsh, and Rachel Greenstadt. Privacy detective: Detecting private information and collective privacy behavior in a large social network. In *Proceedings of the Workshop on Privacy in the Electronic Society*, pages 35–46, 2014.

[5] George Danezis. Inferring privacy policies for social networking services. In *Proceedings of the ACM Workshop on Security and Artificial Intelligence*, pages 5–10, 2009.

[6] Cailing Dong, Hongxia Jin, and Bart Knijnenburg. Predicting privacy behavior on online social networks. In *Proceedings of the AAAI Conference on Web and Social Media*, 2015.

[7] Lujun Fang and Kristen LeFevre. Privacy wizards for social networking sites. In *Proceedings of the International Conference on World Wide Web*, pages 351–360, 2010.

[8] Arik Friedman, Bart P. Knijnenburg, Kris Vanhecke, Luc Martens, and Shlomo Berkovsky. *Privacy Aspects of Recommender Systems*, pages 649–688. Springer US, Boston, MA, 2015.

[9] Bo Gao and Bettina Berendt. Circles, posts and privacy in egocentric social networks: An exploratory visualization approach. In *Proceedings of the IEEE/ACM Conference on Advances in Social Networks Analysis and Mining*, pages 792–796, 2013.

[10] Kambiz Ghazinour, Stan Matwin, and Marina Sokolova. Monitoring and recommending privacy settings in social networks. In *Proceedings of the Joint EDBT/ICDT Workshops*, pages 164–168, 2013.

[11] Alastair J. Gill, Asimina Vasalou, Chrysanthi Papoutsi, and Adam N. Joinson. Privacy dictionary: A linguistic taxonomy of privacy for content analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3227–3236, 2011.

[12] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235, 2004.

[13] Taraneh Khazaei, Lu Xiao, Rober Mercer, and Atif Khan. Detecting privacy preferences from online social footprint: A literature Review. In *Proceedings of the iConference*, 2016.

[14] Taraneh Khazaei, Lu Xiao, Rober Mercer, and Atif Khan. Privacy Behaviour and Profile Configuration in Twitter. In *Proceedings of the Conference on World Wide Web - Companion Volume*, 2016.

[15] Alfred Kobsa. Privacy-enhanced personalization. *Communications of ACM*, 50(8):24–33, 2007.

[16] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a social network or a news media? In *Proceedings of the International Conference on World Wide Web*, pages 591–600, 2010.

[17] Jiwei Li, Alan Ritter, and Eduard H Hovy. Weakly supervised user profile extraction from Twitter. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 165–174, 2014.

[18] Heather Richter Lipford, Andrew Besmer, and Jason Watson. Understanding privacy settings in Facebook with an audience view. In *Proceedings of the Conference on Usability, Psychology, and Security*, pages 2:1–2:8, 2008.

[19] Kun Liu and Evimaria Terzi. A framework for computing the privacy scores of users in online social networks. *ACM Transactions on Knowledge Discovery from Data*, 5(1):6:1–6:30, 2010.

[20] Yabing Liu, Krishna P. Gummadi, Balachander Krishnamurthy, and Alan Mislove. Analyzing Facebook privacy settings: User expectations vs. reality. In *Proceedings of the ACM SIGCOMM Conference on Internet Measurement Conference*, pages 61–70, 2011.

[21] Momin Malik and Jürgen Pfeffer. Identifying platform effects in social media data. In *Proceedings of the AAAI Conference on Web and Social Media*, pages 241–249, 2016.

[22] Alessandra Mazzia, Kristen LeFevre, and Eytan Adar. The PViz comprehension tool for social network privacy settings. In *Proceedings of the Symposium on Usable Privacy and Security*, pages 13:1–13:12, 2012.

[23] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001.

[24] Seth A. Myers, Aneesh Sharma, Pankaj Gupta, and Jimmy Lin. Information network or social network? The structure of the twitter follow graph. In *Proceedings of the International Conference on World Wide Web*, pages 493–498, 2014.

[25] Kaweh Naini Djafari, IsmailSengor Altingovde, Ricardo Kawase, Eelco Herder, and Claudia Niederée. Analyzing and predicting privacy settings in the social web. In Francesco Ricci, Kalina Bontcheva, Owen Conlan, and Samus Lawless, editors, *User Modeling, Adaptation and Personalization*, volume 9146 of *Lecture Notes in Computer Science*, pages 104–117. Springer International Publishing, 2015.

[26] James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577, 2003.

[27] SAS. Finding the right balance between personalization and privacy. *SAS Report*, 2015.

[28] M. Shehab, G. Cheek, H. Touati, A.C. Squicciarini, and Pau Cheng. User centric policy management in online social networks. In *Proceedings of the IEEE Symposium on Policies for Distributed Systems and Networks*, pages 9–13, 2010.

[29] Mohamed Shehab and Hakim Touati. Semi-supervised policy recommendation for online social networks. In *Proceedings of the Conference on Advances in Social Networks Analysis and Mining*, pages 360–367, 2012.

[30] A. Squicciarini, S. Karumanchi, D. Lin, and N. DeSisto. Automatic social group organization and privacy management. In *Proceedings of the Conference on Collaborative Computing: Networking, Applications and Worksharing*, pages 89–96, 2012.

[31] A.C. Squicciarini, Dan Lin, S. Sundareswaran, and J. Wede. Privacy policy inference of user-uploaded images on content sharing sites. *IEEE Transactions on Knowledge and Data Engineering*, 27(1):193–206, 2015.

[32] Anna Squicciarini, Sushama Karumanchi, Dan Lin, and Nicole DeSisto. Identifying hidden social circles for advanced privacy configuration. *Computers & Security*, 41:40 – 51, 2014.

[33] A. Srivastava and G. Geethakumari. Measuring privacy leaks in online social networks. In *Proceedings of the Conference on Advances in Computing, Communications and Informatics*, pages 2095–2100, 2013.

[34] Katherine Strater and Heather Richter Lipford. Strategies and struggles with privacy in an online social networking community. In *Proceedings of the British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction*, pages 111–119, 2008.

[35] Juliana Sutanto, Elia Palme, Chuan-Hoo Tan, and Chee Wei Phang. Addressing the personalization-privacy paradox: An empirical assessment from a field experiment on smartphone users. *MIS Quarterly*, 37(4):1141–1164, 2013.

[36] Eran Toch, Norman M. Sadeh, and Jason Hong. Generating default privacy policies for online social networks. In *Extended Abstracts on Human Factors in Computing Systems*, pages 4243–4248, 2010.

[37] Eran Toch, Yang Wang, and LorrieFaith Cranor. Personalization and privacy: A survey of privacy risks and remedies in personalization-based systems. *User Modeling and User-Adapted Interaction*, 22(1-2):203–220, 2012.

[38] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. Twitterrank: Finding topic-sensitive influential twitterers. In *Proceedings of the Conference on Web Search and Data Mining*, pages 261–270, 2010.

[39] Pamela Wisniewski, Bart P Knijnenburg, and Heather Richter Lipford. Profiling Facebook users privacy behaviors. In *Symposium on Usable Privacy and Security*, 2014.

[40] Xiaowei Xu, Nurcan Yuruk, Zhidan Feng, and Thomas AJ Schweiger. SCAN: A structural clustering algorithm for networks. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 824–833, 2007.

[41] Faiyaz Al Zamal, Wendy Liu, and Derek Ruths. Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. In John G. Breslin, Nicole B. Ellison, James G. Shanahan, and Zeynep Tufekci, editors, *ICWSM*. The AAAI Press, 2012.

[42] Sergej Zerr, Stefan Siersdorfer, Jonathon Hare, and Elena Demidova. Privacy-aware image classification and search. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 35–44, 2012.

Chapter 5

Privacy Preference Inference via Collaborative Filtering

5.1 Introduction

Providing personalized content can mutually benefit businesses and customers. Effective user profiling, however, hinges on collecting large amounts of data from users. Hence, privacy is the cornerstone of any personalization activity that, if disregarded, will influence gratifications derived from the personalization and can lead to irrecoverable trust issues. This issue may be especially problematic in the context of social networks, wherein massive numbers of daily social interactions are taking place nowadays.

Even though users are enabled to protect their data by using privacy controls, such privacy decisions are often complex, requiring careful examination of the trade-offs between the potential social gain and possible privacy risks. Therefore, many users avoid the hassle of privacy configuration and follow the default settings [12]. However, they are normally unaware that the default setting is often open and permissive. Even among the ones that make the effort to manage their privacy, many are still unaware of the implications of their decisions [7].

Although limited in quantity, earlier research on privacy behaviour has had modest success in predicting user's privacy preferences by relying on their social footprints [4, 2]. Here, we propose a Collaborative Filtering (CF) approach that combines neighbourhood-based techniques with a latent factor model inferred from the social graph of users. Our

A version of this chapter has been published in the *proceedings of the International Conference on Web and Social Media (ICWSM)* as a poster paper.

privacy prediction approach is novel and aims to detect privacy-concerned users with publicly available profiles. The contributions of this paper are as follows: 1) adapting a hybrid CF method to infer social privacy preferences, 2) exploring the usefulness of profile attributes in privacy preference detection, 3) establishing and analyzing the properties of a privacy-enhanced social graph, and 4) discovering a set of latent factors related to privacy attributes.

Our study is conducted on Twitter, wherein privacy control follows a binary specification. In Twitter, users can either follow the default *public* setting, which indicates that their tweets and contacts are publicly available, or they can change the setting to *protected*, which makes their tweets and contacts accessible only by their approved followers.

5.2 Related Work

CF methods have shown great promise in the development of recommendation systems, where unknown preferences of users are identified using known preferences of other users. Such approaches can be particularly valuable in the context of social media due to the existence of additional social relations. However, limited attempts have been made to adapt CF techniques for the prediction of privacy preferences.

For instance, Squicciarini et al. [11] first form social circles based on users' characteristics (e.g., gender and hobbies). When a new object is uploaded by the focal user, the system then seeks the social circles that are most likely to deal with the object in a similar way as the user. Then the privacy policies used by the selected circle are the basis for predicting the policy for the added object.

In [10], active learning and the properties of the social graph are first used to detect a set of informative contacts to be labeled as training samples. In the labeling process, the user specifies whether he/she is willing to share a specific item with the selected contact. Then labels are propagated from labeled instances to unlabeled ones in the graph. This propagation is guided by the user similarity metric that is computed based on contacts' profile information, along with their network and community metrics.

CF is also followed in [3], where a set of profile features, users' interests, and privacy configurations are used to find a set of users similar to the focal user. In their work, users are first characterized according to their privacy preference as privacy fundamentalist, privacy pragmatist, or privacy unconcerned. Users are assigned to these categories based on the number of their public, customized, and private photo albums. Then K-nearest neighbour algorithm is used to determine which privacy categorization the focal user belongs to.

CF-based techniques to privacy prediction mainly used neighbourhood-based tech-

niques, where privacy observations from neighbouring or similar users are the basis of the prediction. However, latent factor models, which are the current state-of-the-art for CF [1], are yet to be explored. Earlier work also lacks the study of hybrid techniques that have shown significantly better results compared to pure neighbourhood-based and pure latent factor models in other domains [6].

5.3 Neighbourhood-based Latent Factor Model

Two popular techniques to CF are neighbourhood-based methods and latent factor models. In neighbourhood-based methods, observations from neighbouring and/or similar users is used to detect attributes and preferences of the focal user. Detecting privacy preferences using such user-oriented techniques introduces unique challenges due to the lack of observable information associated with private accounts. In Twitter, a limited number of profile attributes are visible for both *public* and *protected* accounts. Hence, such profile attributes can be used to measure user similarity. The privacy preference of the focal *public* user i can then be determined using the following:

$$y_i = \frac{\Sigma_{i' \in \Omega_i} \omega_{ii'} x_{i'}}{\Sigma_{i' \in \Omega_i} \omega_{ii'}}$$

where $\omega_{ii'}$ measures the similarity between user i and its neighbours $i' \in \Omega_i$. In addition, $x_{i'} \in \{\varepsilon, 1\}$ indicates the actual privacy setting of the neighbour, wherein ε represents a very small value:

$$x_{i'} = \begin{cases} \varepsilon & \text{if } i' \text{ is } \textit{public} \\ 1 & \text{if } i' \text{ is } \textit{protected} \end{cases}$$

Based on this formula, a larger value of y_i indicates a higher level of privacy concern for user i. To gain insight into the potentials and limitations of Twitter profile attributes for our task, we carried out a set of experiments. These experiments, which are briefly presented in the next section, suggest the value of profile attributes in the inference of privacy preferences and indicate their potential for similarity measurement in a neighbourhood-based approach.

Meanwhile, latent factor models have gained tremendous success in the context of recommender systems. Such methods aim to discover a set of informative latent factors regarding users and later use such attributes to infer preferences. We can calculate the likelihood that the *public* user i is privacy-concerned using the following:

$$p(y_i) = p(y_i|\theta_1, ..., \theta_k) = \Pi_{j=1,...,k} p(y_i|\theta_j)$$

where $u_i = \theta_1, ..., \theta_k$ and $p(y_i|\theta_j)$ indicates the probability of user i being privacy-concerned if user i is associated with the latent factor θ_j. Later in this document, we propose a technique to discover such latent variables from a social graph of users. In this graph, users are first transformed into a set of latent variables. The probability of a latent attribute being associated with private people can then be calculated based on the number and/or the ratio of its *protected* neighbours. Finally, these two approaches can be merged in a single model to effectively capture privacy preferences [6].

5.4 Profile Attributes for Privacy Prediction

To analyze the relations of profile attributes and privacy behaviour, we first built a directory of Twitter users by collecting the followers of several famous Twitter accounts (e.g., "Facebook", "Katy Perry", "Obama"). For each account, we then calculated the percentage of the *protected* followers to the total number of followers. The results indicate that the percentage is considerably higher for "CNN Breaking News"(11%) compared to the other follower sets (between 5% to 7%) and the average percentage in Twitter (4.8%). We thus selected this set for our study since the privacy attitude-behaviour dichotomy seems to be minimized.

We then analyzed this user set, which includes a balanced set of almost 1M users, to gain insight into the potential differences in how profile attributes of *public* and *protected* accounts are configured. We focused our analysis on a set of profile features that are readily available from Twitter[1] accounts, along with additional features developed based on the existing profile attributes. For instance, we used a directory of English names to analyze if the declared name includes an actual person name. In addition, we examined linguistic attributes of profile *descriptions* using LIWC[2] and keyword frequency analysis.

As a result, a feature set of size 27 is developed, a summary of which is provided in Table 5.1. The first column in the table shows the Twitter API features. The second column presents the linguistic attributes extracted from profile *descriptions*. In addition to a set of LIWC categories[2], this list includes four keyword-based features that indicate the presence of the corresponding keyword in the *description*. The differences between *protected* and *public* accounts are statistically significant across all these features (as determined by chi-square or t-test results depending on the feature type). For four of these features, the

[1] https://dev.twitter.com/overview/api/users
[2] http://liwc.wpengine.com/

Table 5.1: Profile features to detect *protected* accounts.

Twitter Attributes	Linguistic Attributes
Tweet Count*	Six Letter Words
Friend Count	Function Words
Favirote Count	Clout
List Count	Emotional Tone
Actual Name Count	Authentic
Is Geo-Enabled*	Analytical Thinking
Has Actual Name	Affect Words
Username Has Name	Social Processes
Has URL	Cognitive Processes
Has Location*	Relativity
Has Default Image	Drivers and Needs
Has Default Profile*	"follow"
	"business"
	"smile"
	"@username"

differences are practically significant as well (as determined by Cramer's V or Cohen's d depending on the feature type). These four features are marked by asterisk in the table. Hence, profile attributes are distinctive across *protected* and *public* users and proved to be of value for our task of privacy preference inference in a neighbourhood-based approach. Utilizing this feature set in a regression-based supervised algorithm resulted in an F-score of 0.72, outperforming a random baseline by over 20%. The details of the feature set and machine learning experiments can be found in [5]

5.5 Privacy Graph and Latent Attributes

5.5.1 Graph Construction and Properties

Our approach is intended to exploit homophily [8] and is based on the fact that people of similar interests tend to connect with each other. Therefore, instead of using the asymmetric follow or friend relation in Twitter, we built an undirected mutual graph of users that only includes the edges that are reciprocated. Reciprocated relations are expected to indicate a stronger relationship between the two users, and they distinguish the social network section of the Twitter sphere from its information network [9].

Starting from a random *public* user, we iteratively built a mutual graph of users in a Breadth First Search (BFS) manner. For each *public* user, we first counted the number of *protected* mutual neighbours as well as the ratio of *protected* to all mutual neighbours.

Figure 5.1: A snapshot of the privacy graph.

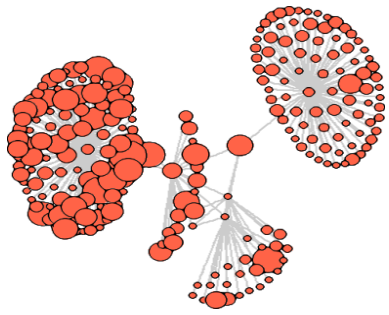
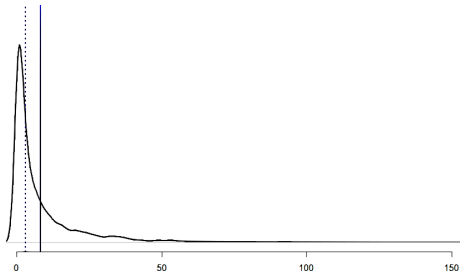


Figure 5.2: The distribution of *protected* contacts.



This user is then annotated with these metrics and is added to the graph. We then check if the new node has a reciprocated relationship with any existing node in the graph and add the corresponding edges. This process is repeated with a new *public* user pulled from the BFS queue. Users with less than 10 tweets or less than 30 followers/friends are considered inactive and thus are not added to the graph. In addition, *verified* users and users with more that 1K followers/friends are not included since they often represent brands and celebrities and are not from the general public. We collected the total of 3K *public* nodes that are annotated based on their privacy ratio metric. Figure 5.1 shows a snapshot of a small portion of the graph visualized using a force-directed layout, wherein the privacy ratio metric is mapped to the node size. In this dataset, each Twitter account is mutually connected to an average of 77 contacts. Among these neighbours, an average of 69 are *public* and 8 are *protected*. Figure 5.2 shows the distribution of *protected* neighbours, where mean and median are marked by a solid and a dotted line, respectively. As can be seen, the distribution is skewed to the right, indicating that despite the smaller number of users with a large number of *protected* neighbours, these numbers are considerably large so that the mean is dragged to the right.

To ensure that homophily applies in the context of privacy, we calculated the correlation between the privacy ratio of each node and the average privacy ratio of the neighbours. That analysis of about 700 users with at least 10 mutual contacts in the graph showed a strong

positive correlation between the two variables (Pearson correlation coefficient $r = 0.88$). This result indicates that users' privacy behaviour is either influenced by their close social contacts or individuals with similar privacy behaviour tend to cluster together in social networks. In either case, this finding implies the great potential of CF for privacy preference prediction.

5.5.2 Latent Attribute Detection

To discover a set of latent attributes that are of interest to privacy-concerned users, we transformed each node into a set of attributes. For each attribute node, we then calculated the number and the ratio of *protected* neighbours. The resulting graph is expected to enable the computation of $p(y_i|\theta_j)$, which indicates the likelihood of user i being private in case he/she is labeled with attribute θ_j. For instance, the privacy ratio of each latent factor may serve as such a probability value. We conducted experiments using unigrams and hashtags extracted from user tweets as latent variables. These attributes are selected from the 500 most recent tweets published by each user. Table 5.2 displays the top 10 hashtags and unigrams that are associated with at least 10% (n = 30) of the users. These features are extracted based on the privacy ratio metric, which is also provided in the table.

The top hashtag in table is "#neverforget", which is a commemorative political slogan that encourages remembrance for national and international tragedies. The high privacy ratio of this hashtag, along with the top-ranked hashtags of "#USA" and "#respect" can indicate that communities interested in political topics tend to be more private. This finding is interesting and can be considered inline with a high percentage of *protected* CNN followers. However, given that the graph was collected in a particular timeframe, these results can be time-specific. Another time-oriented hashtag apparent in the list is "#aquarius". Collecting data over time and clustering hashtags into high-level topics can shed light on the potential relations between users' interests in politics and zodiac signs and their privacy preference.

We also explored the collected tweets to understand the context in which the top unigram "separate" is used. The majority of these tweets are related to user's love lives and relationships. Interestingly, "#relationshipgoals" and "#lastrelationshiptaughtme" have a high privacy ratio in our list. Therefore, one may conclude that users who share about their relationships in social media are more likely to be privacy-concerned. Again, grouping hashtags into general concept and topics can provide more accurate results.

Surprisingly, in the ranked list of hashtags, "#personal" was placed last with the lowest privacy ratio of 0.25. Although this finding may seem counterintuitive, our examination

Hashtags		Unigrams	
#neverforget	19.89	seperate	32.07
#USA	19.57	reward	31.54
#respect	19.25	dull	31.1
#cantwait	18.49	deepest	29.55
#FML	17.62	unforgettable	28.66
#YOLO	16.61	activities	27.98
#instantfollow	16.52	lyric	27.55
#aquarius	16.46	circumstances	25.96
#winning	15.83	somethings	25.83
#soundcloud	15.52	forgiving	25.62

Table 5.2: Latent factors extracted from the privacy graph.

of the tweets showed that this hashtag is mainly used in conversations, wherein users are asked to provide some information, but they refuse to do so by replying a tweet that contains "#personal". Hence, it is likely that they are privacy-aware people that deliberately use the *public* setting, which is inline with what the latent factor model probability indicates. Overall, despite the limited size of the graph, the results are sensible and can reveal interesting information about private neighbourhoods in Twitter.

5.6 Conclusion

To predict one's privacy preference, we proposed a CF method that utilizes both neighbourhood-based and latent factor models. We analyzed the benefits of using profile attributes to measure user similarity and the use of hashtags and unigrams as latent features. The data collection process is currently ongoing, and we will run similar studies on multiple graphs built using different seed users. We also will examine a variety of other user attributes for the latent factor model. Robust evaluation methods will be developed to verify the usefulness of the approach. While we are focused on a simplied form of privacy here (i.e., binary specification), attempts will be made to analyze complex forms and strategies of privacy protection.

Bibliography

[1] Apoorv Agarwal, Fadi Biadsy, and Kathleen R. Mckeown. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic N-grams. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pages 24–32, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[2] Cailing Dong, Hongxia Jin, and Bart Knijnenburg. Predicting privacy behavior on online social networks. In *Proceedings of the AAAI Conference on Web and Social Media*, 2015.

[3] Kambiz Ghazinour, Stan Matwin, and Marina Sokolova. Monitoring and recommending privacy settings in social networks. In *Proceedings of the Joint EDBT/ICDT Workshops*, pages 164–168, 2013.

[4] Taraneh Khazaei, Lu Xiao, Rober Mercer, and Atif Khan. Detecting privacy preferences from online social footprint: A literature Review. In *Proceedings of the iConference*, 2016.

[5] Taraneh Khazaei, Lu Xiao, Rober Mercer, and Atif Khan. Privacy Behaviour and Profile Configuration in Twitter. In *Proceedings of the Conference on World Wide Web - Companion Volume*, 2016.

[6] Yehuda Koren. Factorization meets the neighborhood: A multifaceted collaborative filtering model. In *Proceedings of the Conference on Knowledge Discovery and Data Mining*, pages 426–434, 2008.

[7] Yabing Liu, Krishna P. Gummadi, Balachander Krishnamurthy, and Alan Mislove. Analyzing Facebook privacy settings: User expectations vs. reality. In *Proceedings of the ACM SIGCOMM Conference on Internet Measurement Conference*, pages 61–70, 2011.

[8] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001.

[9] Seth A. Myers, Aneesh Sharma, Pankaj Gupta, and Jimmy Lin. Information network or social network? The structure of the twitter follow graph. In *Proceedings of the International Conference on World Wide Web*, pages 493–498, 2014.

[10] Mohamed Shehab and Hakim Touati. Semi-supervised policy recommendation for online social networks. In *Proceedings of the Conference on Advances in Social Networks Analysis and Mining*, pages 360–367, 2012.

[11] Anna Squicciarini, Sushama Karumanchi, Dan Lin, and Nicole DeSisto. Identifying hidden social circles for advanced privacy configuration. *Computers & Security*, 41:40 – 51, 2014.

[12] Katherine Strater and Heather Richter Lipford. Strategies and struggles with privacy in an online social networking community. In *Proceedings of the British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction*, pages 111–119, 2008.

Chapter 6

Computational Analysis of Collective Intelligence: A Review

6.1 Introduction

The concept of intelligence in the human brain has been extensively explored in psychology research for years. In recent decades, there has been a shift in intelligence research from only considering cognition as something in one person's brain to social and collective intelligence that arises from local interactions within groups of individuals [28]. A well-studied form of intelligence in social structures is the phenomenon of "swarm intelligence" that has been widely observed in nature. Swarm intelligence studies explain how animal groups can perform highly complex and intellectual tasks, which are far beyond cognitive abilities of single animals.

Ant colonies, for example, have long fascinated scientists by exhibiting complex problem solving and intelligent features such as finding the shortest route to a food source or or finding the best nest among the multiple options [37]. In addition to ant colonies, bird flocking, animal herding, bacterial growth, and fish schooling are other instances of self-organized living groups that exhibit signs of intelligence [18].

Even though collective intelligence in humans and swarm intelligence in animals has the "emergent" attribute in common, they differ in two main features [36]. First, animal swarms involve cognitively simple agents, whereas collective intelligence involves agents with complex cognitive abilities, i.e., humans. Second, agents of each type use different

A version of this chapter has been published in the *proceedings of the Hawaii International Conference on System Sciences (HICSS).*

methods to interact with one another. Animal swarms usually use stigmergy-based interactions between individuals [56, 36]. Stigmergy is a form of indirect interaction that takes place through subsequent modifications of a shared environment [41]. For example, ants lay down pheromone on their way back to the nest when they have a found food source. Other ants then get stimulated by the pheromone and follow the trails with stronger scent. Since ants pass shorter paths in less time, their collective actions finally lead to finding the shortest route to food.

In contrast with animal swarms, interactions in human collectives can either be stigmergy-based (i.e., indirect) or direct. In addition, the interactions of both forms can take place through two different channels: verbal communication or non-verbal action. Following these two binary distinctions, interactions among humans can be classified into four different groups: direct verbal communication, direct non-verbal action, indirect verbal communication, and indirect non-verbal action. Deliberation and discussion, pedestrians forming bi-directional lanes in a crowded sidewalk [43], joint authorship, and finding the optimized route through active trail formation [25] are traditional examples of each form, respectively.

Meanwhile, the advances in computer-mediated communication and the emergence of Web 2.0 have enabled mass data exchange at a global scale, leading to new and advanced forms of collective intelligence. Wikipedia, Google, Q&A forums, and social networks are all examples of Web-based collective intelligence platforms. For instance, Wikipedia uses verbal stigmergy-based interactions, where users can modify local parts of their shared virtual environment. In open source projects, participants' actions are made on top of each other to create complex products such as the Linux operating system. On the other hand, in the environments such as Q&A forums, social networks, and deliberation systems, collective intelligence may be triggered by direct conversation and online dialogue among users. Figure 6.1 illustrates different types of interaction in collective intelligence along with their corresponding Web-based examples.

Web-enabled collective intelligence platforms allow crowds to gather and contribute in solving the problems that may be difficult or impossible to solve by even the smartest individuals and fastest computers. Various analytical techniques are developed to understand how intelligence emerges from within social interactions and to determine various micro and macro level factors that may positively or negatively influence the collective intelligence phenomenon.

Hence, the primary focus of this thesis is to perform empirical analysis of varied and large-scale user-generated datasets and to develop computational tools and models to better understand collective intelligence. The resulting analytics will then be utilized along with information visualization and human computer interaction principles to design and

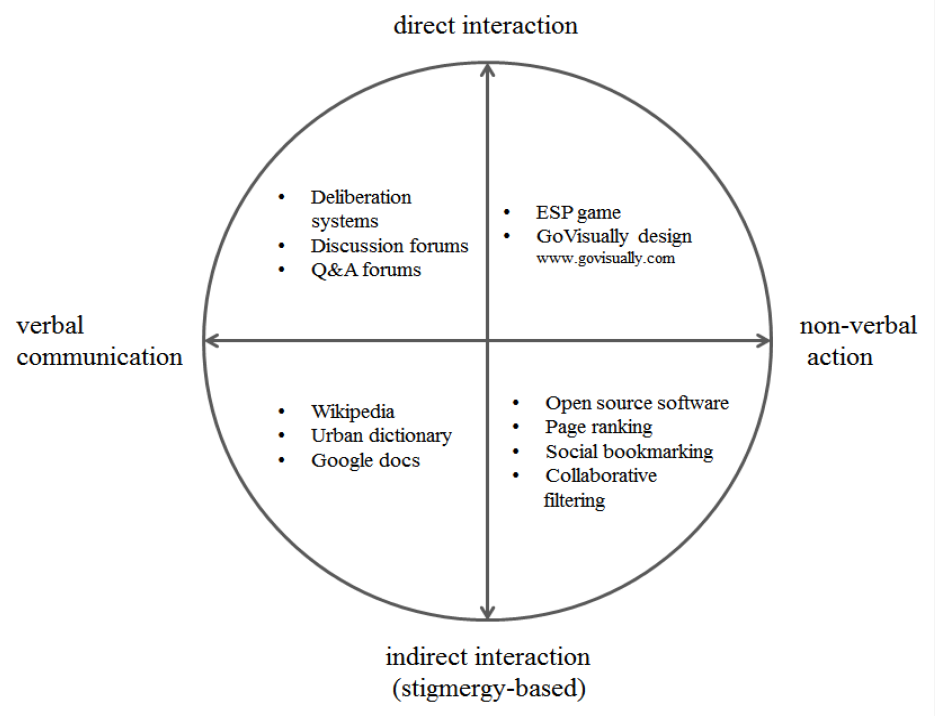


Figure 6.1: Collective intelligence can be categorized into four different groups based on the interaction types among humans.

develop Web-based tools that effectively harness the intelligence power of human collectives. This research mainly focuses on online discourse-centric data that involves at least some direct interrelations among the participants (left upper quadrant in Figure 6.1). Even though computational tools have been developed to extract different aspects of intelligence from collections of product and movie reviews, microblogs, and news headlines, they are independent pieces of writings created by individuals and so are excluded from this study.

To set a research agenda in the direction of fostering and analyzing collective intelligence on the Web, we reviewed the existing computational approaches that utilized automated methods to study large-scale Web-based collective intelligence platforms. We specifically focused on the collective intelligence that emerges from verbal interactions. Therefore, our review excluded the platforms for online reviews, microblogs, and news headlines since these text units are independent pieces of writings created by individuals. In addition, we focus on potentially repeated interactions in bounded environments, so the studies about Web and social networks in general, such as [34, 32], are not included. Chat and email tools are also excluded because they are primarily designed to support one-to-one and small-scale interactions and are not normally seen as large-scale platforms.

In the next section, we describe the group study framework used to classify the existing literature. Then we provide more detailed explanations of the studies in each category.

We next present the review and the analysis of the literature and discuss a set of identified research gaps.

6.2 Methodology

6.2.1 Framework for the Systematic Review

Although there are various models about small group processing (e.g., McGraths group process model [42], Endsleys model of situation awareness [16], and Tuckmans stages model [57]), we are not able to find a model that is specific to large-scale group activities. McGrath defines groups as "social aggregates that involve mutual awareness and potential mutual interaction" [42]. Originally intended to address relatively small and structured groups, McGraths model is not merely a model or theory about small group activities, but a framework for studying groups systematically by considering all the factors that potentially influence the process and outcome of a group activity and how they interplay. We assume that these factors also exist in larger online collectives and share similar interplay structures as outlined in McGraths model. Therefore, we chose this model for our systematic review and applied it to research on larger online collectives that involve mutual interaction among individuals.

A simplified version of the McGrath framework is illustrated in Figure 6.2. In this framework, *group interaction process* is considered the central feature and the essence of a group and it refers to "patterned relations among the behaviors of individuals" [42]. In addition to this primary component, the framework encompasses several elements that can somehow relate to group interaction processes. For instance, participants influence and might be influenced by interaction processes; thus, *member's properties* is incorporated into the framework. The pattern of relations among participants is an important factor in understanding interactions within individuals and so it appears as *standing group* or *group structure* in the framework. Two other important facets included are the *environment* where group interactions takes place and the *tasks* (i.e., informally assumed goals) that group members pursue.

In addition to the overall framework of group studies, McGrath describes interaction processes at a micro level, viewing it in terms of three different stages. The first stage is considered with the structure and distribution of communications among interacting people, which is referred to as the *communication pattern*. As the second stage, each communication is considered in regard to its *content*. The content itself is then viewed in terms of two different aspects: *interpersonal relationship* and *task performance*. Finally, the third

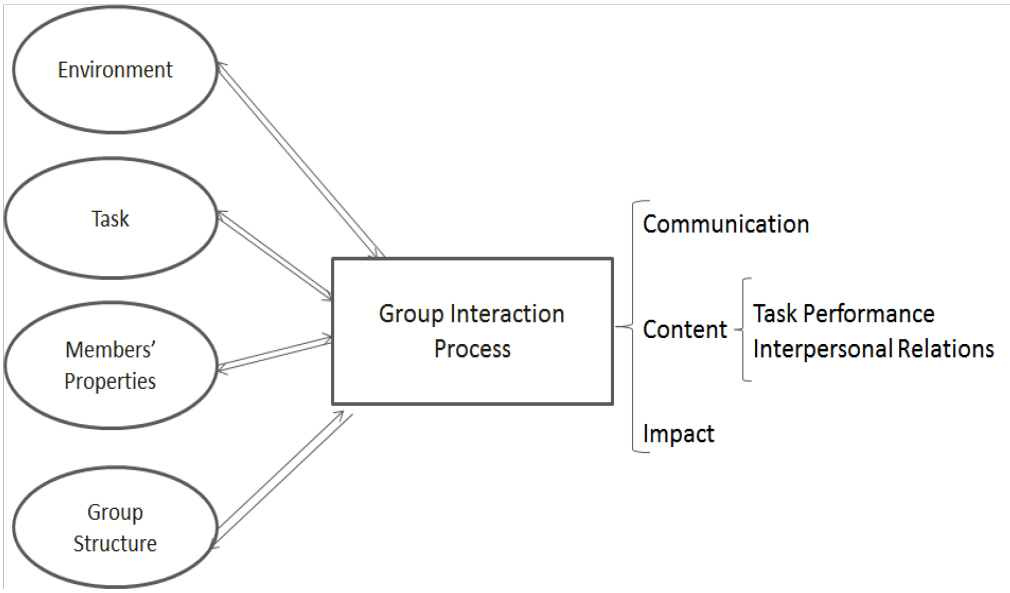


Figure 6.2: McGrath framework can used to classify literature on online collectives.

stage is related to the *impact* of different group interaction processes on each other and on participants.

Current research direction on analysis of collective intelligence commonly treat members' properties (e.g., gender), group structure (e.g., hierarchical), environment (e.g., discussion forum), and task (e.g., decision making) as the context and independent variables of the study and then facilitates various computational methods to study the group interaction process component. Therefore, related works on collective intelligence analysis are categorized into three main groups: communication pattern analysis, content analysis, and impact analysis. Content analysis is then divided into two categories of research. Due to the different nature of virtual spaces, interpersonal relationships component is seen at a higher level, focusing on studies that analyze emotional and sentimental aspects of the content. In addition, the task performance component of content analysis is called task/purpose performance since such settings can be less task-oriented.

6.2.2 Article Selection Process

Since this review is focused on computational methods and algorithms to analyze collective intelligence, we started the article selection process by conducting several keyword searches on ACM Digital Library [2]. Queries were formed by borrowing keywords from McGrath framework of group studies, such as "communication pattern analysis" and "sentiment analysis". An initial analysis of the top 100 documents retrieved by each query was

performed by reading titles, abstracts, and introductions to filter the set to only include those articles that are within the scope of this study. Therefore, to identify the relevant articles, we used the following criteria for each study: 1) proposed automated approaches with minimal human coding and effort, and 2) performed the analysis on large-scale Web-based platforms that involve sharing of information and dialogue among multiple participants.

Once the initial set of documents was selected, backward and forward chaining techniques were used to cover relevant articles in other venues. The purpose was not to cover every single study, but to provide a representative sample of articles to gain a sufficient understanding of the research themes and gaps in the current state of the literature.

6.3 Review of the Literature

6.3.1 Communication Pattern Analysis

Communication pattern refers to the form or structure of a series of interactive behaviours or communications that takes place among people [42]. A recent study by Woodley et al. [64] has shown that the patterns of turn taking among participants is correlated with the collective intelligence of the group. This finding implies the importance of the analysis of communication patterns in understanding and fostering intelligence among human collectives.

Even though earlier works in such contexts have relied on basic statistical measures [63, 52, 66, 59], more sophisticated computational techniques on analyzing communication patterns have followed two main directions: social network analysis and analysis of conversation threads as tree structures. In addition, instead of focusing on group level patterns, a number studies have investigated communication patterns as individual's behaviour.

A large amount of studies attempt to generate a social network of participants and then employ various graph analysis methods proposed by social network analysts to better understand the underlying communication patterns [49, 21, 31, 20, 14, 65, 47, 3, 35]. In social networks, each participant corresponds to a node $i \in V$ in a graph $G = \langle V, E \rangle$, where an edge $(i, j) \in E$ in the graph indicates some sort of social relation between the two participants i and j. When studying communication patterns, the edges normally represent messages and online dialogues exchanged among participants. Different types of social networks can be formed based on different interpretations of how two participants should be linked such as whether the edges should be directed, based on the message sender and receiver, or whether the edges should be weighted, based on the number of messages exchanged.

For example, Gomez et al. [21] studied three different types of social networks obtained

from the comment replies in the Slashdot news sharing website which correspond to different strategies for specifying an edge between a pair of nodes. They further analyzed different attributes of these networks such as maximum number of clusters, average path length, maximal distance between two participants, reciprocity, degree distribution, and assortative mixing by degree and score. Their finding implies that even though Slashdot shares some of its communication pattern features such as small world property and high clustering with traditional social networks, its participants showed neutral mixing by degree, identical in and out distribution, and merely moderated reciprocity, which are rarely observed in other social networks.

In another study, interaction patterns in Wikipedia talk pages are analyzed [31], where three different networks according to different types of interactions are extracted: article reply network, which includes direct replies in article discussion pages; users talk network, which includes direct replies in user talk pages; and wall network, which contains personal messages posted on talk pages of another user. These three networks are compared using an edge overlap metric. The results show a relatively high overlap between the two networks extracted from the talk pages, while the network from personal communications was found to be substantially different from the other two. In addition, the analysis of the directed assortatively by degree showed that the users who reply to many other users tend to reply to inexperienced users, while those who receive comments from many users are more likely to interact with each other.

Rather than forming social networks from conversations, some researchers have explored communication patterns as information cascades or threads [21, 31, 30, 22, 23, 49], where an action occurs due to the influence of others' activities [8]. In these studies, the cascades are modeled as a tree $T = \langle V, E \rangle$, where each node $i \in V$ represents an action by a participant, and there is an edge $(i, j) \in E$ between i and j if one of these actions is in response to or influenced by another.

In the study by Gomez et al. [21] on Slashdot discussions, discussion threads are mapped to radial trees, where a post is a central node and its comments are attached in different nesting levels based on their level to the central post. By analyzing the branching factor b, i.e., the number of comments provoked by a given comment, they found a high level of heterogeneity in discussion threads. In addition, the analysis results indicate a difference in the process underlying the generation of the initial level comments and the comments of the subsequent levels. Furthermore, the h-index metric used in scientific community [26] is adopted to determine the controversy level of a discussion.

In [30], the structure of threaded conversations in three different datasets are analyzed, namely Usenet groups, Yahoo! Groups, and Twitter. The relationships between the size

and the depth, the degree distribution and the thread level, and the size and the number of authors in threads are explored. A mathematical model is also developed based on the basic branching process. In the branching process, each node generates k children with probability p(k), where p is a probability distribution function. Taking the initial analysis into account, the authors have incorporated node degrees, their recency, and author identity into their model.

In their later work, Gomez et al. [22] analyzed four different social datasets and showed that a preferential attachment model with a bias toward initial message is able to capture most of the structural properties as well as evolution patterns of conversation cascades. In the preferential attachment model, the probability of an existing node to be linked to a new node is proportional to its degree distribution, meaning that the nodes with higher degrees are more likely to be linked to the new ones. To provide a more comprehensive model, recency information is utilized in [23], along with both bias toward initial message and degree distribution factors.

Laniado et al. [31] studied the shape and size of the discussion trees at the article level in Wikipedia and suggested different metrics to characterize discussion pages such as chains of direct replies, maximal depth, and the h-index of the tree. After categorizing Wikipedia articles and performing their analysis on different topics, they found a significant difference in discussion structures from different semantic areas.

While the aforementioned studies have focused on communication patterns at the group level, some attempts have been made to characterize individuals' communication behaviour in large-scale social settings. For example, a large number of studies have been conducted focusing on modeling and analyzing the time intervals of individual' communication activities [39, 7, 38, 58]. In addition, ego networks have been used to understand the communication patterns of individuals [3, 62, 20] such as identification of "discussion persons" and "answer persons" [3, 62] or identification of leaders [20]. Besides ego networks, Adamic et al. [3] introduced an entropy metric, capturing the concentration of a participant's reply patterns across topics in a Q&A forum.

6.3.2 Content Analysis

Sentiment Analysis

There has been a large amount of work on sentiment analysis, with researchers investigating various methods to identify subjective sentences, rate the sentiment level either at a binary (e.g., positive/negative) or ordinal scales (e.g., from 1 to 5), to detect moods and emotions (e.g., angry or happy), and finally to understand the source, target, and complex attitude

types [45]. Denoting the same field of study, many other terms have been used, among which subjectivity analysis and opinion mining are the most common.

Two main methods have been employed for automatic analysis of sentiment. The lexicon-based techniques have relied on annotated dictionaries that include sentiment polarity and ratings to determine the sentiment of a text unit based on the words and phrases it contains. Instead of using annotated dictionaries, supervised learning approaches build classifiers using labeled data to predict the sentiment of sentences, phrases, or documents. Various forms and combinations of these methods have been used and studied on generic text and online reviews, while relatively less attention has been paid to online discussions on social platforms.

Previous works focusing on lexicon-based methods have used both dictionaries built by human annotators, as well as automatically constructed ones built using unsupervised learning methods [55]. A traditional approach to use such dictionaries is to determine the sentiment score of a text unit by averaging out sentiment values of individual words extracted from the dictionary. Wanner et al. [60], for example, used a sentiment-bearing dictionary to assess sentiment scores.

Due to the importance of context in sentiment analysis, later approaches moved beyond negations and have considered discourse structure and other contextual features in the sentiment calculation process. For instance, Li and Wu [33] used a sentiment-labeled Chinese dictionary including lists of positive and negative words, privatives, and sentimentally weighted modifiers to assign sentiment rates to discussion posts. For each keyword in a given post, the sentiment rating is first assessed based on the dictionary. The algorithm then checks whether a privative is present within words before the keyword. If this is the case, the word's sentiment score is negated; otherwise, the algorithm moves to its last step. In this step, if a modifier is present in the words surrounding the keyword, the keyword sentiment rating is multiplied by the modifier sentiment weight extracted from the dictionary. The sentiment analysis results is then incorporated into feature vectors for k-means clustering, using which hotspots could be detected. In addition, a Support Vector Machine (SVM) classifier is trained on the sentiment feature vectors and the results of k-means to predict the next hot forum.

The increasing availability of the labeled instances on Web along with the emergence of crowdsouring platforms (e.g., Amazon Mechanical Turk), wherein labeled data can be effectively and efficiently collected, has led to the use of a variety of supervised classification methods in sentiment research. The feature sets normally used in sentiment learning techniques include term presence and frequency, term position, Part-Of-Speech (POS), syntax, and negation [45]. For example, Rosenthal and McKewon [48] adapted an exist-

ing method [4], which uses lexical scores of words, N-gram analysis, and polarity of surrounding syntactic constituents, to social media data by taking into account the emotions, acronyms, and misspellings.

Although supervised learning methods have been proven to be useful and accurate in sentiment classification tasks, when they are trained on a different domain than the domain of interest, their performance drops considerably [45]. In [54], a domain-independent algorithm [53], called reasoning through search, is used to classify the mood (i.e., angry, sad, happy) of news comments on Yahoo!Buzz based on the labeled instances obtained from the LiveJournal comment dataset.

To automatically detect sentences with attitude toward other participants in online discussions, Hassan et al. [24] took a somewhat different approach than the mainstream research on sentiment analysis. In their approach, a language model is constructed using the labeled data. However, rather than assigning labels to the text units, the language model assigns probability values [45]. The most relevant part of sentences with attitudes is where the second person pronoun is included. As such, this fragment is first extracted. Then using lexical items, POS tags, word polarity tags, and dependency relations, a set of patterns associated with each fragment is extracted. The patterns are then used to build Markova models for sentences with and without attitude. A new sentence can then be labeled as a sentence with or without attitude.

Task/Purpose Performance Analysis

The value of groups lies in their potential abilities to effectively perform wide range of tasks. Therefore, there has been an extensive classic research in social psychology conducted through field studies, surveys, and laboratory experiments, employing manual methods to explore how and in what conditions different groups can be task effective [5, 51]. In recent years, however, the proliferation of Web-based social platforms has called for automated approaches to process and asses the large-scale textual content generated by the crowd. A set of studies on online social environments attempts to address the information overload problem by organizing and describing the features of the underlying information space. Such studies can be seen as preliminary steps that can lead to easier and more accurate task performance analysis by machines or users. Another set of works, on the other hand, have devised novel computational methods to directly evaluate the quality and relevancy of conversational text on Web.

One of the early studies on describing online dialogues is conducted by Sack [49], in which basic natural language processing methods are used to extract and present a list of most frequently used terms in discussions along with their semantic network. Clustering

techniques [50, 13, 46] have also been utilized to describe content-based features of the communication space. Clustering refers to the groupings of objects based on some measures of similarity. In general, text clustering methods encode document contents in appropriate data structures, such as vector space model, and then employ various distance metrics to compute the similarity measure and to form document clusters. The underlying principle behind content-based clustering methods is that the more keywords the documents share, the more similar they are.

Document clustering is a well-studied topic; however, lack of structure and editorial supervision in conversational text imposes new challenges and complexities. As such, adoptions and extensions of methods used to address generic text is often proposed to cluster conversational text. For example, Said and Wannas [50] proposed Leader-based Posts Clustering (LPC) method, a modification to the Leader algorithm [6], to cluster posts in discussion threads. Their approach starts with the head post of the thread as the first leader. Using their suggested distance metric that involves both content and communication pattern (i.e., inter-post tagging) features, the post similarity to the existing clusters is calculated. If the resulting measure is below a certain threshold, the post is assigned to the corresponding cluster; otherwise, it becomes a new leader. This process continues until all the posts are clustered, leading to the detection of off-topic and outlier posts.

A closely related approach to document clustering is topic modeling, where researchers have used a variety of supervised and unsupervised mining methods to segment, label, and track different topics [67]. Regardless of the challenges of topic modeling for conversational text, some scholars have applied novel methods in small-scale discussions (e.g., [27, 19]); however, similar to clustering, topic modeling is less explored for large-scale Web-based conversations. One example is provided by Zhue et al. [67], which extends the basic Topic Detection and Tracking (TDT) algorithm to suit this method for data generated in online communities.

Instead of clustering and topic modeling, another line of work on descriptive linguistic analysis of online discussions have focused on the task of finding and labeling utterances and units of text that are of value in performance analysis such as identification of claims [48, 40], agreements or disagreements [1], justifications or argumentations [9], and ideas [11]. The approaches used in this research direction share some common ground with sentiment analysis studies as both are concerned with mapping a piece of text to a label from a pre-specified list. Additionally, some of the approaches in this direction utilize subjectivity analysis in identifying certain features (e.g., claims and agreements/disagreements). Thus, sentiment calculation is sometimes included in the approaches.

In the work presented by Rosenthal and McKewon [48], claims are automatically de-

tected using regression analysis on various features including sentiment, belief words, POS, and N-grams. They found that sentiment and POS tags were more important in LiveJournal, while committed belief and sentiment were more important for Wikipedia forums. This difference indicates the presence of distinctive linguistic characteristics in various social contexts.

Given a pair of sentences, with the first one being a claim, Biran and Rambow [9] devised a machine learning algorithm to automatically detect whether the second sentence is a justification of the claim. Their approach extracts a list of indicators of Rhetorical Structure Theory (RST) relations they deemed relevant for the task. For each indicator, a list of co-occurring word pairs is then extracted from English Wikipedia and forms the basis of the features they used in their supervised learning approach.

In [11], two sense making tools are designed and developed for idea management systems. The first tool, called idea spotter, uses speech act theory to automatically mark the idea core within unstructured textual information. The second tool, called comment interpreter, seeks specific linguistic characteristics within the comments made for an idea and then assigns the comment to one of the three pre-specified comment categories: reactions related to the content, value, or status of the idea. Using these labels, the system proactively suggests reactions to these comments based on the category the comment belongs to.

Another direction of research on task/purpose performance analysis employs computational methods to directly evaluate the user-generated content [54, 44, 12, 61, 17]. For instance, in order to calculate the relevance of comments to the actual post, [54] uses TF-IDF, in which comments are considered as queries and posts are seen as documents. Since they are applying their technique on a news-story commenting site, the corpus of Reuters news is used as the underlying collection to calculate the IDF metric. Similarly, [44] presents a novel method for detecting off topic posts in forums. In this techniques, a set of terms is extracted using BLTR word informativeness method to represent each thread. Each post is also represented with a set of terms. The similarity measure of these two vectors is combined with three other measures computing the similarity of the post with the lead post, with the preceding posts, and with all of the preceding posts to calculate the relevancy of each post to the thread.

Borrowing the HITS algorithm from information retrieval research [29], Feng et al. [17] proposed a graph-based method to detect the most important posts in a threaded discussion. To form a graph from conversational text, each message is represented by a node in the graph. A feature-based link generation method is proposed to place links between a pair of nodes based on the three features of lexical similarity, poster trustworthiness, and speech act relations. Then, for each node, two scores of hub (i.e., the quality of a node as a pointer

to others) and authority (i.e., the quality of a node as a resource) are iteratively calculated till the algorithm converges. Weimer et al. [61] applied SVM classifier based on various intrinsic features such as lexical, syntactic, and content similarity to assess the perceived quality of posts (i.e., good or bad) in discussion forums.

6.3.3 Impact Analysis

Impact analysis aims to understand the potential effects of group interaction processes, i.e., communication patterns, performance process, and sentimental aspects, on each other. In addition, understanding how these processes influence participants' behaviours such as their attitude, perception, judgement, and learning is of great interest. Employing computational methods, two different procedures have been followed to study impact in online environments. A set of works have focused on employing the existing methods to calculate the variables independently and then use basic statistical methods to assess if they correlate. A few recent approaches utilized more complex methods to directly assess the potential impact of one variable on another.

Analyzing a news commenting site, Diakopoulos and Naaman [15] examined the relationships between the topicality, temporality, sentiment, and quality of news comments. In their topicality analysis, the topics that aroused negativity and thus required more moderation activities were identified. A correlation was found between negativity and the number of deleted comments, suggesting that sentiment can be an indicator of the comment quality. In addition, in their temporal analysis, they found a positive correlation between frequency of commenting and the negativity of users comments, which also suggests a potential relation between sentiment and quality of comments. In a relevant study, the relationship between participants' activity behaviour and their sentiment is explored in BBC forums [10], finding that negativity boost participants' activity and more active users tend to express negative emotions.

Chmiel et al. [10] conducted a relatively complex analysis of online posts on different platforms to understand the impact of current emotions and sentiments expressed in a an online community on the emotions of the following posts. In their approach, chains of posts are first clusters into groups of consecutive posts with the same sentiment value (i.e., positive, neutral, or negative). A comparison of the generated clusters with the clusters formed based on random data showed a considerable difference, with clusters of the actual data being of a larger size. Their statistical analysis indicates that conditional probability for consecutive posts grow as a power law with cluster size, which is similar as preferential processes. Overall, they concluded that online posts tend to trigger post of similar

sentiment. In addition, they investigated the relation between thread size and the emotions expressed in the thread, showing that shorter threads tend to start with less negative sentiment and longer threads have larger sentiment variations.

6.4 Discussion

To identify the research gaps from the existing literature, we examined the papers primary and low-level purpose, the computational methods used, the environments of their focus, and the secondary benefits that they may provide to Web 2.0 technologies. The results are summarized in Table 6.1.

As shown in the table, despite the value and capabilities that lie in the task-oriented environments, majority of the previous works have been focused on environments primarily designed to facilitate social interactions (e.g., discussion fora and news/media comment sets). Little effort has been made to understand task-oriented online environments such as deliberation tools and idea management systems. There might have been a lot of usability and user experience (UX) studies that examine the impact of the environments through traditional UX methods. A new and underexplored direction in the literature is the development of computational techniques to study these environments by detecting and analyzing the "traces" of collective intelligence. It is expected that there are various factors that influence the "traces" such as the characteristics of the participants, the kinds of intelligence tasks, and the design of the environments.

The lack of sophisticated automated methods in the analysis of impact is also apparent in the table. While a large set of works have focused on employing the existing methods to calculate the variables independently and using basic statistical methods to assess if they correlate; only a few recent approaches utilized more complex methods to directly assess the potential impact of one interaction process on another or on participants. Supervised and unsupervised learning methods can be valuable and effective tools for modeling and potentially predicting such impacts.

In addition, as can be seen in the table, some attempts have been made to analyze the impact of interaction processes, i.e., communication pattern, task/purpose performance, and sentiment, on each other. However, less attention has been paid to understand how these processes influence participants' individual and collective behaviors over time, such as their perception, learning, and judgment.

Primary Purpose	Aspects of Focus	Main Methods	Environments	Secondary Benefits
communication analysis	interaction structure thread structure interaction behaviour	social network analysis tree analysis machine learning	comment sets social networks discussion fora Q&A fora	browse & navigation popularity detection & prediction controversy detection & prediction
content: sentiment analysis	subjective polarity & rating emotions and moods attitudes	lexicon-based machine learning language modeling	discussion fora comment sets	understanding social relations social action/behaviour detection community management marketing browse & navigation
content: performance analysis	topicality social actions & behaviours quality & relevance	clustering topic modeling supervised learning information retrieval	comment sets social networks discussion fora Q&A fora deliberation tools idea management	search & navigation marketing technology impact assessment content filtering & summarization
impact analysis	performance (topicality) & communication performance (topicality) & sentiment performance (quality) & sentiment sentiment & sentiment communication & learning	basic statistics clustering	comment sets discussion fora Q&A fora learning fora	understanding social & psychological processes technology impact assessment community management

Table 6.1: Prior research is summarized according to the McGraths framework.

In order to gain insights into how intelligence emerges from online interactions and to understand the various factors that may influence the phenomenon, further research is required to fill these gaps. We envision that a combination of computational techniques that analyze the traces and the content of collective intelligence, and visualization techniques that reveal the topical and temporal patterns over time is one research direction to address this gap. However, even though some attempts have been made to provide definitions of collective intelligence [6, 64], the lack of an agreed-upon operational definition for this concept has led to the lack of an empirical evidence to examine the potential influence of each of these factors on collective intelligence. Therefore, proposing well-grounded methods to measure collective intelligence can substantially contribute to the field.

6.5 Conclusion

Human societies have always been suffering from and dealing with "wicked" problems such as climate change, natural hazards, and healthcare. By enabling discussion and deliberation at a massive scale, Web 2.0 social platforms move beyond problem solving capabilities of a small group of authorities and can effectively harness the collective power of unique individuals at unprecedented scales. As such, it is of great value and interest to identify different patterns of discourse-based activities that correspond to intelligence.

The proliferative literature body of analyzing dialogue-based collective intelligence is promising. We conducted a systematic review of the prior studies that proposed computational techniques to analyze collective intelligence in large-scale user-contributed text of discourse. Our analysis reveals a set of research gaps in this area including the lack of focus on task-oriented environments, the lack of sophisticated methods to analyze the impact of group interaction process on each other, as well as the lack of focus on the study of the impact of group interaction processes on participants over time. We call for research activities to address these gaps and we believe that such work may contribute to improving our understanding of humans' collective behavior, fostering the development of collective work skills, and providing valuable insights on the design of collective intelligence systems.

Bibliography

[1] Rob Abbott, Marilyn Walker, Pranav Anand, Jean E. Fox Tree, Robeson Bowmani, and Joseph King. How can you say such things?!?: Recognizing disagreement in informal political argument. In *Proceedings of the Workshop on Languages in Social Media*, pages 2–11, 2011.

[2] ACM. ACM Digital Library. `http://dl.acm.org/`.

[3] Lada A. Adamic, Jun Zhang, Eytan Bakshy, and Mark S. Ackerman. Knowledge sharing and yahoo answers: Everyone knows something. In *Proceedings of the International Conference on World Wide Web*, pages 665–674, 2008.

[4] Apoorv Agarwal, Fadi Biadsy, and Kathleen R. Mckeown. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic N-grams. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pages 24–32, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[5] F.H Allport. *Social Psychology*. 1924.

[6] Ravindra T Babu and Narasimha M Murty. Comparison of genetic algorithm based prototype selection schemes. *Pattern Recognition*, 34(2):523–525, 2001.

[7] A. Barabiási. The origin of bursts and heavy tails in human dynamics. *Nature*, 435:207–211, 2005.

[8] Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades. *Journal of Political Economy*, 100(5):992–1026, 1992.

[9] Or Biran and Owen Rambow. Identifying justifications in written dialogs. In *Proceedings of the IEEE International Conference on Semantic Computing*, pages 162–168, 2011.

[10] A. Chmiel, J. Sienkiewicz, G. Paltoglou, K. Buckley, M. Thelwall, and J.A. Holyst. Negative emotions accelerating users activity in BBC forum. *Physica A*, 390(16):2936–2944, 2011.

[11] Gregorio Convertino, Ágnes Sándor, and Marcos Baez. Idea spotter and comment interpreter: Sensemaking tools for idea management systems.

[12] Constantin Daniil, Mihai Dascalu, and Stefan Trausan-Matu. Automatic forum analysis: A thorough method of assessing the importance of posts, discussion threads and of users' involvement. In *Proceedings of the International Conference on Web Intelligence, Mining, and Semantics*, pages 37:1–37:9, 2012.

[13] Kushal Dave, Martin Wattenberg, and Michael Muller. Flash forums and forumReader: Navigating a new kind of large-scale online discussion. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, pages 232–241, 2004.

[14] Munmun De Choudhury, Winter A. Mason, Jake M. Hofman, and Duncan J. Watts. Inferring relevant social networks from interpersonal communication. In *Proceedings of the International Conference on World Wide Web*, pages 301–310, 2010.

[15] Nicholas Diakopoulos and Mor Naaman. Topicality, time, and sentiment in online news comments. In *Extended Abstracts on Human Factors in Computing Systems*, pages 1405–1410, 2011.

[16] Mica R. Endsley. Toward a theory of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 37(1):32–64, 1995.

[17] Donghui Feng, Erin Shaw, Jihie Kim, and Eduard Hovy. Learning to detect conversation focus of threaded discussions. In *Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference*, pages 208–215, 2006.

[18] L. Fisher. *The Perfect Swarm: The Science of Complexity in Everyday Life*. Basic Books, 2009.

[19] Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. Discourse segmentation of multi-party conversation. In *Proceedings of the Annual Meeting on Association for Computational Linguistics*, volume 1, pages 562–569, 2003.

[20] Peter A. Gloor, Rob Laubacher, Scott B. C. Dynes, and Yan Zhao. Visualization of communication patterns in collaborative innovation networks - Analysis of some

W3C working groups. In *Proceedings of the International Conference on Information and Knowledge Management*, pages 56–60, 2003.

[21] Vicenç Gómez, Andreas Kaltenbrunner, and Vicente López. Statistical analysis of the social network and discussion threads in slashdot. In *Proceedings of the International Conference on World Wide Web*, pages 645–654, 2008.

[22] Vicenç Gómez, Hilbert J. Kappen, and Andreas Kaltenbrunner. Modeling the structure and evolution of discussion cascades. In *Proceedings of the ACM Conference on Hypertext and Hypermedia*, pages 181–190, 2011.

[23] Vicenç Gómez, HilbertJ. Kappen, Nelly Litvak, and Andreas Kaltenbrunner. A likelihood-based framework for the analysis of discussion threads. *Proceedings of the International Conference on World Wide Web*, 16(5-6):645–675, 2013.

[24] Ahmed Hassan, Vahed Qazvinian, and Dragomir Radev. What's with the attitude? Identifying sentences with attitude in online discussions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1245–1255, 2010.

[25] Dirk Helbing, Frank Schweitzer, Joachim Keltsch, and Péter Molnár. Active walker model for the formation of human and animal trail systems. *Physical Review E*, 56:2527–2539, 1997.

[26] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46):16569–16572, 2005.

[27] Joshua E. Introne and Marcus Drescher. Analyzing the flow of knowledge in computer mediated teams. In *Proceedings of the Conference on Computer Supported Cooperative Work*, pages 341–356, 2013.

[28] J.F. Kennedy, J. Kennedy, R.C. Eberhart, and Y. Shi. *Swarm Intelligence*. Evolutionary Computation Series. Morgan Kaufmann Publishers, 2001.

[29] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Jornal of the ACM*, 46(5):604–632, 1999.

[30] Ravi Kumar, Mohammad Mahdian, and Mary McGlohon. Dynamics of conversations. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 553–562, 2010.

[31] David Laniado, Riccardo Tasso, Yana Volkovich, and Andreas Kaltenbrunner. When the Wikipedians talk: Network and tree structure of Wikipedia discussion pages. In Lada A. Adamic, Ricardo A. Baeza-Yates, and Scott Counts, editors, *International AAAI Conference on Weblogs and Social Media*, number 177-184, 2011.

[32] K. Lerman and R. Ghosh. Information contagion: An empirical study of the spread of news on Digg and Twitter social networks. In *Proceedings of International Conference on Weblogs and Social Media*, 2010.

[33] Nan Li and Desheng Dash Wu. Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision Support Systems*, 48(2):354 – 368, 2010.

[34] David Liben-Nowell and Jon Kleinberg. Tracing information flow on a global scale using internet chain-letter data. In *Proceedings of the National Academy of Sciences*, volume 105, page 46334638, 2008.

[35] A. De Liddo, S. Buckingham Shum, I. Quinto, M. Bachler, and L. Cannavacciuolo. Discourse-centric learning analytics. In *Proceedings of the International Conference on Learning Analytics and Knowledge*, pages 23–33, 2011.

[36] Shuangling Luo, Haoxiang Xia, Taketoshi Yoshida, and Zhongtuo Wang. Toward collective intelligence of online communities: A primitive conceptual model. *Journal of Systems Science and Systems Engineering*, 18(2):203–221, 2009.

[37] E. Mallon, S. Pratt, and N. Franks. Individual and collective decision-making during nest site selection by the ant Leptothorax albipennis. *Behavioral Ecology and Sociobiology*, 50(4):352–359, 2001.

[38] R. D. Malmgren, D. B. Stouffer, A. E. Motter, and L.A.N Amaral. A Poissonian explanation for heavy tails in e-mail communication. In *Proceedings of the National Academy of Sciences*, volume 105, pages 18153–18158, 2008.

[39] R. Dean Malmgren, Jake M. Hofman, Luis A.N. Amaral, and Duncan J. Watts. Characterizing individual communication patterns. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 607–616, 2009.

[40] Alex Marin, Bin Zhang, and Mari Ostendorf. Detecting forum authority claims in online discussions. In *Proceedings of the Workshop on Languages in Social Media*, pages 39–47, 2011.

[41] Leslie Marsh and Christian Onof. Stigmergic epistemology, stigmergic cognition. *Cognitive Systems Research*, 9(1-2):136 – 149, 2008.

[42] J.E. McGrath. *Groups: Interaction and Performance*. Prentice-Hall, 1984.

[43] S Miligram and H Toch. *Hanbook of Social Pyschology*, volume 4, chapter Collective Behavior: Crowds and Social Movements, pages 507–610. 1969.

[44] Wanas Nayer, Amr Magdy, and Heba Ashour. Using automatic keyword extraction to detect off-topic posts in online discussion boards. In *Content Analysis in Web 2.0 Workshop*, 2009.

[45] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

[46] Mari-Sanna Paukkeri and Tanja Kotro. Framework for analyzing and clustering short message database of ideas. In *Proceedings of Knowledge Management and Knowledge Technologies*, 2009.

[47] H. Rangwala and S. Jamali. Defining a coparticipation network using comments on Digg. *IEEE Intelligent Systems*, 25(4):36–45, 2010.

[48] Sara Rosenthal and Kathleen McKeown. Detecting opinionated claims in online discussions. In *2012 IEEE Sixth International Conference on Semantic Computing*, pages 30–37, 2012.

[49] Warren Sack. Conversation map: A content-based usenet newsgroup browser. In *Proceedings of the International Conference on Intelligent User Interfaces*, pages 233–240, 2000.

[50] Dina A. Said and Nayer M. Wanas. Clustering posts in online discussion forum threads. *International Journal of Computer Science and Information Technology*, 3(2):1–14, 2011.

[51] Marjorie E. Shaw. A comparison of individuals and small groups in the rational solution of complex problems. *The American Journal of Psychology*, 44(3):491–504, 1932.

[52] Marc A. Smith. Invisible crowds in cyberspace: Mapping the social structure of the usenet. In Marc A. Smith and Peter Collock, editors, *Communities in Cyberspace*, pages 195–219. Routledge, 1999.

[53] S. O. Sood, S. Owsley, and Birnbaum L Hamoon, K.J. Reasoning through search: A novel approach to sentiment calssification. *Northwestern University Tech Report*, 2007.

[54] S.O. Sood and E.F. Churchill. Anger management: Using sentiment analysis to manage online communities. *Grace Hopper Celebration*, 2010.

[55] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Compututational Linguistics*, 37(2):267–307, 2011.

[56] Guy Theraulaz and Eric Bonbeau. A brief history of stigmergy. *Artificial Life*, 5(2):97–116, 1999.

[57] Bruce W Tuckman. Developmental sequence in small groups. *Psychological bulletin*, 63(6):384, 1965.

[58] Pedro Olmo S. Vaz de Melo, Christos Faloutsos, Renato Assunção, and Antonio Loureiro. The self-feeding process: A unifying model for communication dynamics in the Web. In *Proceedings of the International Conference on World Wide Web*, pages 1319–1330, 2013.

[59] Fernanda B. Viégas and Judith S. Donath. Chat circles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 9–16, 1999.

[60] Franz Wanner, Thomas Ramm, and Daniel A. Keim. Foravis: explorative user forum analysis. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, pages 14:1–14:10, 2011.

[61] Markus Weimer, Iryna Gurevych, and Max Mühlhäuser. Automatically assessing the post quality in online discussions on software. In *Proceedings of the Annual Meeting on Association for Computational Linguistics (Interactive Poster and Demonstration Sessions)*, pages 125–128, 2007.

[62] Howard T. Welser, Eric Gleave, Danyel Fisher, and Marc Smith. Visualizing the signatures of social roles in online discussion groups. *The Journal of Social Structure*, 8(2):1–31, 2007.

[63] Steve Whittaker, Loren Terveen, Will Hill, and Lynn Cherny. The dynamics of mass interaction. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, pages 257–264, 1998.

[64] A. N. Woolley, C. F. Chabris, A. Pentland, N. Hashmi, and T. W. Malone. Evidence for a collective intelligence factor in the performance of human groups. *Science*, 330:686–688, 2010.

[65] Min Wu, Hui Li, Ke Zhang, and Lijuan Qin. An evolutionary model of reply networks on bulletin board system. In *International Conference on Information Technology, Computer Engineering and Management Sciences*, volume 2, pages 92–95, 2011.

[66] Rebecca Xiong and Judith Donath. PeopleGarden: Creating data portraits for users. In *Proceedings of the ACM Symposium on User Interface Software and Technology*, pages 37–44, 1999.

[67] Mingliang Zhu, Weiming Hu, and Ou Wu. Topic detection and tracking for threaded discussion communities. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 77–83, 2008.

Chapter 7

Determinants of Online Persuasion

7.1 Introduction

Our everyday social interactions often involve a complex process of persuasion and influence. Such interactions range from informal settings such as hallway conversations to formal situations like business negotiations and presidential election campaigns. Therefore, the ability to persuade is a key skill to possess and can greatly contribute to our social life. Dating back to the time of Aristotle, there has been a great deal of research on persuasion and persuasive discourse in such traditional communications [23, 7]. However, with the advent and ongoing development and growth of social networking websites, a considerable amount of our social engagement and interaction is taking place through new and advanced forms of communication. The generated communication data, mainly in the form of text messages and comments, provide a great resource for understanding the mechanisms behind online persuasion and its similarities and differences from traditional persuasion processes. Yet, such messages have not been sufficiently leveraged to study persuasive acts in online communication environments, mainly due to the lack of labeled data.

A niche subreddit community offers a substantial resource for studying online persuasion, facilitating opinion change through online comments and requiring the members to annotate the successful comments. This subreddit, called Change My View (CMV), is intended for users who have an opinion on a subject but are willing to listen to different voices and to change their opinion. The users who post their views in the CMV are known as Original Posters (OPs) in the community. After a view is posted by an OP, the other

A version of this chapter has been submitted for publication to *ACM Transactions on Social Computing*.

users provide comments with the aim of changing the OPs' opinions. CMV rules require that an OP mark the comment(s) that successfully changed his/her view and briefly explain why the view is changed. In earlier work on persuasion and text, persuasion is broadly defined as an interactive process through which a given message alters an individuals perspective by changing the knowledge, beliefs, or interests that underlie that perspective [32]. This definition draws on the associated literature in literacy, psychology, and philosophy and is applicable to what CMV users are asked to achieve. Therefore, the CMV subreddit provides a set of comments that are written with the purpose of persuasion. In addition, the CMV rule on marking the comments enables us to quickly develop a persuasion corpus as the comments are already indirectly labeled as persuasive or non-persuasive by the OPs.

By focusing on this corpus, we investigate various dimensions of persuasive comments in comparison with the non-persuasive ones. We first study the text of the comments regarding their relevance to the content of the original post. Due to the importance of timing and order in asynchronous communications within multiple participants, we also take into account the temporal attributes of the comments and their entry order. According to our results, these two features play an essential role in the success of a comment in belief change. Therefore, we first control for these two variables and further analyze the linguistic indicators that are not necessarily adding to the propositional content of the comment but are intended to help the reader organize, interpret, and evaluate the information given. These linguistic features are often referred to as metadiscourse or non-topic material and are known to influence the persuasion process [12, 27].

In this work, many of such linguistic indicators are captured by the LIWC[1] dictionary that provides psychologically meaningful categories of words [45]. In addition, writing quality and the organization of the text are known to facilitate comprehension and recall of the argumentative text, which subsequently contributes towards persuasion [14]. Therefore, we analyzed a set of attributes relevant to the coherence and sophistication of the writings such as the use of transitional phrases, the use of punctuation marks, as well as lexical diversity and overlap in our controlled sets and found differences between persuasive and non-persuasive comments. In addition, given that the text provided from a credible source is more likely to lead to persuasion, we take into account a set of user attributes. The predictive power that each of these features carries is further analyzed.

Finally, as mentioned earlier, CMV requires users to include an explanation as to why and how a comment could change their view. With the aim of gaining insight into the perceived reasons for belief change, we conducted a preliminary analysis of the collected explanations. We first extracted particular sentences from explanations that we deemed

[1]Language Inquiry and Word Count: `http://liwc.wpengine.com/`

relevant to be the subject of further analysis. These sentences are then grammatically parsed to extract certain clues that can reveal which characteristics of the comment are perceived as persuasive. The explanation analysis results are linked to earlier analysis of the comments to understand the differences and similarities between the perceived and actual persuasive cues in online comments.

In the remaining sections, we first review the related studies (Section 7.2) and then describe our dataset (Section 7.3). We then present our data analysis for understanding the linguistic properties and user characteristics that are specific to persuasive comments, along with the prediction models to identify persuasive comments (Section 7.4). The study of user explanations are then presented (Section 7.5). Next, we discuss the finding and limitations of the study (Section 7.6) and conclude with a summary of the work and our future research plans (Section 7.7).

7.2 Related Work

A few research directions are related to this study: works on the detection and analysis of influential users in social media, studies focused on request fulfillment on the Web, research on persuasive essays and essay scoring, and finally attempts to analyze and find persuasive text in online posts and comments.

Influence in Social Media. Influential users are expected to persuade more often compared to the non-influential ones. Thus the unique attributes in their texts can offer insights on the linguistic indicators of persuasive texts. Quercia et al. [38] studied the linguistic attributes of users' tweets and found that influential users tend to be individuals who express the negative sentiment in part of their tweets.

Biran et al. [5] hypothesized that influencers are more likely to engage in certain conversational behaviours such as persuasion, agreements, and disagreements. To detect influencers in conversations in LiveJournal and Wikipedia, they first built a feature set based on the human annotations about these three conversational behaviours and the dialogue pattern of the discussions. They then developed classifiers to detect the influencers in these two environments. Their best classifier resulted in an F-score of 59.3% for Wikipedia discussions and an F-score of 74.3% for LiveJournal weblogs.

Web-based Request Fulfillment. Another related line of research is focused on the analysis of online requests. Request fulfillment is central to many social platforms such as Q&A websites (e.g., *StackOverflow.com*) and crowdfunding and philanthropy communities (e.g., *Kickstarter.com*). For such requests to be fulfilled by other participants, the requester commonly needs to employ persuasion strategies, convincing others to help. In a study

on a crowdfunding platform [31], the linguistic characteristics that led the crowd to fund specific projects are explored. By analyzing 9M phrases extracted from *Kickstarter.com*, they found that the language used to describe projects has predictive power, accounting for roughly 58% of the variance around successful funding. They also analyzed 59 other features commonly present in crowdfunding websites, such as the existence of a video, the number of comments, and the number of updates, and evaluated their predictive power.

Websites like *Kickstarter.com* provide incentives to promote contribution rate. Another set of platforms allows for altruistic requests. Althoff et al. [3] investigated what motivates people to give when they receive no tangible reward in return. By examining the request text, they found that providing a narrative that clearly communicates the needs is an essential factor for the success of requests. In addition, providing indications of gratitude as well as reciprocity enhanced the chance of success. Finally, the high status of the asker was found to be related to the request fulfillment process. This feature set is enhanced in [22] by exploring the centrality and role of the requester as well as the topical aspects of the request. Centrality characterizes the extent of user interactions in the community, and the role attribute reflects the role of user interactions in the network communities. By adding these features to the one proposed in [3], they improved the classification performance for predicting the success of an unseen request.

Essay Scoring and Evaluation. Automated essay evaluation is one of the main applications of natural language processing research in education and often includes techniques to assess and judge the quality of the text. Essay scoring studies have analyzed written text from a variety of dimensions such as prompt adherence [35], coherence and sophistication of the language [36, 13], technical quality and relevance [34], as well as argument strength and quality [36, 43]. For instance, Persing and Ng [36] proposed a rich set of features to study and score argumentative essays. This feature set includes a POS n-gram feature as a syntactic generalization of word n-grams as well as semantic frames of the text as a semantic generalization. In addition, they examined the use of transitional phrases (e.g., furthermore and likewise) and coreferences as an indication of coherence. Argument-related features such as prompt agreement and argument errors are also studied.

Persuasion in Social Media. There has been limited work focusing on the automatic detection of persuasive acts in social media. Anand et al. [4] presented a corpus of blog posts that are marked with the attempts to persuade in online interactions. They developed a classification of these attempts based on their preliminary examination of the data and the literature. This classification included Cialdini's [11] six principles of influence, Marwell and Schmitt's [29] twelve strategy types for securing behavioral compliance, and some of the Walton, Reed, and Macagno's [47] argumentation schemes, specifically, arguments

from causal reasoning, arguments from absurdity, and arguments from example.

The closest to our work is the recent study conducted by Tan et al. [44] that is focused on the analysis and detection of persuasive comments from CMV. In CMV, users can comment on the original post or form a discussion thread by replying to others' comments. By studying these discussion threads, the authors found interaction dynamics that can affect the persuasion process, such as user entry-order and the degree of back and forth exchange. In addition, they examined the language interplay between the OP and the comments by investigating the lexical overlap and similarity of the two texts. Their attempt to predict which of the two comments will succeed in belief change resulted in an accuracy of around 65% for the root reply and 70% when considering the full discussion path. In addition, they explored a set of linguistic features from the original posts and found a small set of attributes that may signal whether the OP is malleable or resistant to persuasion.

Our work is informed by these research directions and draws from social psychology literature on belief change and persuasion. Different from Tan et al. [44]'s work, we focused exclusively on the root comments but conducted a more comprehensive and in-depth study of such comments. Our experimental control is more robust as we include the temporal aspect of the comments besides the relevance to the original post. Except for several linguistic features that are in common between the two works, we introduce a new set of variables that explore the psychological aspects of the language as well as the readability, cohesion, and sophistication of the text. In addition, our methods for measuring certain attributes (e.g., content relevance and similarity) are different from theirs. Our prediction model improves the performance of their classifier in identifying the persuasive comments by 10%. Finally, we have studied user explanations to gain insight into the persuasive factors perceived by the users themselves. We then compared the perceived reasons to the features found in the direct analysis and comparison of the persuasive and non-persuasive comments.

7.3 Reddit Dataset

Change My View (CMV) is an active subreddit with over 23K subscribers. CMV users provide their beliefs on a subject, are open to different views, and will change their beliefs if convinced by the others' comments. According to the CMV rules, the original post needs to explain the reasoning behind the view, not just what the view is. Therefore, the posts are required to be at least 500 characters. In addition, CMV rules clearly indicate that the OP "must personally hold the view and be willing to have it changed".

Once a view is posted to CMV, other users can provide arguments and reason against

the OP's initial view reflected in that post. The poster can then choose to accept the comments' assertion and change his/her belief, ignore them, or engage in a discussion in the form of conversation threads and defer action until he/she is certain about the view. When a comment successfully changes the poster's view, he/she is required to reply to that comment. This reply must include the delta (Δ) character that indicates the belief change and should provide an explanation of why and how the comment has changed his/her view. CMV can function as a controlled experimental setting for persuasion studies as it controls for a set of persuasion-related variables such as the belief that is intended to change and the attributes of the user who holds the belief. Specifically, since this platform controls for the *personal relevance* of the belief that has shown to be greatly influential in the persuasion process [48]. Meanwhile, CMV allows variations in another set of variables including the attributes of the users who post the comments and different types of language and writing styles used in the comments. CMV is heavily moderated, controlling for any posts or comments that are not following the rules. The CMV's delta mechanisms along with its strict rules provide a unique and a valuable dataset for studying online persuasion.

Figure 7.1 shows an example post submitted to CMV, along with three selected comments that are provided to change the original view. The top section of the first block shows the title of the post along with the time of submission and the OP's username. The bottom part is the OP's view and reasoning. The next three blocks with a solid border are comments provided to change the view. As can be seen, each comment is associated with the commenter's username and overall delta score (i.e., the number of times this user has received a delta so far), and the comment score (i.e., the number of upvotes deducted by the number of downvotes the comment has received so far), and the timestamp of the comment submission. Finally, the block with the dotted border is a reply to the first comment, with the delta (Δ) character and the rationale for giving the delta (Δ) to the comment. It should be noted that the metadata presented in the example and their placement is the same as the actual CMV interface.

As mentioned, CMV comments can occur in the form of threaded discussions, where CMV users can reply to other comments and build a discussion tree. A delta can be awarded to comments at different levels of the tree. In our study, however, we have only focused on the top-level comments, which are the direct replies to the post and so serve as the root comment in each thread. By focusing only on the root comments, we can ensure that the persuasive aspects lie within the comment itself but not the conversational context that takes place through a comment thread. Given that we are focused on the root comments, from now on, whenever we say comment, we refer to a root comment. In addition, by persuasive or successful comments we refer to the comments that have won a delta, whereas the non-

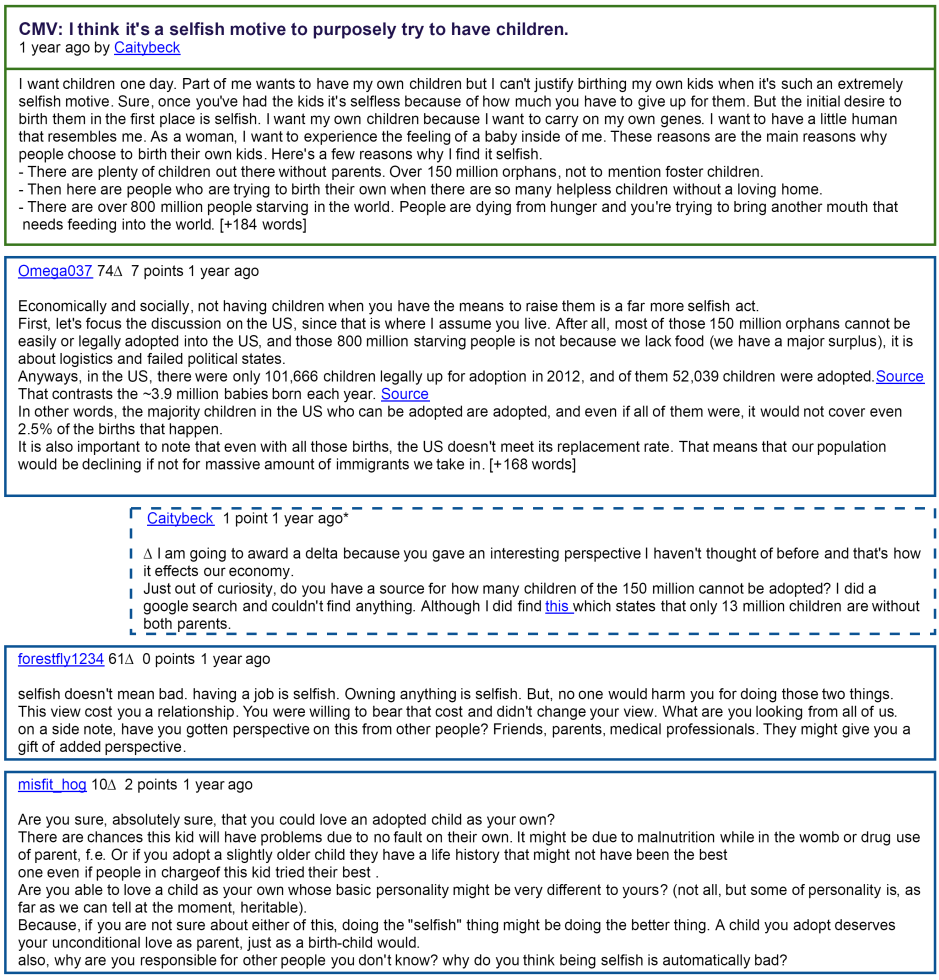


Figure 7.1: Example of a CMV post, along with the comments provided to change the view.

Entity	Total	Unique Users
post	4,853	3,592
root comment	58,242	13,920
persuasive root comment (Δ)	1,771	1,019
persuaded post (Δ)	1,272	1,108

Table 7.1: Data statistics

persuasive ones are the ones that have not received a delta.

From February 2015 to January 2016, we collected posts and their comments from CMV. In Reddit, when a user or a textual content of a comment or post is no longer available, the corresponding data field is assigned the value of *[deleted]* or *[removed]*. We report the data statistics after the cleaning of such data points. Also, only a small portion (less than 5%) of the attempts to persuade through root comments are perceived persuasive and thus are given a delta. To reliably compare the persuasive and non-persuasive comments, we need to control a set of variables. One of the most important factors to take care of is the post that the comments are provided for, since the topic of the post and the OP's characteristics may affect the persuasion process. Therefore, we only focused on the posts that are associated with at least one successful comment. In addition, for the sake of comparison, we filtered this set to only include the posts that have at least 5 comments. This filtering process resulted in 1,249 posts and 18,534 comments associated with these posts, among which 1690 comments have received a delta. The statistics of our finalized data set is summarized in Table 7.1.

7.4 Persuasive Comments: How are they Different?

7.4.1 Relevance to the Original Post

One of the main metrics we deemed relevant is the degree to which the comments are on-topic and related to the post provided by the OP. To calculate the relevance of the comments to the main post, we employed the standard information retrieval technique TF-IDF, which is widely used to retrieve and rank documents given a query. TF-IDF determines the importance of a given query word to a document based on the frequency of the word in the document (TF) as well as the distribution of the word in the underlying corpus (IDF). Based on TF-IDF, the words that appear more often in the document but rarely appear in the corpus are assigned the larger values.

To determine the relevance of a comment to a post, we first conducted all the necessary

cleaning and pre-processing steps on both posts and comments, such as conversion to lowercase, emoticon replacement, and elision control. Next, we treated the collection of posts as the underlying corpus and each comment as a query. The relevance of a comment to a post (i.e., a document) is then calculated as the sum of the tf-idf scores of the words in the comment [41].

Since the raw tf-idf score of comments depends on the features of the main post, a direct comparison of the relevance scores may not be a reliable approach. Instead, we separately analyzed each set of comments provided for a post and studied how the persuasive comments ranked compared to the non-persuasive one based on their tf-idf score. Focusing on the raw rank of the comments will result in a bias for comment sets of different sizes. Instead, we calculate the percentage of the non-persuasive comments that have a higher score than the persuasive ones using the following:

$$Relevance\ Grade = \frac{Rank_{\Delta} - 1}{N} * 100 \tag{7.1}$$

$Rank_{\Delta}$ is the raw rank of a persuasive comment when the comments are sorted from the highest to the lowest tf-idf score. N is the total number of comments in the comment set. For instance, consider a post with 20 root comments. If a persuasive comment has the second highest tf-idf score in the set, then it will be given a relevance grade of 5%. This metric implies that 5% of the comments are placed above the persuasive comment regarding their tf-idf score. Figure 7.2 shows the distribution of the relevance grade for different persuasive comments. The plot clearly depicts a descending pattern, indicating that the most relevant comments in the set have a higher chance of receiving a delta. The vertical red line shows the mean value and implies that, on average, the persuasive comments are within the top 25% of the comments according to their relevance. One may argue that tf-idf results are biased and favor longer comments. While this statement is generally true, for our dataset there is no correlation between the relevance grade of a comment and the word count (Pearson r=0.002). However, it is worth noting that the raw tf-idf score is correlated with the word count (Pearson r=0.65). This shows another advantage of using the relevance grade instead of the raw rank or the raw tf-idf score.

To analyze the interplay of the comments and the posts, Tan et al. [44] employed similarity metrics based on word overlap. As a result, they found that comments that are dissimilar to the original post in terms of the content word usage have a higher chance of success. However, this metric overlooks the fact that some words might be common in this particular community and should not be given the same value in case of an overlap com-

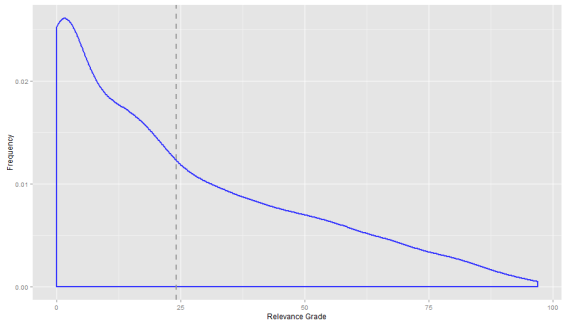


Figure 7.2: Density of relevance grade distribution for persuasive comments.

pared to the rare words. For instance, based on our IDF calculation, the words *people* and *believe* appear frequently in CMV posts in general, thus their overlap may not necessarily indicate a strong topical similarity. Even though general dissimilarity may have a positive effect on persuasion, our results indicate that it is vital to use words that are specific to the view of focus. This finding is in line with earlier research on persuasion that states that matching the content of a persuasive message to the functional basis of people's attitudes enhances the chances of persuasion [37].

7.4.2 Timing and Order

Similar to the majority of Q&A and news sharing social platforms, Reddit primarily facilitates asynchronous communications. Therefore, timing and entry-order may influence the persuasion process. We first examined the time lag between the submission of the main post and the submission of the comments. The results indicate that the majority of the comments are provided fairly quickly once the main view is posted, with over 90% of the comments provided within a day after the submission of the post. In other words, timing is similar for all the comments in a thread.

Therefore, instead of the time lag, we analyzed the entry order of the comments. Given a post, we calculated a temporal grade for the persuasive comments in the same manner as the relevance grade. Each persuasive comment is given a value which is the percentage of the non-persuasive comments that have been submitted to the post prior to the comment of focus. Figure 7.3 shows the distribution of the temporal grade across persuasive comments. Even though the frequency drops as the temporal grade increases, the change is neither as steady nor as sharp as the relevance grade. In particular, there is a relatively significant rise as we go from the first 10% of the comments to the ones that are between 10-20%. However, overall, the grade frequency drops as the temporal grade increases, indicating that a comment that is submitted after a larger set of comments have a lower chance for

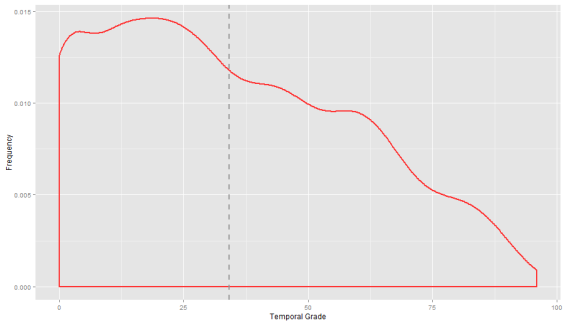


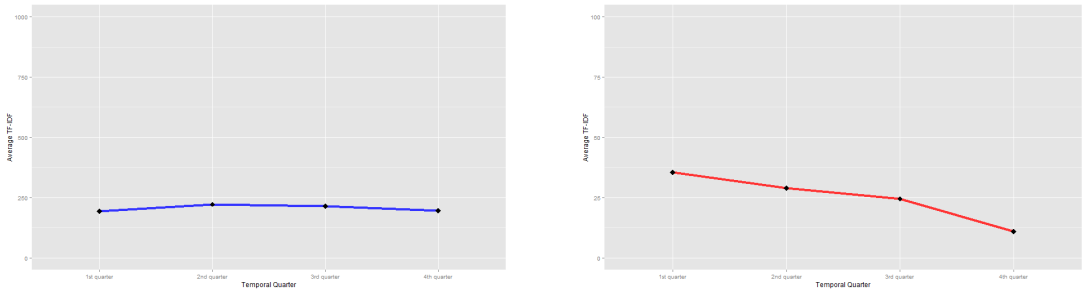
Figure 7.3: Density of temporal grade distribution for persuasive comments.

receiving a delta. The mean value for the temporal grade is 34% as shown by the vertical red line.

One may argue that this trend is due to a possible change in relevance as new comments are submitted. As more users comment on a post, a topic shift may occur due to the potential influence of earlier comments. To test this hypothesis, we quartered all of our comment sets based on their temporal order and calculated the average TF-IDF score in each quarter. As Figure 7.4 (a) shows, the changes in relevance is subtle, defeating the argument that relevance may be a latent variable affecting our temporal results. However, as Figure 7.4 (b) implies, the percentage of comments that are perceived persuasive steadily decreases as we move from one quarter to the next. A few possible reasons may cause this effect. For instance, this change may be attributed to the *backfire effect* [33], which suggests that beliefs may actually get stronger in the face of contradicting evidence. An increase in the number of contradictory comments and arguments might increase the likelihood of the backfire effect. Further investigations are required to validate this claim in the context of CMV. In addition, users may lose faith in the ability of the community after seeing multiple unsuccessful comments and thus may not follow the upcoming comments as closely. Finally, those who have already changed their views by an earlier comment might be less motivated or even possibly reluctant to monitor the forthcoming comments.

7.4.3 Psychological Attributes of the Language

According to the earlier analysis, both relevance and temporal aspects of comments can affect whether they are perceived as persuasive. Therefore, to analyze the potential differences in the use of off-topic linguistic indicators and writing quality of the comments, it is imperative to control for these two variables first. For each persuasive comment, we choose a non-persuasive comment that 1) is the most similar to the persuasive one with respect to the relevance score, and 2) is preceding the persuasive one. By adding the tem-



(a) (b)

Figure 7.4: Even though the average relevance score in each temporal quarter is similar (a), the percentage of persuasive comments steadily decrease as we move from one quarter to the next (b).

poral constraint, we are ensuring that the non-persuasive ones were seen by the OP but failed to change his/her belief. By disregarding this aspect, a comment may not receive a delta simply because it may have been missed by the OP. When the persuasive message is the earliest comment, we choose the most similar comment in terms of relevance from the next three comments that have been submitted. Following this method led to the selection of a set of non-persuasive comments of the same size as the persuasive ones, enabling further comparison and analysis of the root comments.

We used LIWC as one of our linguistic analysis tools because it is capable of providing a broad range of social and psychological insights from the language. The LIWC program has two central components, namely the processing component and the dictionaries. The processing component processes a text file word by word. Each word is then compared against the dictionaries to determine which LIWC category the word belongs to. After processing all the words in the text, LIWC calculates the percentage of each LIWC category. LIWC categories belong to four main processes: linguistic processes (e.g., prepositions, pronouns, and auxiliary verbs), psychological processes (e.g., family, anxiety, and health), personal concerns (e.g., death, work, and achievement), and spoken categories (e.g., assents, nonfluencies, and filters). The definition and examples of each of these categories can be found at the LIWC website[2].

In addition, a set of LIWC variables is measured independent of the dictionaries. The use of a variety of punctuation marks is examined in the text by calculating their usage percentage. Besides these, four non-transparent language variables (analytical thinking, clout, authenticity, and emotional tone) and general descriptive variables (words per sentence and percent of words that are longer than six letters) are taken into account. While the majority

[2] http://www.liwc.net/descriptiontable1.php

of the LIWC variables are analyzed here, the ones that can contribute greatly towards writing quality (e.g., punctuation marks) are included in the next section, along with a newly introduced set of features.

To use the LIWC tool, we treated each comment in the collection as a separate input. We then analyzed each LIWC category using a two-tailed two sample t-test. The attributes and features for which the differences are statistically significant are presented in Table 7.2. The difference is also practically significant for the variables that are marked with an asterisk (Cohen's $d > 0.2$).

According to the LIWC manual, a high value for the analytical thinking variable reflects formal, logical, and hierarchical thinking, while the lower numbers reflect more informal, personal, and narrative thinking. The authentic category captures the degree to which the language is more honest, personal, and disclosing. Aristotle's Rhetoric describes three modes of persuasion: logos, pathos, and ethos. Logos is the logical appeal of the persuasive act that can be characterized by the LIWC analytical thinking score. As expected, the persuasive comments are scored higher on analytical thinking (both statistically and practically significant), confirming that users often value arguments that are based on facts, logic, and analytics. Pathos refers to the persuasive acts that are intended to stir the feelings and emotions of the audience. According to the analysis, persuasive comments are scored lower on authenticity, indicating that logos is valued more in online communications compared to pathos. Ethos is an appeal to the authority or credibility of the presenter and is discussed in Section 7.4.5.

Emotional tone captures the sentiment level of the language. A high number is associated with a more positive style, whereas a low number reveals greater anxiety, sadness, or hostility. A number around 50 suggests either a lack of emotionality or different levels of ambivalence. Even though the emotional tone is lower for persuasive comments, the average values for both comment sets are close to 50, making it difficult to interpret the results. In addition, no difference was observed regarding the use of positive and negative words captured by the LIWC dictionary. Therefore, to draw reliable conclusions regarding the sentiment-related features and persuasion in online text, more sophisticated sentiment analysis tools may be required. Earlier persuasion work states that messages that match the emotional state of the receiver have a higher chance of persuasion [15]. Therefore, instead of sentiment analysis of the comments in isolation, studying the sentiment interplay of the main post and the comments may give us insight into the potential links of persuasion and sentiment in online communities.

While we analyzed the content of comments based on their relevance to the main post, LIWC linguistic processes allow the inspection of function words, also known as style

Summary Variables	Welch T-Test P-Value
Analytical thinking*	+++
Authentic	- -
Emotional tone	- -
Linguistic processes	
Function words	- - -
Total pronouns	- - -
Personal pronouns	- -
1st person singular (I)	- -
2nd person (you)	- -
Impersonal pronouns	- -
Prepositions	- -
Auxiliary verbs*	- -
Negations	- -
Interrogatives	- -
Numbers	+++
Psychological processes	
Cognitive processes	- - -
Cause	- -
Discrepancies	- - -
Tentativeness	- -
Differentiation	- - -
Perpetual processes	- -
Time Orientation	
Past focus	++
Present focus	- - -

Table 7.2: Welch T-Test for the two comment groups. ++ for $P < 0.05$ and +++ for $P < 0.005$ when the values are greater for persuasive comments, whereas - - and - - - are used when the values are smaller for persuasive comments. The variables with a small effect size (Cohen's d > 0.2) are marked with an asterisk.

words, in the text. From a psychological and social perspective, style words reflect *how* people are communicating, whereas content words convey what *they* are saying. Therefore, function words are much more closely linked to measures of writers' and readers' social and psychological worlds [45]. We found statistically significant differences between the two comment groups in the use of function words. The persuasive comments, in particular, include fewer pronouns (both personal and impersonal), prepositions, and auxiliary verbs (see Table 7.2).

Negations appear more frequently in the comments that failed to persuade. According to earlier psycholinguistic research, negative grammatical transformations are more complicated than the positive ones, and they often require a longer time and more cognitive resources to process [19]. Interrogatives, often used when asking questions, are also observed more often in non-persuasive messages. Numbers, on the other hand, appear more frequently in persuasive comments, which can be associated with referrals to quantified evidence and statistics. Using quantification instead of descriptive modifiers is thought to provide integrity to a communication due to the credibility associated with numbers [48]. Even though some earlier studies showed negative effects of quantification on persuasion when the source is of low credibility (mainly advertisements) [21], quantification is positively associated with persuasion in CMV, wherein source credibility is largely unknown.

The words representing cognitive processing seem to appear less in persuasive comments. In particular, the words related to the discrepancy (e.g., should and would) and the tentative (e.g., maybe and perhaps) categories are less frequently used in successful comments. Findings related to these two categories can be explained by earlier research on persuasion from two different perspectives. First, these two categories show hesitation in language, which is associated with a powerless style of communication. Powerless language can adversely affect reader's judgments of author credibility and subsequently affect the persuasion process or may have a direct negative impact on persuasion [21, 42]. Secondly, persuasion can be influenced by how clear or vague the message is presented. Some works found vague language to be adversely associated with persuading others, while some found strategic ambiguity to enhance the chances of persuasion [21]. The reason behind the latter finding is that the receiver of the message could not as easily reject vague messages. This reason, however, is not applicable in the context of CMV, wherein the OP personally holds the view, and the purpose is to change that view. Therefore, the ambiguity of the message is expected to have a negative influence in CMV. Differentiation-related words (e.g., without and despite) are also more frequent in unsuccessful comments, which may be related to the added syntactic complexity of using negative grammar. In addition to the potential effects of syntactic complexity, the rhetoric strategies associated with these words

can greatly influence the persuasion process.

Similarly, perceptual processes (i.e., processes related to seeing, hearing, and feeling) are used less often in persuasive comments. This finding could also be linked to the dominant power of logos compared to pathos in online platforms since perceptual processes are often linked with the description of personal experiences and narratives compared to analytical and logic-based text. Finally, past tense is used more often in persuasive comments, whereas present tense is used more commonly in non-persuasive ones. Our observation of the comments indicates that the past tense is used mainly when providing evidence and examples from the past. Some instances of the parts of comments that include such referrals are provided below:

- "All told, 60,000,000 people (at least) died in World War II. There was significant evidence at the time that an invasion of Japan would have tacked on another 10,000,000 to 15,000,000."

- "First off, it's ok to use an old symbol if it isn't used anymore. The Nazis could have used the swastika without it being cultural appropriation if the Hindus still weren't using it. "

Even though the presence of evidence and factual information can directly lead to opinion change, providing such historical cases can also contribute towards verbal imagery. The ability of words to elicit images in readers have shown to positively influence persuasion compared to abstract language [39, 8]. However, the present tense can also offer verbal imagery. Further experiments are required to confirm that it is indeed the verbal imagery aspect of these historical case that leads to persuasion.

7.4.4 Writing Sophistication and Comprehensibility

The credibility of the source is known to be one of the main influential factors in the persuasion process [32]. Variations in language can affect readers' judgments of the author credibility and attractiveness [21]. Therefore, we analyze the comments in terms of the sophistication of their language. Language sophistication can be measured based on the length of the text [13], lexical difficulty and diversity [13, 20], as well as the presence of spoken and informal linguistic elements. In addition, a well-written text enhances recall and comprehension, allowing readers to process the text deeply and increasing the chances of understanding and persuasion [17]. Proper use of punctuation marks and the use of cohesive language can both contribute towards comprehensibility. Cohesion can be characterized by the use of connectives and transitional phrases [13] as well as lexical overlap

between the consecutive sentences [13, 20].

According to the results, persuasive messages tend to be longer and include longer sentences. In addition, they include difficult words (i.e., words with more than six letters) more often. Besides, persuasive comments are more diverse in terms of their lexical choices. We calculated lexical diversity by dividing the number of unique words (types) to the total number of words (tokens) in the comment, a metric commonly known as the type-token ratio. A lower diversity metric indicates that the language is relatively redundant, while a higher number indicates a diverse and rich vocabulary usage. Lexical diversity and difficulty are known to influence readers' judgments of the author through a *principle of preference for complexity* [21, 6]. This finding was later confirmed as lexical diversity and sophistication are shown to be directly related to the readers' judgments of the author's competence and socioeconomic status [21, 6]. In addition, informal categories of words including netspeak (e.g., plz and ppl), assent (e.g., yep and okay), and nonfluencies (e.g., mmm and er) are often associated with unsuccessful comments. Not only are these features related to a lack of language sophistication, but the use of nonfluencies also represents hesitation and a powerless style of language, leading to negative effects on persuasion.

We further calculated the Flesch-Kincaid readability score and grade level [30] (see Table 7.3). The Flesch-Kincaid readability score indicates how difficult a reading passage is to understand based on the number of sentences, words, and syllables. Higher scores indicate that the text is easier to read, whereas lower numbers mark passages that are more difficult to read. The Flesch-Kincaid readability grade level is a transformation of the readability score that directly presents a score as the US grade level required to understand the text. One may expect the comments that are easier to read to persuade more often since they are easy to comprehend. On the other hand, complex comments may be associated with the credibility and education level of the author and may lead to persuasion more often. Our readability-related results are in line with the earlier findings on the relations of complex language and the effectiveness of the message. Overall, the results indicate that language complexity and sophistication impacts users' judgments of the credibility, status, and even education level of the author. Such perceptions may be even more essential in online persuasion when little or no information is available about the author, and the language is often the main reflection of author credibility.

One of the main linguistic tools that aids text comprehension and understandability is the use of connectives and transitional phrases. Therefore, we analyzed the frequency of transitional phrases in persuasive versus non-persuasive comments. The set of transitional phrases used is compiled by Study Guides and Strategies [3] and includes 14 different types

[3] http://www.studygs.net/wrtstr6.htm

Sophistication	Welch T-Test P-Value
Word count**	+++
Word per sentence	+++
Words>6 letters	+++
Lexical diversity	++
Netspeak	- -
Assent	- -
Nonfluencies	- -
Readability	
Readability score*	- - -
Readability grade level	+++
Cohesion	
Direction phrases	+++
Emphasis phrases	++
Exemplify phrases	+++
Summarizing phrases	+++
Lexical overlap	- -
Punctuation Marks	
All Punctuation	+++
Commas	++
Colons	+++
Question marks	- - -
Dashes	+++
Parentheses (pairs)	++
Other punctuation*	+++

Table 7.3: Welch T-Test for the two comment groups. ++ for $P < .05$ and +++ for $P < .005$ when the values are greater for persuasive comments, whereas - - and - - - are used when the values are smaller for persuasive comments. * and ** placed beside the variable names indicate a small effect size (Cohen's d>0.2) and a moderate effect size (Cohens'd>0.5), respectively.

of transitional phrases. Our analysis shows that four types of transitional phrases had more occurrences in the persuasive comments than the non-persuasive ones: direction phrases (e.g., beyond and nearly), emphasis phrases (e.g., above all and particularly), exemplify phrases (e.g., including and such as), and summarization phrases (e.g., after all and in conclusion). In addition to the contribution of these phrases toward language cohesion, these words can be associated with a set of rhetorical strategies that influence the persuasion process. For instance, summarizing phrases can aid in the comprehension of the text or exemplify phrases can lead to verbal imagery.

Another metric for assessing language cohesion is the lexical overlap between adjacent sentences. A larger degree of overlap between adjacent sentences indicates a more cohesive language. To calculate lexical overlap between the two sentences, we first tokenized and stemmed nouns found in the two sentences. Then, to account for the length of the sentences, we employed the Jaccard similarity metric between the two noun sets and measured the average similarity across sentences of a comment:

$$\textit{Lexical Overlap} = \frac{1}{N}\sum_{i=1}^{N-1}\frac{|S_i \cap S_{i+1}|}{|S_i \cup S_{i+1}|} \tag{7.2}$$

where N is the total number of sentences in the comments and S_i is the set of stemmed nouns in the ith sentence of the comment. Interestingly, non-persuasive comments exhibit a higher degree of overlap between sentences. This unexpected result could be attributed to the difference in length of the persuasive versus non-persuasive comments. Given that the persuasive ones are longer in general, they may present more arguments from different perspectives, leading to a lower overall lexical overlap. Therefore, we filtered the comment sets to include only the comments that have more than 10 sentences and re-ran the experiment. As expected, no significant difference was observed in the two sets.

Our results also show that persuasive comments tend to have more punctuation marks. Given that parentheses, colons, and dashes are commonly used as linguistic strategies to provide further information related to the core of the text, it is possible that persuasive comments involve more elaborations, explanations, and clarifications, thus making them more convincing. Unlike all the punctuation marks, persuasive comments tend to include fewer question marks. Furthermore, our observation of a subset of comments with question marks shows that what follows the questions seems to differ in the two groups as well. In the persuasive comment group, the person who raised the question often provided explanations and/or answers immediately after the question, whereas in the non-persuasive comment group, often the user provided no further information. For instance, the first item below is a part of the comment that received a delta in Figure 7.1 (this part is hidden in the figure

due to space limitations), while the second one is from an unsuccessful comment (the last sentence of the last comment in Figure 7.1):

- "Now, why is this important? Because our economic and especially our social systems depend on new productive workers to enter the workforce as old ones retire. Without them, the economy would collapse, social programs would go bankrupt, and we would basically experience a long depression until our population rates stabilized or increased."
- "also, why are you responsible for other people you don't know? why do you think being selfish is automatically bad?"

7.4.5 User Delta and Karma Score

Besides the language used in the comments, a user's delta score can serve as an indicator of the users credibility in the community. A user's delta score reflects the total number of times he/she has successfully changed an OPs belief and is displayed in the CMV interface as part of the comments metadata (see Figure 7.1), potentially affecting how others evaluate the comment. We examined the delta scores associated with the users in the two sets of comments and found statistically significant differences ($P < 0.05$).

In addition to the delta score that is specific to CMV, Reddit users are associated with two other metrics related to their history in the Reddit community, namely their link karma and comment karma. Link karma refers to the total number of points that the posts submitted by the focal user have received so far in different subreddits. Points are calculated based on the total number of upvotes deducted by the total number of downvotes. Likewise, comment karma refers to the total points that the comments posted by the focal user have received across different subreddits. Even though the total number of points received by a comment is shown in the CMV interface, link and comment karma information of the author are not directly available (see Figure 7.1). Therefore, we expect these two features to have little influence on the success of a comment. The correlation coefficients of the delta score and link karma (Pearson r=0.002) and the delta score and comment karma (Pearson r=0.11) of the users confirmed our expectation.

7.4.6 Prediction Evaluation

The analysis of persuasive and non-persuasive comments indicates the presence of certain features that are specific to each comment group. To study the predictive power of these features we ran several experiments. For each persuasive comment, we randomly selected

Algorithm	Precision	Recall	F-score
Naive Bayes	0.609	0.779	0.684
LibSVM	0.523	0.996	0.686
Classification via Regression	0.714	0.740	0.727
Ada Boost	0.744	0.679	0.710
Random Forest	0.744	0.760	**0.752**

Table 7.4: The evaluation results of a set of supervised machine learning algorithms for the detection of persuasive comments.

a non-persuasive one from the same line of arguments provided to change a view, leading to a balanced set of comments. We then utilized supervised learning algorithms (using Weka[4]) to predict whether an unseen comment would receive a delta. We evaluated them using 10 fold cross-validation. These models are selected from different categories of algorithms (e.g., Bayesian, tree-based, and regression-based) and are built on 49 features that are related to the relevance of the comments, their entry order, as well as the presence of linguistic indicators that have shown to be effective in opinion change. Table 7.4 presents the results of the classifiers.

The evaluation results of the supervised algorithms are promising as we obtained an F-score of 0.752 using the Random Forest algorithm (as our best classifier) and an average F-score of 0.713 across all classifiers. These results suggest that surface-based attributes can contribute significantly toward predicting the persuasiveness of a message, regardless of the underlying claims and arguments. The accuracy of Random Forest is 0.749, which is about a 10% improvement over the feature set used by Tan et al. [44] for the prediction of successful root replies. However, given that our underlying dataset is different from theirs, the results may not be directly comparable. Using our feature set on their dataset can shed light on the contribution of our newly introduced features and is included in our future research plans.

To gain insight into the predictive power of individual features, we used a feature selection and ranking algorithm. Given that the best performance is achieved by Random Forest, Boruta is employed for such analysis [26]. In the Random Forest algorithm, instances are classified by the votes of multiple decision trees built independently on different bagging samples of the training data. The importance measure of an attribute is the loss of accuracy of classification caused by the random permutation of attribute values. The Boruta algorithm runs in an iterative fashion, wherein the importance of the original attributes is compared with the importance of their randomized copies [26]. By following this ap-

[4] http://www.cs.waikato.ac.nz/ml/weka/

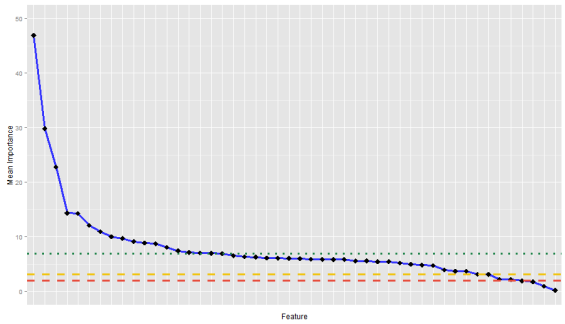


Figure 7.5: The result of the feature analysis.

proach, it is expected that the potential impact of random fluctuations and correlations can be minimized.

The results obtained with the Boruta algorithm are presented in Figure 7.5 in which the x-axis represents the features and the y-axis represents the mean importance of the features assigned by Boruta. Depending on the feature mean importance compared to the importance of all the features, Boruta also decides to either *Reject* or *Confirm* a feature. In addition, the algorithm may not be able to make a clear decision regarding a feature and marks it as *Tentative*. The features below the bottom horizontal line in Figure 7.5 are rejected, indicating that they have no predictive power. These features are the emotional tone of the language, the percentage of internet specific language (known as netspeak in LIWC), as well as the frequency of summarization and emphasis phrases. Also, the three features placed between the midline and the bottom line are the ones that are marked as tentative. These features are the percentage of assents, I pronoun, and the readability grade of the text.

Even though the remainder of the features is confirmed by Boruta, many of them are very similar regarding their mean importance. However, the mean importance starts increasing noticeably after it passes seven (as marked by the top green line). Generally, when the number of classification features is too high relative to the training sample size, the classification performance may decrease since the data set can be undersampled compared to the feature set [24]. Hence, we evaluated the classifiers using the 18 features placed above the top line. Interestingly, the average performance (F-score=0.718) and the performance of Random Forest (F-score=0.763) increased slightly.

The top selected features represent all of the feature categories discussed. For instance, user delta score is placed at the top, serving as an explicit indicator for author credibility. All of the features related to content relevance (i.e., raw TF-IDF, raw relevance rank, and relevance grade) are among the top five features. The temporal grade is also ranked high in the list. Word count, lexical diversity, the count of direction phrases, and the percentage

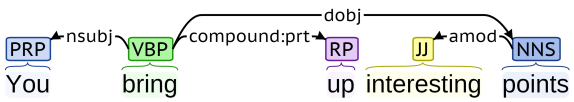


Figure 7.6: An example explanation parsed by a dependency parser.

of parentheses represent writing sophistication and comprehensibility. Also, the percentage of negations, discrepancies, and interrogatives are among the features having possible psychological effects on the belief change process.

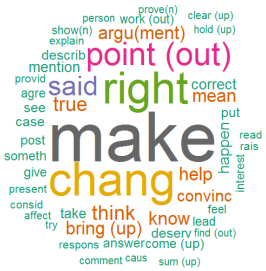
7.5 Persuaded Users: Why are they Persuaded?

The actual reasons leading to belief change and persuasion can be different from the reasons perceived to be persuasive by people [16]. While the analysis of the comment groups helped us understand the actual reasons behind persuasion, the analysis of user explanations can reveal the perceived causes. As discussed earlier, in CMV, those whose belief has changed are required to provide a comment, explaining how and why the other comment changed their view.

These explanations are often short and sometimes include further discussions related to the topic of focus. Therefore, we first identified the relevant sentences to be the subject of further analysis. These sentences are the ones that explicitly refer to the comment that changed their belief and explain how the comment succeeded. Therefore, we extracted the sentences with explicit referrals to the earlier comments (e.g., use of *you*, *comment*, and *argument*), wherein these tokens are the nominal subject of the sentence. The Stanford dependency parser [25] is then used to extract the grammatical structure of the sentences.

The output of the dependency parser is a set of binary relations that are grammatical relations holding between a *governor* and a *dependent*. The governor of the nominal subject is often a verb. However, when the verb is a copula verb, the root of the clause is the complement of the copula verb, which can be an adjective or a noun. Therefore, by the analysis of the governors of the selected tokens, we can gain insight into the perceived reasons and attributes of persuasion in the comments. An example of a selected sentence and its dependency-parsed tree is shown in Figure 7.6.

After the extraction of 2554 relevant sentences from explanations, we analyzed the governors of the subjects and counted their frequency. Before counting, we ensured that the governors are stemmed. In addition, when the governor is a verb, we checked whether it is a compound verb and extracted both words in case it is (e.g., *bring up* in the instance sentence). Figure 7.7 (a) is a tag cloud representation of the 50 most frequent governors



(a) The tag cloud representation of the governors extracted from explanations.

(b) The tag cloud representation of the objects extracted from explanations.

Figure 7.7: The grammatical analysis of the user explanations can be of value in understanding the perceived reasons behind persuasion.

found in the set. In addition, we extracted the objects when such governors are transitive verbs (e.g., *point* in the instance sentence). A tag cloud of the objects is shown in Figure 7.7 (b).

The most frequent words are related to the sentences users provide to describe that a belief change has happened. Instances of these words include *make* and *change* from the governor set and *delta* and *view* from the object set. For example, sentences such as "You change my view" or "You deserve a delta" are among such descriptions. However, some of the extracted words reveal perceived reasons behind persuasion. Some of these reasons are in common with the actual reasons found in the comment analysis, while some can give use new perspectives as to why people decide to change their beliefs. For example, the verbs *present* and *provide*, along with the objects *example* and *source* imply that users value examples and statistics in the persuasion process. This finding is in line with the common presence of numbers, exemplify phrases, as well as past tense in the language of successful comments. In addition, the word sum up is frequent within the governor set, confirming our earlier finding regarding the frequent use of summarizing phrases in persuasive comments. In addition, some of the actual reasons found can be lower-level factors, contributing to what is perceived as the reason. For instance, the proper use of punctuation marks or the lack of powerless linguistic indicators can contribute towards the clarity of the message, which can be associated with the verb *clear up* frequently used in the explanations.

On the other hand, the perceived reasons can uncover certain features of the persuasive text that are difficult or even impossible to infer using computational linguistic tools. For instance, the presence of right, true, *correct*, and *prove* among the governors imply the importance of validity and soundness of the arguments and reasoning in the community. Furthermore, verbs such as *bring up*, *point out*, and *raise* along with objects such as *perspective*, *insight*, and *point* show the value of providing arguments from a new perspective

or providing new pieces of information originally missed by the OP's reasoning in the main post.

7.6 Discussion

An extensive body of research has been devoted to understanding persuasion and persuasive processes in traditional settings. Despite variations evident in the literature, there are certain principles about persuasion that different research communities generally hold in common. Murphy [32] outlines four widely observed features of the text that are known to lead to persuasion. These features include: Argument structure, content, comprehensibility, and credibility. Argument structure is known to be one of the key features of persuasive text [10]. Despite the sound theories proposed over the years [46, 18], automated analysis and discovery of these elements is still a challenging task, and there exist no reliable implementations of such frameworks. Therefore, here, we discuss our findings with respect to the other three factors and highlight the limitations of the study.

The content of arguments has shown to be profoundly effective in the persuasion process [9]. We analyzed comment content in terms of its relevance to the content of the post and found that the most relevant ones have a higher chance of winning a delta. In addition, according to Boruta, our content-related features hold a strong predictive power for persuasive comments. Besides, the examination of the explanations showed the importance of validity of the arguments. Earlier work has shown that the content that provides both sides of an argument and proves the falsity of one tend to be more persuasive [2]. However, the majority of these studies are focused on scientific readings and text. Even though a direct analysis of such refutational text requires sophisticated language processing tools, certain cues found in this study may indicate that this may not be the case for online deliberations. For instance, negations and differentiation words (e.g., although and whereas) are among the linguistic indicators that can be used to reject an argument. However, such indicators are more frequently observed in the non-persuasive comments.

Another factor contributing to the persuasion process is the comprehensibility of text [1, 32]. Comprehensibility refers to the extent to which the text as a whole is easy to read, and it allows the reader to grasp the meaning effortlessly [32]. Some of the linguistic features found in this work such as the use of punctuation marks and the use of transitional phrases can lead to comprehensibility. These findings, therefore, are in line with earlier research on persuasion in text and indicates the importance of comprehensibility in belief change in online environments. In fact, the effect of comprehensibility may be amplified in the context of online communities, where users may be exposed to a large number of

comments in a short amount of time. Therefore, they may selectively pay more attention to the comments that require less effort and cognitive burden to process. Further research on the direct comparison of comprehensibility in traditional and online text forms and its effects on persuasion is warranted to validate this claim.

Finally, the perceived credibility of the communication plays an essential role in the persuasion process. The delta score of the user is the only piece of information that can serve as an explicit indicator of author credibility. As expected, significant differences have been found between the two comment sets regarding the delta score of the author, and this metric has shown to be the most predictive of persuasion. Due to the lack of direct information about the author, the linguistic cues that can signal credibility are shown to be of great value and importance. For example, persuasive comments are scored higher in terms of their readability grade, which can be a potential indicator of the author's education level. The powerless style of language (e.g., use of tentative indicators) is less common in successful comments. Finally, all the features that contribute toward the sophistication of writing (e.g., lexical diversity and difficulty) are more dominant in persuasive comments.

The Reddit API limits the number of posts that can be retrieved from the website to the latest 1000 posts. Therefore, after collecting the posts and their associated comments for over a year, we collected 1690 eligible persuasive comments to be the subject of study. However, to obtain reliable results, larger sets of data may be required. The data collection is currently ongoing and testing the validity of our findings on larger datasets is a part of our future research plans. Besides the data size limitation, we are focused on a single platform with a particular design and audience. Even though CMV provides an appealing environment to study online belief change and persuasion, how a social platform is designed can shape and influence user behavior [28]. Hence, another limitation of the study is the fact that findings may be specific to CMV and might not be generalizable to other online platforms. Therefore, we plan to validate and test our models across different datasets and explore the application of the models in other online deliberation environments.

7.7 Conclusion

Even though a variety of persuasion frameworks and theories have long been established in traditional settings, online persuasion and belief change has received relatively little attention from different research communities. In this study, we mainly focused on the variations of online comments that can lead to persuasion. By analyzing a set of human-annotated comments extracted from Reddit.com, we have identified a set of attributes that are specific to persuasive comments and examined their predictive power for persuasion.

Except for one temporal feature and one attribute associated with the history of users in the community, the remainder of the features are associated with the persuasive impact of various components of the language. These features resulted in a classification of persuasive and non-persuasive comments with a reasonable performance of 75%. Finally, a preliminary examination of user explanations allowed the analysis of the percieved reasons behind their beilef change.

Nearly all variations of the language are important in the persuasion process [21]. Therefore, although we take into account 47 linguistic features, there exist other dimensions of language that need to be studied further. For instance, syntactic varitions, part-of-speech tags, and language intensity levels are among the features we plan to study in the future. In addition to the features of the text and the writer, the characteristics of the reader have shown to be significant in the persuasion literature [17]. One notable attribute is the background knowledge of the reader on the topic of focus. In particular, individuals with more prior knowledge about the topic are less likely to change their beliefs. Saltiel and Woelfel [40] maintained that the strength of a person's attitude depends on the number of incoming messages that are related to that attitude that the person processed. Since the author of the original post and the reader who awards delta are the same in CMV, one may consider the length of the original post as an indicator of the reader's background knowledge and analyze its potential influence on the persuasion process. Analysis of the topical features and their potential influence on the persuasiveness of messages is warranted.

Lastly, persuasion studies have shown that there are various contextual factors that influence the persuasiveness of the act [11]. Such factors include social influence, the task of focus, the comments made by others, and the design of the underlying environments. As a starting point towards a comprehensive computational model of the detection of persuasive messages in online interactions, we have focused on the root comments in the context of belief change in Reddit. We plan to improve our model by taking these contextual factors into consideration.

Bibliography

[1] Patricia A. Alexander and Tamara L. Jetton. The role of importance and interest in the processing of text. *Educational Psychology Review*, 8(1):89–121, 1996.

[2] Mike Allen. Meta-analysis comparing the persuasiveness of onesided and twosided messages. *Western Journal of Speech Communication*, 55(4):390–404, 1991.

[3] Tim Althoff, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. How to ask for a favor: A case study on the success of altruistic requests. In *Proceedings of the AAAI Conference on Web and Social Media*, 2014.

[4] Pranav Anand, Joseph King, Jordan Boyd-Graber, Earl Wagner, Craig Martell, Doug Oard, and Philip Resnik. Believe me-we can do this! annotating persuasive acts in blog text. In *The AAAI Workshop on Computational Models of Natural Argument*, 2011.

[5] Or Biran, Sara Rosenthal, Jacob Andreas, Kathleen McKeown, and Owen Rambow. Detecting influencers in written online conversations. In *Proceedings of the Second Workshop on Language in Social Media*, pages 37–45, 2012.

[6] James J. Bradac, Roger J. Desmond, and Johnny I. Murdock. Diversity and density: Lexically determined evaluative and informational consequences of linguistic complexity. *Communication Monographs*, 44(4):273–283, 1977.

[7] Michelle M. Buehl, Patricia A. Alexander, P. Karen , and Christopher T. Sperl. Profiling persuasion: The role of beliefs, knowledge, and interest in the processing of persuasive texts that vary by argument structure. *Journal of Literacy Research*, 33(2):269–301, 2001.

[8] Alvin C. Burns, Abhijit Biswas, and Laurie A. Babin. The operation of visual imagery as a mediator of advertising effects. *Journal of Advertising*, 22(2):71–85, 1993.

[9] Marilyn J. Chambliss. Text cues and strategies successful readers use to construct the gist of lengthy written arguments. *Reading Research Quarterly*, 30(4):778–807, 1995.

[10] Marilyn J. Chambliss and Ruth Garner. Do adults change their minds after reading persuasive text? *Written Communication*, 13(3):291–313, 1996.

[11] Robert Cialdini. *Influence: The Psychology of Persuasion*. New York, NY: Collins, 2007.

[12] Avon Crismore, Raija Markkanen, and Margaret S. Steffensen. Metadiscourse in persuasive writing:A study of texts written by American and Finnish university students. *Written Communication*, 10(1):39–71, 1993.

[13] Scott A. Crossley and Danielle S. McNamara. Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, 35(2):115–135, 2012.

[14] Marion Crowhurst. Interrelationships between reading and writing persuasive discourse. *Research in the Teaching of English*, 25(3):314–338, 1991.

[15] David DeSteno, Richard E. Petty, Derek D. Rucker, Duane T. Wegener, and Julia Braverman. Discrete emotions and persuasion: The role of emotion-induced expectancies. *Journal of Personality and Social Psychology*, 86(1):314–338, 2004.

[16] James Price Dillard, Lijiang Shen, and Renata Grillova Vail. Does perceived message effectiveness cause persuasion or vice versa? 17 consistent answers. *Human Communication Research*, 33(4):467–488, 2007.

[17] J.A Dole and G.M Sinatra. Reconceptualizing change in the cognitive construction of knowledge. *Educational Psychologist*, 33:109128, 1998.

[18] Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357, 1995.

[19] J.B. Gleason and N.B. Ratner. *Psycholinguistics*. Harcourt Brace College Publishers, 1998.

[20] ArthurC. Graesser, DanielleS. McNamara, MaxM. Louwerse, and Zhiqiang Cai. CohMetrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2):193–202, 2004.

[21] Lawrence A Hosman. *The persuasion handbook: Developments in theory and practice*, chapter Language and persuasion, pages 371–390. 2002.

[22] Hsun-Ping Hsieh, Rui Yan, and Cheng-Te Li. *Advances in Knowledge Discovery and Data Mining*, chapter Will I Win Your Favor? Predicting the Success of Altruistic Requests, pages 177–188. 2016.

[23] James Jaccard. Conversation as a resource for influence: evidence for prototypical arguments and social identification processes. *Journal of Personality and Social Psychology*, 40(2):260–269, 1981.

[24] Eamonn Keogh and Abdullah Mueen. *Curse of Dimensionality*, pages 257–258. Springer US, 2010.

[25] Dan Klein and Christopher D. Manning. Accurate unlexicalized parsing. In *Proceedings of the Annual Meeting on Association for Computational Linguistics*, pages 423–430, 2003.

[26] Miron B. Kursa, Aleksander Jankowski, and Witold R. Rudnicki. Boruta - a system for feature selection. *Fundamenta Informaticae*, 101(4):271–285, 2010.

[27] Bernadette Longo. The role of metadiscourse in persuasion. *Technical Communication*, 41(2):348–352, 1994.

[28] Momin Malik and Jürgen Pfeffer. Identifying platform effects in social media data. In *Proceedings of the AAAI Conference on Web and Social Media*, pages 241–249, 2016.

[29] Gerald Marwell and David R. Schmitt. Dimensions of compliance gaining behavior: An empirical analysis. *Sociomety*, 30:350–364, 1967.

[30] G. M. McClure. Readability formulas: Useful or useless? *IEEE Transactions on Professional Communication*, PC-30(1):12–15, 1987.

[31] Tanushree Mitra and Eric Gilbert. The language that gets people to give: Phrases that predict success on kickstarter. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 49–61, 2014.

[32] P. Karen Murphy. What makes a text persuasive? comparing students and experts conceptions of persuasiveness. *International Journal of Educational Research*, 35(7-8):675 – 698, 2001.

[33] Brendan Nyhan and Jason Reifler. When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330, 2010.

[34] Isaac Persing and Vincent Ng. Modeling thesis clarity in student essays. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 260–269, 2013.

[35] Isaac Persing and Vincent Ng. Modeling prompt adherence in student essays. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1534–1543, 2014.

[36] Isaac Persing and Vincent Ng. Modeling argument strength in student essays. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 543–552, 2015.

[37] Richard E. Petty and Duane T. Wegener. Matching versus mismatching attitude functions: Implications for scrutiny of persuasive messages. *Personality and Social Psychology Bulletin*, 24(3):227–240, 1998.

[38] Daniele Quercia, Jonathan Ellis, Licia Capra, and Jon Crowcroft. In the mood for being influential on twitter. In *Proceedings of the IEEE SocialCom*, pages 307–314, 2011.

[39] John R. Rossiter and Larry Percy. Visual imaging ability as a mediator of advertising response. *Advances in Consumer Research*, 50:621–629, 1978.

[40] John Saltiel and Joseph Woelfel. Inertia in cognitive processes: The role of accumulated information in attitude change. *Human Communication Research*, 1(4):333–344, 1975.

[41] S.O. Sood and E.F. Churchill. Anger management: Using sentiment analysis to manage online communities. *Grace Hopper Celebration*, 2010.

[42] John R. Sparks, Charles S. Areni, and K. Chris Cox. An investigation of the effects of language style and communication modality on persuasion. *Communication Monographs*, 65(2):108–125, 1998.

[43] Christian Stab and Iryna Gurevych. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the Conference on Emprical Methods in Natural Language Processing*, pages 46–56, 2014.

[44] Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the International Conference on World Wide Web*, pages 613–624, 2016.

[45] Yla R. Tausczik and James W. Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54, 2010.

[46] Stephen Toulmin. *The Uses of Argument*. Cambridge University Press, 1958.

[47] Doug Walton Walton, Chris Reed, and Fabrizio Macagno. *Argumentation Schemes*. Cambridge University Press, 2008.

[48] Richard F. Yalch and Rebecca Elmore-Yalch. The effect of numbers on the route to persuasion. *Journal of Consumer Research*, 11(1):522–527, 1984.

Chapter 8

RST Cue Extraction and Analysis

8.1 Introduction

Rhetorical relations, also known as discourse relations and coherence relations, are paratactic or hypotactic relations that hold between spans of text, explaining the construction of coherence in discourse [17]. For years, researchers have been focused on building robust theoretical foundations and taxonomies for rhetorical relations [8, 9, 10, 12]. However, automatic identification of such relations among text spans has remained a difficult and a challenging task.

Facilitating the automatic identification of rhetorical relations can contribute to various Natural Language Processing (NLP) tasks, most notably text summarization, text generation, and essay scoring [19]. Therefore, it is necessary to conduct thorough corpus-based studies in order to derive knowledge regarding the feature sets that are of value in detection of both explicit and implicit relations.

Explicit relations are the ones that are signaled by cues, while no cue is present in implicit relations. In general, various kinds of cues can signal the existence of a relation, including lexical cues, mood, modality, and intonation [17]. In this study, we are focused on explicit relations in written text that are signaled by the presence of lexical cues. We aim to gain insight into the ability of such cues in the automatic detection of rhetorical relations. We also hope to understand whether the effectiveness of cue-based approaches varies based on the nature of the relation and the underlying text genre.

A version of this chapter has been published in the *proceedings of the IEEE International Conference on Semantic Computing (ICSC).*

The present study is part of a larger effort to identify rationales in the text of discourse. Given that the three rhetorical relations of CIRCUMSTANCE, EVALUATION, and ELABORATION are commonly present in rationales [22], we focus on these three relations in this study. We provide initial results for the following tasks: (1) understanding the features of corpus-based cues (2) exploring the value of such lexical cues in the detection of different rhetorical relations and (3) analysis of the potential differences and similarities among the cues extracted from two corpora that belong to different genres.

The experiments are conducted on two RST annotated corpora: RST corpus [6] and Simon Fraser University (SFU) review dataset [18]. The altered version of TF-IDF proposed in [3] is used to extract a set of key ngrams from each corpus as potential lexical cues, signaling the presences of rhetorical relations. By analyzing such corpus-based cues, this study moves beyond the fixed cue lists that are normally proposed and used in the prior literature. Such cue sets are commonly bound to a few well-defined syntactic categories such as conjunctions (e.g., *but* and *because*) and prepositions (e.g., *since* and *by*) [15], and are referred to as discourse markers or discourse connectives.

To obtain empirical evidence regarding the potential value of using lexical cues in identification of rhetorical relations, the list of top cues for each relation and extracted from each corpus was used to detect the instances of the same relation in the other corpus. For each of the experiments, the classification measures of accuracy, precision, recall, and F are calculated and analyzed.

The remainder of this paper is organized as follows: An overview of the previous research on automatic detection of rhetorical relations is provided in Section 8.2. In Section 8.3, an explanation of the underlying corpora and the method used to extract the cues is provided. The experiment results are described in Section 8.4, followed by a discussion of the findings given in Section 8.5. A conclusion is provided is Section 8.6, along with an overview of future research activities toward improving the cue-based approaches and the broader context of our research efforts.

8.2 Related Work

Work on automatic detection of rhetorical relations falls into two categories of research. A set of studies has relied on human-annotated corpora and proposed different models with various feature sets to detect rhetorical relations. Other researchers have followed a pattern-based approach, where a set of patterns are utilized to construct labeled artificial corpora. The built corpora are then used to construct relation extraction models.

Using Discourse Graph Bank [21], Wellner et al. [20] developed a classifier based on a

rich set of syntactic and lexico-semantic features, some of which are constructed based on external knowledge sources. Focusing on the recognition of implicit rhetorical relations, Lin et al. [11] proposed a classifier based upon Penn Discourse Treebank, by taking into account the context of the two spans, word pair information, as well as the spans internal constitution and dependency parses. Penn Discourse Treebank is also used by Pitler et al. [14], where several linguistically informed features such as word polarity, verb classes, and word pairs are used to build a model for automatic detection of relations.

Following the pattern-based approach to identify implicit relations, Marcu and Echihabi [13] used a small set of patterns built on a set of pre-specified discourse markers to extract the relations and to automatically form an annotated corpus of sentences. A corpus of non-relations is also formed by randomly selecting non-adjacent sentences that at least three sentences apart. These training sets are then used to learn pairs of words that are likely to co-occur in conjunction with each relation. Finally, to detect relations among text spans, word pair information is utilized, along with the set of discourse markers. Blaire et al. [4] extended this approach by refining the training and classification process using parameter optimization, topic segmentation, and syntactic parsing. Saito et al. [16] extended Marcu's approach [13] as well, showing that the phrasal patterns extracted from text spans can be valuable for the task of relation identification.

Biran and Rambow [3] followed a hybrid approach. In their study, rather than using a pre-specified list of discourse markers, a list of lexical cues (called relation indicators) are extracted from a human-annotated corpus (i.e., RST corpus). This list of corpus-based cues is used to build patterns and to form an artificial corpus. Word pair features are then utilized to extract a subset of rhetorical relations.

Similar to [3], in our work, we have chosen to use annotated corpora to extract a set of relation cues, moving beyond the pre-specified lists of discourse markers used in most of the previous works. However, instead of using this list to build an artificial corpus, we focused on the ability of corpus-based lexical cues in identification of different relations and the potential influence of genre-specific factors on the quality and performance of these cues.

8.3 Methodology

8.3.1 Underlying Corpora

We used two human-annotated corpora as our underlying datasets for the experiments: The RST corpus [6] and the SFU review dataset [18]. Both corpora are annotated in the RST

framework and are constructed using the RSTTool [2]. RST is one of the most widely accepted frameworks for discourse analysis and understanding. In the RST framework, elementary discourse units (i.e., atomic spans) can be related through particular discourse relations to form embedded text spans. The embedded spans can then participate in a new relation, leading to a hierarchy of relations that can be represented as a discourse tree.

The RST corpus, which has been made available by the Linguistic Data Consortium over the years, includes 385 Wall Street Journal articles and covers more than 178,000 words. Among the relations used for annotation, there exist three different variations of the EVALUATION relation. EVALUATION-S is used when the assessment occurs in the satellite (i.e., the less essential span in the relation) and EVALUATION-N is used when the assessment occurs in the nucleus (i.e., the more essential span in the relation). When the assessment is of equal weight in both spans, the instance is tagged with an EVALUATION label. In this study, the text spans tagged by each of the three labels are considered to be an EVALUATION instance. In addition, six sub-relations of ELABORATION (e.g., ELABORATION-ADDITIONAL, ELABORATION-SET-MEMBER, and ELABORATION-PART-WHOLE) are used to annotate the corpus. Relation instances tagged by each of these variations is treated as an ELABORATION instance.

SFU review corpus is a collection of 400 review documents from movie, book, and consumer products. This dataset contains over 303,000 words and was collected in 2004 from the Epinions Web site [1]. To ease the process of genre-specific comparisons and analysis in the rest of the manuscript, we refer to the RST corpus as the news corpus and the SFU dataset will be referred to as the review corpus. Table 8.1 shows the percentage of the relations of focus among the two corpora.

8.3.2 Lexical Cue Extraction

To extract lexical cues associated with a given relation, a relation document labeled according to its corresponding rhetorical relation (e.g., CIRCUMSTANCE, EVALUATION, and ELABORATION) is first created to include all the text spans that are linked with that relation in the underlying corpus. Similarly, we created a non-relation document, containing all the text spans that participate in any relation except for the relation of focus. Each pair of spans linked by a relation are concatenated in their correct order and represented in a single line.

Table 8.1: The percentage of the relations of focus in the underlying corpora

Relation	News	Reviews
CIRCUMSTANCE	3%	8%
EVALUATION	1%	2%
ELABORATION	36%	7%

Table 8.2: Sample lexical cues extracted from the two corpora

Relation	News Cues	Review Cues
CIRCUMSTANCE	when, now, since	while, until, once
EVALUATION	good, high, well	nice, love it, impressed
ELABORATION	who, which, as	which, where, as if

Forming these two documents for each relation allows us to use each corpus either as a training set to extract the cues, or a test set to analyze and evaluate the use of lexical cues in the relation extraction process.

In order to extract potential lexical cues, the approach proposed in [3] was followed. All the ngrams (up to trigrams) were first extracted from the relation document. For each ngram, an altered version of TF-IDF metric was then calculated. The IDF measure was still calculated based on the number of documents that contain the ngram and the total number of documents in the corpus. However, since each line corresponds to one instance of the relation, the TF metric is calculated based on the number of lines that contain at least one instance of the ngram. This altered metric allows us to offset the potential bias that may be caused by the TF metric for the words appearing more than once in a relation instance. The list of the extracted ngrams (i.e., lexical cues) were sorted based on the altered measure of TF-IDF in a descending order. This list was then filtered not to include any pronouns, modal verbs, or auxiliary verbs.

With the purpose of relation extraction, the sorted list of lexical cues extracted from each corpus was applied to the other corpus. As such, the lexical cues extracted from the news corpus were applied to the review corpus and the lexical cues extracted from the review set were used on the news corpus. The measures of accuracy, precision, recall, and F were then calculated for each of the cues independently. Figure 8.1 illustrates the overall process of cue extraction and relation classification for a relation of interest. Based on the experiment results, it could be seen that for all of the relations and for both datasets, those cues ranked after 120 had zero performance. Therefore, only the first 120 cues ordered by TF-IDF were involved in the analysis process. Table 8.2 includes samples of these cues extracted from the two corpora. Later, we will discuss how these two cue sets performed in the extraction of relations from the other corpus and how they overlap.

8.4 Experiment Results

The experiment results are reported in the form of line charts, where the vertical axis represents the measure of focus (i.e., TF-IDF, precision, recall, and F score) and the horizontal

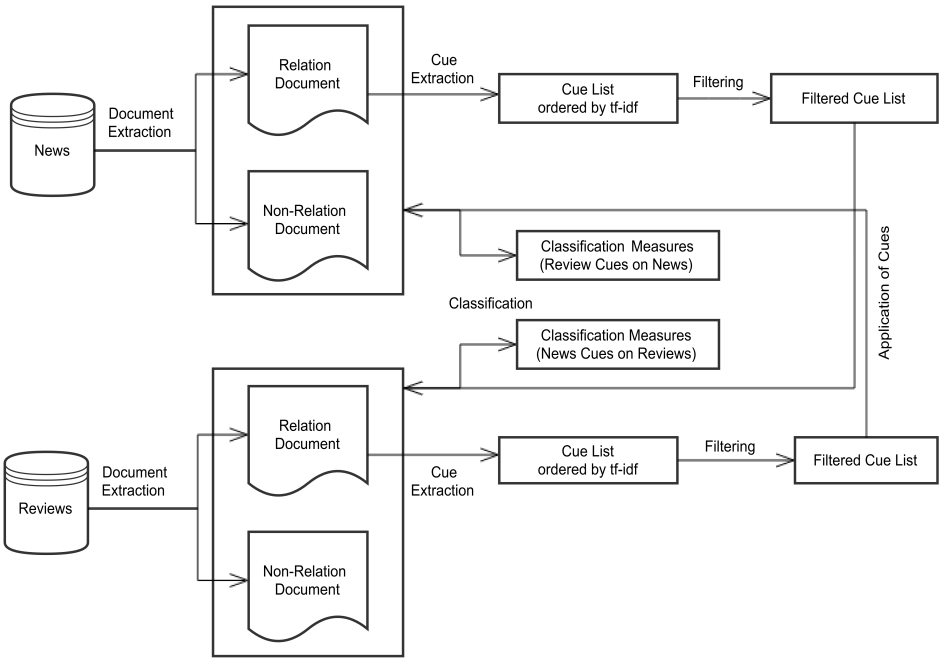


Figure 8.1: The extracted cues from each corpus are applied to the other corpus to classify rhetorical relations.

axis represents the 120 selected cues sorted based on the measure of focus in a descending order. Therefore, similar numeric labels on the charts may indicate different lexical cues. Table 8.3 shows the first two cues in the CIRCUMSTANCE charts. The dataset label in the charts denotes the training set, i.e. where the cues are extracted from. In all the charts, the darker colour represents the news dataset, indicating how lexical cues extracted from the news dataset performed on the review corpus. The lighter colour represents the review dataset, illustrating the measures for the top cues extracted from the review collection. Such a representation facilitates analysis of the trends and capabilities of cues among these relations and across genres.

As mentioned earlier, we created the non-relation documents by collecting all the instances that are linked by any relation other than the relation of focus. As such, the non-relation documents contained a much larger number of instances compared to the relation documents, leading to the large number of true negatives for each cue in our classification results. Therefore, the accuracy results were not reliable and are excluded from the analysis.

Figure 8.2 shows the TF-IDF measure calculated for the top cues extracted for all of the three relations and from both corpora. As can be seen, the TF-IDF measure is consistently lower for the lexical cues extracted from the reviews. This finding can be attributed to the

Table 8.3: Cue labels and their corresponding lexical cues for the CIRCUMSTANCE relation

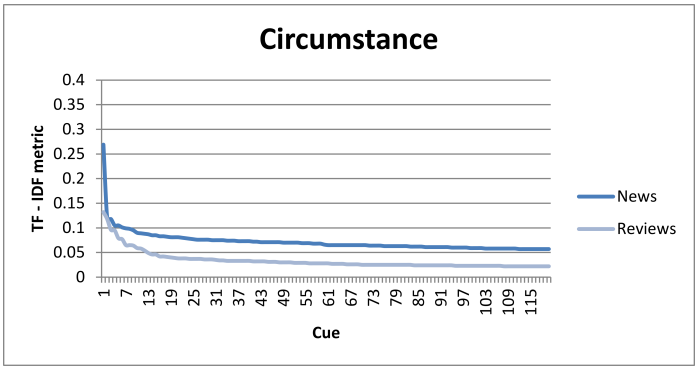
	Cue #1		Cue #2	
Measure	News	Reviews	News	Reviews
TF-IDF	when	when	as	when I
PRECISION	when it	it comes to	when	when it comes
RECALL	when	when	on	after
F SCORE	when	when	after	after

fact that news text is a well-structured formal writing, whereas online reviews are relatively less structured and informal, written by users with a wide range of writing abilities. Therefore, the lexical cues may be used in a more consistent way in the news collection compared to the review set, suggesting that the cue-based approaches might perform better on formal and structured text.

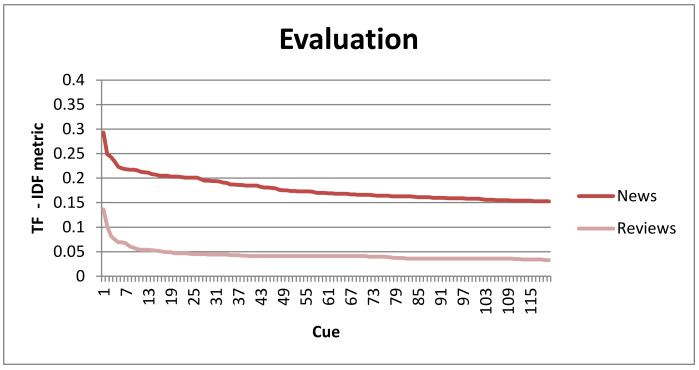
By comparing the plots for the three relations, it can be seen that the ELABORATION cues have a relatively lower TF-IDF for both datasets, indicating that regardless of the genre, ELABORATION may not be well-signaled by lexical cues. This metric is higher for both CIRCUMSTANCE and EVALUATION, especially for the top ranked cues; however, the difference between the cues extracted from the two corpora is more considerable for the EVALUATION relation. These results may indicate that lexical cues might be more genre-specific for the EVALUATION relation.

After the extraction of lexical cues from each set, we carried out further experiments by using these cues to extract rhetorical relations from the other set. Measures of precision, recall, and F were calculated for each cue independently. Figures 8.3 and 8.4 illustrate the precision and recall measures for the top cues, respectively. The results for CIRCUMSTANCE is consistent with the TF-IDF finding since lexical cues extracted from the news set have a relatively high precision score. The results for the EVALUATION could also be expected as it drops considerably after the first couple of cues, confirming that EVALUATION cues can be very specific to the underlying genre. The results for the ELABORATION, however, are not consistent with the TF-IDF metric. Even though TF-IDF measures were low for the ELABORATION cues, the precision is relatively high and is considerably higher for the lexical cues extracted from the review collection. This unexpected result can be attributed to the high percentage of ELABORATION instances in the news dataset. As can be seen in Table 8.1, more than one-third of the relation instances of the news set are annotated as ELABORATION. This large proportion of ELABORATION instances causes even a random cue to have a relatively high precision.

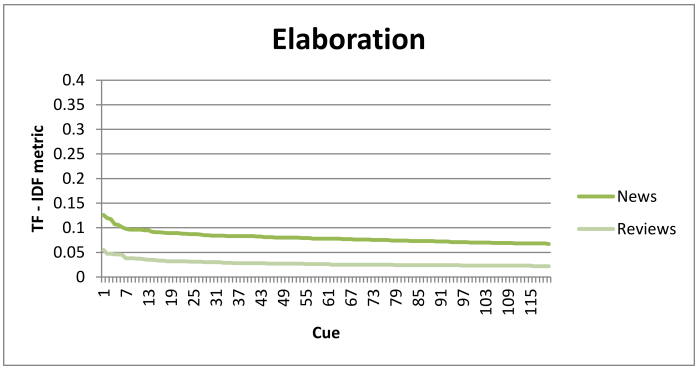
Unlike TF-IDF and precision, the recall metric has a very similar trend across all the relations and datasets (see Figure 8.4). For every cue, the instances classified as false



(a)



(b)



(c)

Figure 8.2: The TF-IDF metric is plotted for the top cues extracted from news and review datasets for the relations of CIRCUMSTANCE (a), EVALUATION (b), and ELABORATION (c).

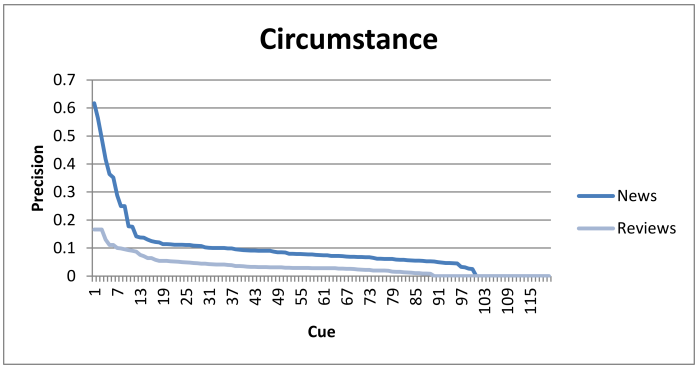
negative consist of those explicit relations that are signaled by any lexical cue other than the cue of focus as well as implicit instances that are not signaled at all. A relatively large proportion of relations are normally implicit [17], which can cause the number of true positives and false negatives signaled by other cues to have little influence on the recall results. Therefore, the recall measure is mainly characterized by implicit relations.

Finally, the results from the calculation of F score, as the most reliable performance metric, are plotted in Figure 8.5. Except for the ELABORATION, the majority of the lexical cues extracted from the news dataset are performing better in the extraction of rhetorical relations. This better performance for these two relations is consistent with the TF-IDF results. As well, the better performance of the extracted cues from the review set on the news corpus can be explained by a large proportion of the news dataset being labeled as ELABORATION. Comparison of results for different relations indicates that the top CIRCUMSTANCE cues have a better performance compared to the top cues from the other two relations, while the difference is less considerable between EVALUATION and ELABORATION.

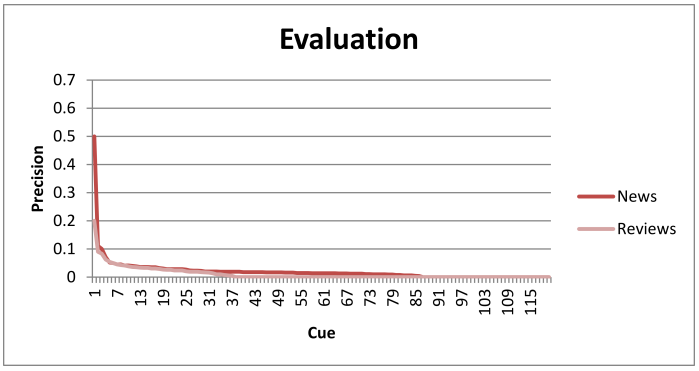
8.5 Discussion

The results of the experiments demonstrates that the top cues for the CIRCUMSTANCE relation performed better that the top cues from the other two relations and the results are consistent for both datasets. This finding can be attributed to the fact that the relations that are typically expressed through subordination are more heavily signaled compared to the relations that hold between two or more sentences (e.g., ELABORATION and EVALUATION) [17]. For the CIRCUMSTANCE relation, the top cues are from the common discourse markers proposed and used in the prior literature. For example, *when*, *since*, *after* and *before* are among the top cues for both corpora. This large number of cues being discourse markers makes the CIRCUMSTANCE relation relatively genre-independent and has led to the similar performance on both datasets. Counting the number of overlapping cues extracted from both corpora confirms this finding since there are 29 similar cues in the two cue lists, with all of them being discourse markers.

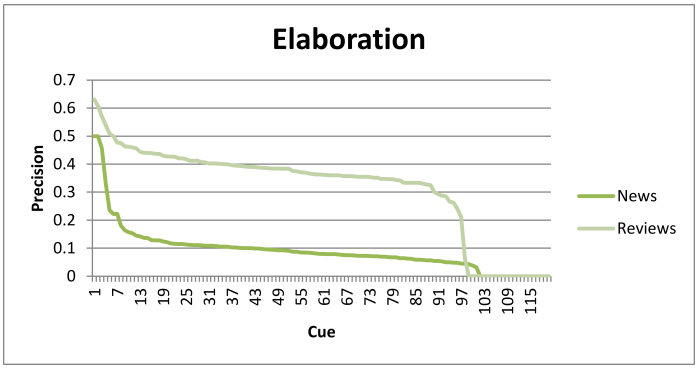
Based on the experiment results for EVALUATION, it could be concluded that the EVALUATION relation may be commonly signaled depending on the text genre and the lexical cues might be more genre-specific compared to the other two relations. The majority of EVALUATION cues extracted from both datasets are not from among the traditional discourse markers (see Table 8.2 for samples), but are from underexplored syntactic groups such as adjectives, making it reasonable to have different cues from two different text genres. For example, *love it* is a lexical cue commonly found in EVALUATION instances in the review cor-



(a)

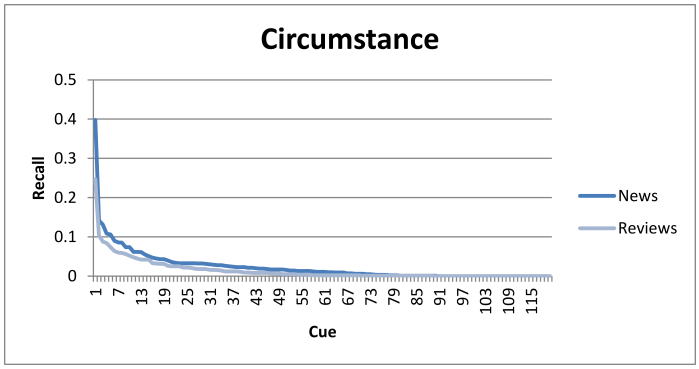


(b)

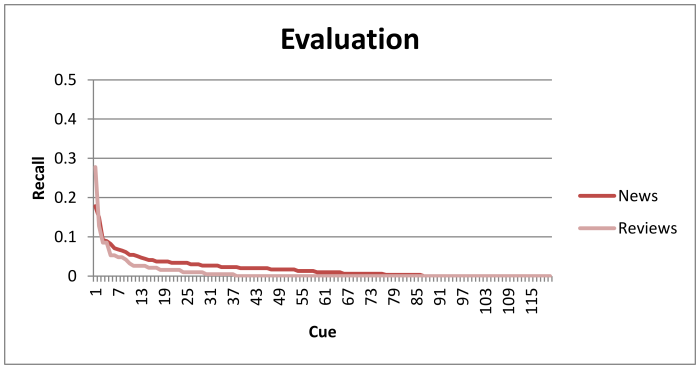


(c)

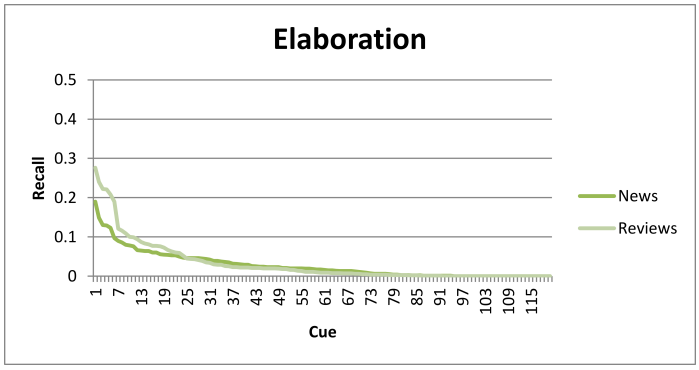
Figure 8.3: The precision metric is plotted for the top cues extracted from the training set, which is represented by the darker colour, for the relations of CIRCUMSTANCE (a), EVALUATION (b), and ELABORATION (c).



(a)

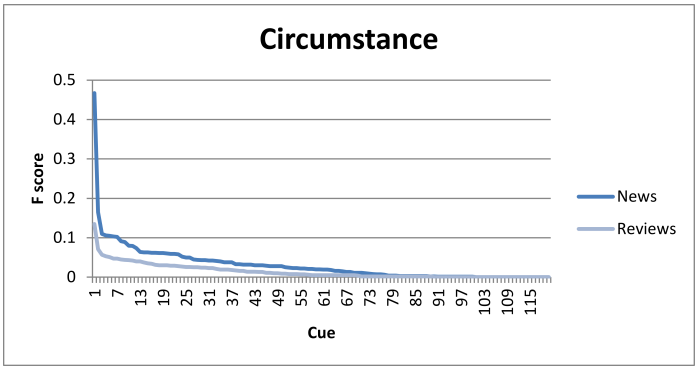


(b)

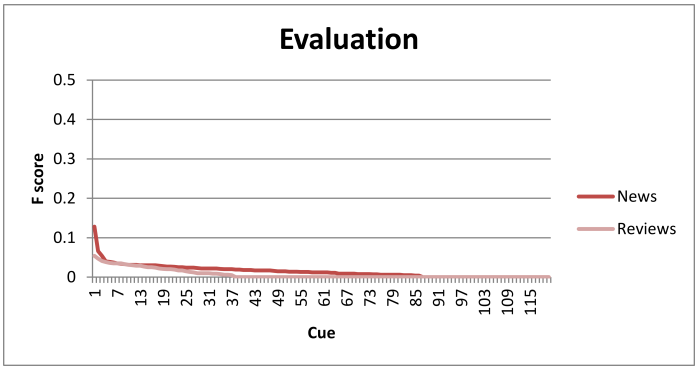


(c)

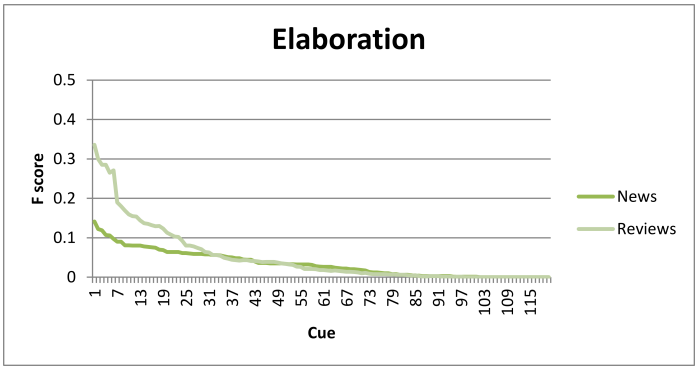
Figure 8.4: The recall metric is plotted for the top cues extracted from the training set, which is represented by the darker colour, for the relations of CIRCUMSTANCE (a), EVALUATION (b), and ELABORATION (c).



(a)



(b)



(c)

Figure 8.5: The F metric is plotted for the top cues extracted from the training set, which is represented by the darker colour, for the relations of CIRCUMSTANCE (a), EVALUATION (b), and ELABORATION (c).

pus (see Table 8.2); however, it does not appear in the news corpus at all. The low number of overlaps (5 overlapping cues) between the two cue sets confirms this finding.

The results for the ELABORATION relation might not be as reliable due to the imbalanced number of ELABORATION instances among the two corpora (see Table 8.1). A large proportion of ELABORATION instances in the news corpus is a result of the structure and nature of the news, where each part of the text is provided to elaborate on the the previous part. However, from the analysis of TF-IDF metric it can be concluded that the ELABORATION relation is lightly signalled, resulting in the cue-based methods not to be well suited to extract ELABORATION instances. Although there are 42 overlapping cues among the two cue sets, the cues are not heavily weighted to be relied on. This finding is consistent with prior linguistic literature, arguing that ELABORATION is too ill-defined to be even considered a relation [9].

Overall, the results suggest that the nature of the relation plays a significant role in the effectiveness of using lexical cues for the relation detection task. In addition, the syntactic group that these cues belong to can determine to what degree a relation is specific to the underlying genre. When the common discourse markers are the dominant cues, the relation can be considered genre-independent, while in the case of adjectives or verbs as cues, the relation is more genre-specific. Our research can inform downstream studies on automatic detection of rhetorical relations. For development of cue-based algorithms, one should think of different relations (or at least relation categories) independently and should accommodate for genre-specific factors for the relations that are signaled by other kinds of cues beyond discourse markers.

When considering cue-based approaches, choosing the optimal number of cues can be a challenging task. We ran some experiments with different number of cues and calculated the cumulative precision, recall, and F score for various cue sets. Table 8.4 includes the results for the CIRCUMSTANCE cue sets of different sizes extracted from the news collection and applied to the review corpus. As can bee seen, the precision is constantly decreasing as new cues are being added, suggesting that the number of false positives added by the new cues is larger compared to the added number of true positives. The recall measure, however, is increasing as the cue set grows larger. The results for recall was expected since we are able to extract more true positives as we add new lexical cues. The F score is decreasing, indicating that the influence of precision is larger than recall.

For this particulate experiment, surprisingly, the optimal F score was reached by the first cue (*when*) alone. When the second cue (*as*) was added, the large number of false positives caused the precision and eventually the F score to drop, despite the improvement in the number of true positives and in the recall metric. These results call for further computational steps to be taken in cue-based approaches to disambiguate the cues and to identify

Table 8.4: The precision, recall, and F score are calculated for cue sets of different sizes. The results are reported for the cues extracted from the news corpus and applied to the review corpus for the CIRCUMSTANCE relation.

Measure	15	45	60	75	90	105	120
Precision	0.24	0.11	0.10	0.9	0.09	0.08	0.08
Recall	0.59	0.85	0.91	0.94	0.94	0.95	0.95
F Score	0.34	0.20	0.18	0.16	0.16	0.16	0.15

in what context the cue is indeed signaling the existence of a relation. After adding such extra steps, further experiments are required to analyze and study the optimal number of cues for different relations and text genres.

In RST, atomic text units are linked to each other through rhetorical relations. These linked units then can participate in new relations with other spans, resulting in a hierarchy of relations. Table 8.5 shows the distribution of atomic and embedded relations for all of the three relations and among both corpora. As can be seen, the number of embedded relation instances constitutes a considerable proportion of instances in all cases. This percentage in consistently higher for the news corpus. This finding was expected since, unlike the news dataset, the review corpus is annotated at the sentence level. One of the potential limitations of this approach is the lack of focus on atomic and embedded relations as two different types of relations. However, the presence of multiple relations inside a relation can indeed affect cue extraction approaches. For example, consider the two following instances of the CIRCUMSTANCE relation from the news corpus:

Instance #1:

[dressed in jeans and a sweatshirt] [as she slogs through the steady afternoon rain]$_{\text{circumstance}}$

Instance #2:

[recently, a contractor saved her from failing three stories][[as she investigated what remained of an old Victorian house][torched by an arsonist]$_{\text{elaboration}}$]$_{\text{circumstance}}$

The first instance consists of two atomic spans. The second instance, however, consists of one atomic span as its first span and a non-atomic span as its second one. Therefore,

Table 8.5: Distribution of atomic and embedded relations for the three relations and among the two corpora.

	News		Reviews	
Relation	Atomic	Embedded	Atomic	Embedded
CIRCUMSTANCE	46%	54%	78%	22%
EVALUATION	21%	79%	69%	31%
ELABORATION	48%	52%	70%	30%

using our approach, *as* and *by* will be assigned the same TF weight for the second instance, while *by* is clearly independent of the CIRCUMSTANCE and is used to elaborate on the previous part. Hence, it is necessary to conduct in-depth analysis to understand how such hierarchical structure can influence the cue-based approaches and how to detect and re-weight less relevant parts of an instance when seeking lexical cues.

8.6 Conclusion

We described experiments carried out to advance our understanding of the potential and limitations of the use of lexical cues for automatic identification of rhetorical relations in text. By extracting lexical cues from annotated corpora, we moved beyond the common use of pre-specified discourse markers and explored if lexical cues from other syntactic categories can signal the presence of relations. In addition, we aimed at understanding how different relations are signaled. Furthermore, we used two different corpora from two different genres to investigate whether genre-specific factors can influence how relations are signaled. Overall, the results indicate that how relations are signaled, both in general and across genres, is largely dependent on the nature of the relation.

One of the main differences of RST with other theories is that rhetorical relations place emphasis on the writers intentions and the effect of the relation on the reader [17]. As such, RST annotations are inherently subjective and are based on the readers' understanding of the text [17]. Therefore, some of the differences found across the two corpora can be a result of the potentially different knowledge of each set of annotators regarding the culture, situation, and language that the text represents.

RST postulates a hierarchical structure on text; thus, participating text spans can be either atomic or may consists of multiple atomic spans. One of our future research actives will be focused on the analysis of embedded and atomic sentences independently. In addition, by focusing on well-signaled relations such as CIRCUMSTANCE, we plan to improve the cue-based approaches by seeking novel methods to disambiguate lexical cues. We are currently building graph-based models to disambiguate the cues by taking into account their syntactic context. Furthermore, we will carry out further experiments to identify the optimal number of lexical cues for different relations.

As mentioned in the introduction Section, CIRCUMSTANCE, EVALUATION, and ELABORATION are the three common RST relations in rationales [22]. As the first approach toward the identification of rationales in online corpora, we attempted to detect these three rhetorical relations. A rationale is an explanation of the reasons underlying decisions, conclusions, and interpretations. Prior studies on rationale articulation and sharing suggest that it con-

tributes to quality control, knowledge management and reuse by extracting the rationales from the text generated through these large-scale ideation and deliberation [24, 23]. There have been a few attempts in applying computational techniques to identify rationales from ill-structured text such as online discourse and reviews [3, 7, 5]. Our study contributes to the research effort in this emerging area by demonstrating the potential and limitations of using rhetorical relations to detect rationales.

Bibliography

[1] Epinions. http://www.epinions.com/.

[2] RSTTool. http://www.wagsoft.com/RSTTool.

[3] Or Biran and Owen Rambow. Identifying justifications in written dialogs. In *Proceedings of the IEEE International Conference on Semantic Computing*, pages 162–168, 2011.

[4] Sasha Blair-Goldensohn, Kathleen McKeown, and Owen Rambow. Building and refining rhetorical-semantic relation models. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Lingustics*, pages 428–435, 2007.

[5] Filip Boltužić and Jan Šnajder. Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 49–58, 2014.

[6] Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. Building a discourse tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, pages 1–10, 2001.

[7] Debanjan Ghosh, Smaranda Muresan, Nina Wacholder, Mark Aakhus, and Matthew Mitsui. Analyzing argumentative discourse units in online interactions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 39–48, 2014.

[8] J.R. Hobbs. *Literature and Cognition*. Center for the Study of Language and Information - Lecture Notes. Cambridge University Press, 1990.

[9] Alistair Knott and Ted Sanders. The classification of coherence relations and their linguistic markers: An exploration of two languages. *Journal of Pragmatics*, 30(2):135 – 175, 1998.

[10] Alex Lascarides and Nicholas Asher. Temporal interpretation, discourse relations and commonsense entailment. *Linguistics and Philosophy*, 16(5):437–493, 1993.

[11] Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. Recognizing implicit discourse relations in the Penn Discourse Treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 343–351, 2009.

[12] William C. Mann and Sandra A. Thompson. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988.

[13] Daniel Marcu and Abdessamad Echihabi. An unsupervised approach to recognizing discourse relations. In *Proceedings of the Annual Meeting on Association for Computational Linguistics*, pages 368–375, 2002.

[14] Emily Pitler, Annie Louis, and Ani Nenkova. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference of the Annual Meeting of the Association of Computational Lingustics and the International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, pages 683–691, 2009.

[15] Rashmi Prasad, Aravind Joshi, and Bonnie Webber. Realization of discourse relations by other means: Alternative lexicalizations. In *Proceedings of the International Conference on Computational Linguistics*, pages 1023–1031, 2010.

[16] Manami Saito, Kazuhide Yamamoto, and Satoshi Sekine. Using phrasal patterns to identify discourse relations. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Lingustics*, pages 133–136, 2006.

[17] Maite Taboada. Discourse markers as signals (or not) of rhetorical relations. *Journal of Pragmatics*, 38(4):567 – 592, 2006.

[18] Maite Taboada, Caroline Anthony, and Kimberly Voll. Methods for creating semantic orientation dictionaries. In *Proceedings of the Conference on Language Resources and Evaluation*, pages 427–432, 2006.

[19] Maite Taboada and William C. Mann. Applications of rhetorical structure theory. *Discourse Studies*, 8(4):567–588, 2006.

[20] Ben Wellner, James Pustejovsky, Catherine Havasi, Anna Rumshisky, and Roser Saurí. Classification of discourse coherence relations: An exploratory study using

multiple knowledge sources. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, pages 117–125, 2006.

[21] Florian Wolf and Edward Gibson. Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31(2):249–288, 2005.

[22] Lu Xiao. Do members converge to similar reasoning styles in teamwork? A study of shared rationales in small team activities. In *Proceedings of the iConference*, pages 524–530, 2013.

[23] Lu Xiao. The effects of a shared free form rationale space in collaborative learning activities. *Journal of Systems and Software*, 86(7):1727 – 1737, 2013.

[24] Lu Xiao. Effects of rationale awareness in online ideation crowdsourcing tasks. *Journal of the Association for Information Science and Technology*, 65(8):1707–1720, 2014.

Chapter 9

RST Cue Disambiguation

9.1 Introduction

A semantically sound text consists of discourse units that are connected through discourse relations, which are also referred to as rhetorical relations. Despite the efforts to build robust theoretical foundations and taxonomies for such relations [8, 13, 14, 16], current methods for their automatic analysis and discovery in written discourse have yet to improve. However, providing robust models to analyze and identify rhetorical relations can benefit various research directions in computational linguistics such as text generation [9] and summarization [17], and machine translation [18].

One of the widely accepted frameworks for discourse analysis and understanding is Rhetorical Structure Theory (RST) [16]. In RST, discourse structure has a form of a tree, where the leaves correspond to elementary discourse units, and the internal nodes correspond to contiguous text spans. Each internal node is marked with a rhetorical relation that holds between its child nodes. Figure 9.1 provides an example of an RST tree taken from the RST corpus [3]. One of the notable differences of RST with other similar theories is that it is structured on the intentions of the writers to use those relations [24]. This distinctive feature may make it even more difficult to build models for automatic identification and classification of rhetorical relations in the context of RST.

Rhetorical relations can be either explicit or implicit. Explicit relations are the ones that are signaled by cues, such as lexical cues, mood, modality, and intonation [24], while no cue is present in implicit relations. In this study, we are focused on explicit relations

A version of this chapter has been published in the *proceedings of the Conference on Emprical Methods in Natural Language Processing (EMNLP LSDSem'5).*

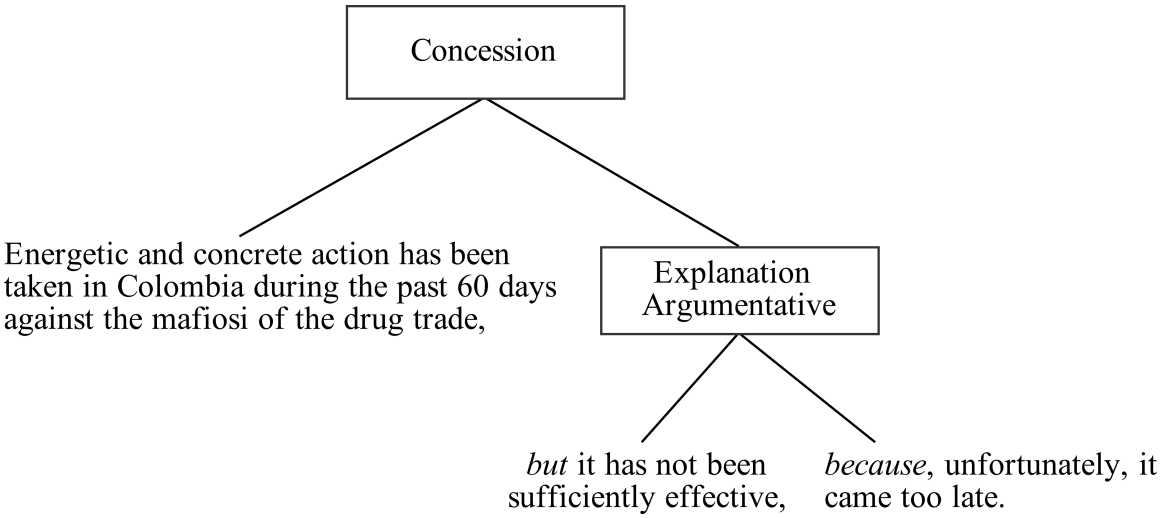


Figure 9.1: An example sentence parsed in the form of RST

in written text that are signaled by the presence of lexical cues. Lexical cues are defined as linguistic expressions that function as explicit indicators of a discourse relation [7]. For example, in the sentence provided in Figure 9.1, *but* and *because* can be considered lexical cues signaling the existence of the CONCESSION relation and the EXPLANATION-ARGUMENTATIVE relation, respectively.

Since this study is part of a larger project to identify rationales in written discourse, we focus on the three relations of CIRCUMSTANCE, EVALUATION, and ELABORATION that are commonly present in rationales [28]. With the aim of proposing a cue-based approach to extract rhetorical relations, we have carried out some corpus-based experiments on RST annotated corpora. As a result of these experiments, we have generated a list of key n-grams as potential lexical cues for each relation. Such a corpus-based method may result in the discovery of underexplored lexical cues.

Even though lexical cues can be exploited to label rhetorical relations, they are not always unambiguous [20]. Some linguistic expressions may or may not function as a lexical cue, or they may signal different types of relations in different sentences. Hence, here, we propose a graph-based probabilistic model that takes into account the syntactic features of sentences. These models are intended to determine in what syntactic context a lexical cue is indeed signaling the presence of a particular relation.

The results of the evaluation of our approach are presented and discussed for the CIRCUMSTANCE relation. CIRCUMSTANCE is chosen as the relation of focus since [11] revealed that the cue-based approaches can be well-suited for the detection of CIRCUMSTANCE across different genres, while the ELABORATION relation is not normally signaled [11]. In addition, the features of the underlying text genre can significantly influence how EVALUATION is sig-

naled [11].

The remainder of this paper is organized as follows: An overview of the previous research on explicit relation classification and lexical cue disambiguation is provided in Section 9.2. In Section 9.3, an explanation of the underlying corpora and the methods used to extract and disambiguate the cues is provided. The evaluation results are presented in Section 9.4. A discussion of the findings is given in Section 9.5, followed by a conclusion of the study in Section 9.6.

9.2 Related Work

The majority of studies focusing on discourse parsing and discourse relation classification report results achieved from both explicit and implicit relations [23, 27, 26]. Among the works that are particularly focused on lexical cue disambiguation, a large proportion are conducted on the Penn Discourse TreeBank (PDTB) [21], while fewer studies have been conducted to study other discourse theories and frameworks.

PDTB annotation is lexically-grounded, and it is theory-neutral with respect to higher-level discourse structure [22]. In the course of the annotation, the annotators were asked to seek lexical items that can signal discourse relations and then annotate their corresponding arguments and relations [22]. Even for implicit relations, annotators were asked to look for adjacent sentences that lacked one of these signals. When a discourse relation could be inferred, they were asked to first label the relation with a lexical item that could serve as a signal and then annotate the relation sense. Such a lexically oriented approach to annotate discourse relations has motivated a lot of work on disambiguation of lexical cues in PDTB.

For example, Miltsakaki et al. [19] have utilized a set of syntactic features along with a maximum entropy model to disambiguate three discourse cues of *while*, *since*, and *when*. Their feature set includes form of the auxiliary *have*, form of the auxiliary *be*, form of the head, and presence of a modal. They obtained an accuracy of 75.5% to classify *since* given its three possible senses, 71.8% for *while* given its four possible senses, and 61.6% for *when* given its three possible senses.

Pitler and Nenkova [20] also used a set of syntactic features to disambiguate discourse cues regarding their discourse and non-discourse usage, as well as sense disambiguation. Their features of focus consist of the syntactic category of the marker itself, its immediate parent, along with its left and right siblings. In addition, two binary features are taken into account to indicate whether the right sibling contains a VP and if the right sibling contains a trace. Their best feature set also included pairwise interaction features between the cues and syntactic features as well as pairwise interactions between the syntactic features

themselves. Their learning algorithm resulted in an F-score of 92.28% and an accuracy of 95.04% for discourse versus non-discourse usage and an accuracy of 94.15% for sense classification.

These results were later improved in [10], where a set of novel surface-level and syntactic features are introduced and are combined with the feature set presented in [20]. The results of a maximum entropy classifier trained using this feature set resulted in an accuracy of 97.78% and an F-score of 96.22%.

Within a broader context of building an end-to-end discourse parser for PDTB, Lin et al. [15] built a cue classifier to identify whether a lexical item functions as a discourse cue or not. In addition to the features used in [20], they also included part-of-speech features as well as features related to the syntactic parse path from the cue to the root of the tree. Using their set of lexico-syntactic and path features resulted in an accuracy of 97.25% and an F-score of 95.36%.

Even though RST is one of the most widely accepted frameworks for discourse analysis, relatively little attention has been paid to RST annotated corpora in regards to lexical cue analysis and disambiguation. Unlike PDTB, annotations following RST are not lexically-grounded, and every relation is defined in terms of intentions that lead authors to use those specific relations [24]. Therefore, an RST diagram represents some of the authors' purposes or intentions for including each part of the text [24]. Such attributes of RST annotations make it a challenging task to study the role of lexical items in relation classification and to potentially disambiguate them.

Marcu [17] attempted to create a rhetorical parsing algorithm, which takes an unrestrictive free text and generates a valid rhetorical structure tree. A corpus study was conducted to understand how cues can be used to identify elementary discourse units and hypothesize their corresponding relation. By utilizing prior studies on discourse analysis, he created a list of 450 discourse cues to start with. An average of 17 text spans associated with each cue was then collected from the Brown corpus.

All of the sentences were then annotated with two sets of metadata: discourse-related information (e.g., marker, usage, and position) and algorithmic features. Using these annotations, which mostly capture the orthographic environments of the cues, a set of regular expressions was created manually to recognize potential discourse cues. If a discourse cue had different discourse functions in different orthographic environments, a separate regular expression was made for each case. The algorithm resulted in an 84.9% F-score from recall of 80.8% and precision of 89% for the sub-task of cue identification. For the sub-task of relation classification, they achieved a recall of 47.0% and precision of 78.4%, resulting in an F-score of 58.76%.

In [4], a set of ambiguous discourse cues is first extracted from the database of Spanish discourse cues. The context of each cue is then extracted from the RST Spanish Treebank and is given to a syntactic parser. The syntactic features of the context of each cue are then manually analyzed to identify potential linguistic regularities and patterns. By using the results of the analysis, linguistic rules are developed to disambiguate the lexical cues. By evaluating their rules on the test corpus of the RST Spanish TreeBank, they achieved an accuracy of 60.65%.

Both of the works on RST annotations are semi-automated and include manual steps. Our work is intended to provide a fully automated approach to detect potential lexical cues that can indicate rhetorical relations, and to analyze whether their syntactic context can be of value for cue disambiguation and sense classification.

9.3 Approach

In this section, we first describe the two RST annotated corpora that are used in the present work: RST corpus [3] and Simon Fraser University (SFU) review dataset [25]. Then, an explanation of the approach used to extract a set of key n-grams as potential lexical cues is presented, which is followed by a description our graph-based approach to disambiguate lexical cues.

9.3.1 Corpora

We used two human-annotated corpora as our underlying datasets for the experiments: the RST corpus [3] and the SFU review dataset [25]. Both corpora are annotated in the RST framework and are constructed using the RSTTool [1].

The RST corpus, which has been made available by the Linguistic Data Consortium over the years, includes 385 *Wall Street Journal* articles and covers more than 178,000 words. Among the relation instances in the RST corpus, there exist around 700 instances of CIRCUMSTANCE, which constitutes almost 3% of the total number of relation instances.

The SFU review corpus is a collection of 400 review documents from movie, book, and consumer products. This dataset contains over 303,000 words and was collected in 2004 from the Epinions Web site [2]. There exist around 1300 CIRCUMSTANCE instances, constituting almost 7% of the annotated instances in the corpus.

[1] http://www.wagsoft.com/RSTTool
[2] http://www.epinions.com/

9.3.2 Lexical Cue Selection

The news text has a well-structured formal writing style, whereas the online reviews are relatively less structured and informal, written by users with a wide range of writing abilities. Therefore, to extract lexical cues associated with a given relation, we used the RST corpus.

First, all the relation instances are extracted from the RST corpus and are collected in a relation document named after the corresponding relation. Then, following the approach proposed in [1], all the n-grams (up to tri-grams) are extracted from the composed relation document. For each n-gram, an altered version of TF-IDF metric is then calculated. The IDF measure is still calculated based on the number of documents that contain the n-gram and the total number of documents in the corpus. However, since each line corresponds to one instance of the relation, the TF metric is calculated based on the number of lines that contain at least one instance of the n-gram. This altered metric allows us to offset the potential bias that may be caused by the TF metric for the words appearing more than once in a relation instance.

The list of the extracted n-grams (i.e., lexical cues) is then filtered to only include the n-grams with their TF-IDF above 0.5. To filter any corpus-specific n-grams that may appear in the list, the n-grams extracted from the RST corpus are applied to the SFU review dataset to identify the corresponding relation. The F-score of each n-gram is then calculated independently. Finally, the n-grams with an F-score of above 0.1 are selected as potential lexical cues. The following provides a step-by-step explanation of the process for the CIRCUMSTANCE relation, which resulted in the selection of seven lexical cues: *When*, *after*, *on*, *before*, *with*, *out*, *as*. These results are extensively analyzed in another document [11].

- Extract all of the CIRCUMSTANCE instances from the RST corpus and form the circumstance document
- Calculate an altered version of TF-IDF for all of the n-grams (up to tri-grams) that appear in the circumstance document.
- Process the n-gram list to remove the ones with their TF-IDF below a a certain threshold.
- Use the updated list of n-grams to extract CIRCUMSTANCE instances from the SFU review corpus and calculate their F-score independently.
- Select the final cues with an F-score above a certain threshold.

9.3.3 Lexical Cue Disambiguation

Our cue disambiguation approach is mainly inspired by the work of [6] on the detection of sentences with attitudes. In their study, the text fragment that includes a second pronoun is first extracted as the most relevant part of a sentence. These fragments are then represented using different patterns, capturing their syntactic features and semantic orientation. For every kind of pattern, graph models are built based on sentences with and without attitude. Finally, the likelihood of a new sentence being generated from these models is used to predict the existence of an attitude. We adopted their approach for lexical cue disambiguation. Our graph models are built on the RST corpus and evaluated on the SFU review corpus and vice versa. Therefore, the graph building procedure explained below is conducted on both underlying corpora.

Data Collection

For every extracted cue, we first create two corresponding documents from the annotated corpora. One document consists of all of the relation instances that contain the cue and are annotated with the relation of focus (e.g., all of the CIRCUMSTANCE instances that are signaled by *when*). From now on, such instances will be referred to as positive instances. The other document consists of all the relation instances that contain the cue and are annotated as any relation except for the relation of focus (e.g., all of the non-circumstance instances that contain *when*). In the rest of this manuscript, we will refer to these instances as negative instances.

RST postulates a hierarchal structure on text, where a relation instance can be embedded in other instances. Therefore, during the extraction of the instances, we ensured not to collect negative instances that include any positive or negative sub-instance. We also ensured not to collect any positive instances that include negative sub-instances. The inclusion of such embedded instances would have resulted in redundant and incorrect data points. For example, consider the following positive instance from the RST corpus:

[*When* Mr. Gandhi came to power,]

[he ushered in new rules for business]$_{\text{circumstance}}$

When collecting negative instances, it was revealed that this instance was embedded in ten negative instances. However, since *when* is in fact functioning as a circumstance cue in all of them, those ten instances could not qualify as negative instances and so were excluded.

Syntactic Representations

After creation of the documents, each instance is processed and transformed into two different representations, capturing the syntactic features of the instance. To create the first syntactic representation, words in instances are replaced with their corresponding Part-Of-Speech (POS) tags, while the cue itself is kept as is. The second representation includes the shortest path from the root element to the cue in the dependency parse tree. The following is an example of the CIRCUMSTANCE relation, along with its two corresponding syntactic representations:

- Positive instance with *when* as the cue:
  When Mr. Gandhi came to power, he ushered in new rules for business
- POS-based representation:
  When NNP NNP VBD TO NN PRP VBD IN JJ NNS IN NN
- Shortest path representation:
  root advmod

We used the OpenNLP[3] toolkit to tokenize and POS tag the instances and the Stanford dependency parser to generate the parse trees [12].

Graph Modeling

We encoded the syntactic information of the instances in graph models. We build the directed weighted graph $G = (V, E), w$, where:

- V is the set of all possible tokens that may appear in the representations. For example, for the POS representations, V is the union of the set of all POS tags and the cue set.
- $E = V \times V$ is the set of all possible ordered transitions between any two tokens.
- $w \rightarrow [0 - 1]$ is a weighting function that assigns a probability value to an edge (i, j), which represents the probability of a transition from token i to token j.

Given a set of syntactic representations, the probability of a transition from token i to token j is calculated following a maximum likelihood estimation. Thus, the probability is calculated by dividing the number of times that token i is immediately followed by token j by the number of times that token i itself appears in the set.

[3] https://opennlp.apache.org/

This method of building the graphs is similar to language modeling but is conducted over a set of syntactic representations [6]. For every kind of representation, we build one graph based on the set of positive instances, and one based on the set of negative instances. As a result, given a cue (e.g., *when*) and its corresponding relation (e.g., CIRCUMSTANCE), we build four graph models based on the following sets: POS representations of positive instances, POS representations of negative instances, dependency parsed representations of positive instances, and dependency parsed representations of negative sentences.

Cue Disambiguation Model

Finally, for our final cue disambiguation model, we utilize the probability values obtained from our graph models as the feature set for a standard machine-learning model. Given an instance and a graph, we calculate the likelihood of its syntactic representations to be generated from the corresponding syntactic graphs. The probability of a syntactic representation R that consists of a sequence of tokens $T_1, T_2, ..., T_n$ being generated from graph G is estimated using the following formula. Note that W is the weighting or probability transition function.

$$\begin{aligned} P_G(R) &= \prod_{i=2}^{n} P(T_i|T_1, ..., T_{i-1}) \\ &= \prod_{i=2}^{n} W(T_{i-1}, T_i) \end{aligned}$$

Given that we have four graph models, we can generate four probability values as our feature set. These features are further used in a standard supervised learning algorithm to disambiguate the cue and to classify the relation of a given instance. Figure 9.2 provides a high-level description of the entire procedure of lexical cue extraction and disambiguation.

9.4 Evaluation

Given that our ultimate goal is to detect rationales from written discourse, our approach is evaluated for the CIRCUMSTANCE relation as the only cue-based relation that is known to be frequently present in rationales [11, 28]. We carried out experiments using different forms of POS representations based on the number of POS tags surrounding the cue and the granularity of the tags. We conducted experiments using the entire POS tagged instance, using two POS tags before and two tags after the cue, and using one before and one after the cue. We also used three levels of POS tag granularity, including the finest that is the

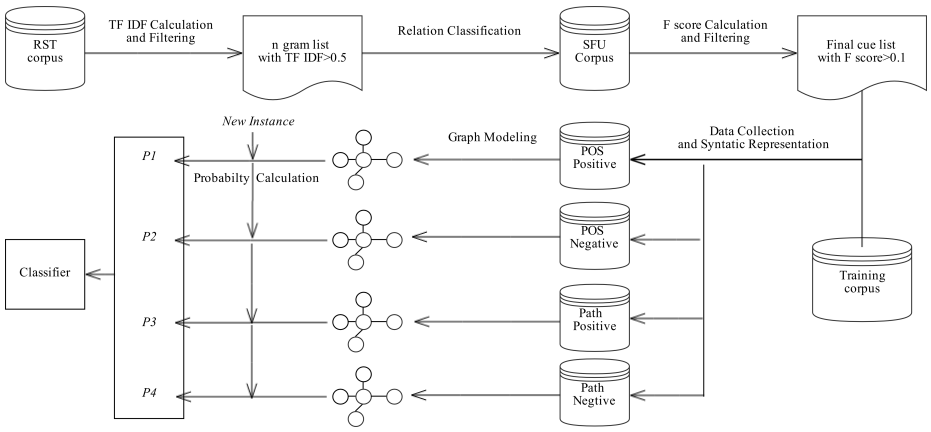


Figure 9.2: A high-level overview of the cue extraction and disambiguation approach

Table 9.1: In addition to the POS tags in the Penn English Treebank tag set, experiments are conducted using tags grouped according to different levels of granularity.

Label	Medium Granularity	Coarse Granularity
JJ	JJ, JJR, JJS	JJ, JJR, JJS, DT, WDT
NN	NN, NNS, NNP, NNPS	NN, NNS, NNP, NNPS
PRP	PRP, PRP$	PRP, PRP$, WP, WP$
RB	RB, RBR, RBS	RB, RBR, RBS, WRB
WP	WP, WP$	-
VB	VB, VBD, VBN, VBP, VBZ	VB, VBD, VBN, VBP, VBZ, MD, VBG

Penn English Treebank[4] tagset used by OpenNLP. We also used a medium and a coarse granularity that are created by categorization of similar tags into one high-level tag. Table 9.1 shows the tags that are categorized in each of these two granularity levels. Note that the tags not mentioned in the granularity levels are used as is.

Using these three variations of the two attributes of POS tags resulted in nine different experiment settings. We achieved our best results on both corpora using one tag before and one tag after the cue and the medium granularity level. In this section, we report results for experiments using this particular POS setting.

The final algorithm built on probability values is executed and evaluated using Weka workbench[5] and classifies instances via regression[6]. A stratified ten-fold cross validation approach is followed to evaluate the model. To gain insight into the effectiveness of our model in the disambiguation of different cues, results are reported for each of the seven cues independently. The SMOTE filter was used when we encountered significant class

[4]http://www.cis.upenn.edu/ treebank/

[5]http://www.cs.waikato.ac.nz/ml/weka/

[6]ClassificationViaRegression algorithm is used with default parameters

imbalance.

Table 9.2 demonstrates the results when the RST corpus was used to build the graphs, and the SFU corpus was used to build and test the final model. Table 9.3 shows the results of the evaluation, where graphs are built on the SFU corpus and used on the RST dataset. As can be seen, the measures of precision, recall, and F-score are reported, along with their average value. The weighted average of F-score is also provided, taking into account the distribution of relation instances that contain the cues in the test set. This metric is provided while bearing in mind that the test set may not be an accurate representative of the general distribution of relations. According to the results, on average, we were able to classify CIRCUMSTANCE with an F-score of 0.66 in the SFU review dataset, while the weighted average of F-score is 0.71. In addition, an average F-score of 0.68 and a weighted average of 0.69 are achieved for the RST corpus.

9.5 Discussion

The use of syntactic context to disambiguate lexical cues has been shown to be useful to disambiguate cues in lexically oriented relations (e.g., PDTB relations). In this study, we have focused our efforts on RST annotated corpora and explored the potential of syntactic context for cue disambiguation in the RST framework. We have demonstrated that syntactic features can be of great value in the classification of explicit rhetorical relations. In addition, unlike the majority of prior studies on cue disambiguation, we encoded the syntactic context of cues in the forms of graphs. This graph-based approach was expected to provide a more generalizable and effective approach.

To verify this hypothesis, we conducted experiments to compare our graph-based model

Table 9.2: Classification results of a ten-fold cross validation on the SFU corpus. Probability values used as the underlying feature set are inferred from the graph models built on the RST corpus.

Cue	Precision	Recall	F-score
When	0.55	0.79	0.65
After	0.46	0.56	0.51
On	0.73	0.75	0.74
Before	0.62	0.60	0.61
With	0.76	0.80	0.78
Out	0.60	0.54	0.57
As	0.71	0.82	0.76
Average	0.63	0.69	0.66
Weighted Average			0.71

Table 9.3: Classification results of a ten-fold cross validation on the RST dataset. Probability values used as the underlying feature set are inferred from the graph models built on the SFU corpus.

Cue	Precision	Recall	F-score
When	0.62	0.69	0.65
After	0.61	0.57	0.59
On	0.77	0.81	0.79
Before	0.70	0.40	0.51
With	0.77	0.81	0.79
Out	0.75	0.70	0.73
As	0.73	0.67	0.70
Average	0.71	0.67	0.68
Weighted Average			0.69

with the concrete usage of syntactic features, where the POS tag and dependency path representations were utilized directly. A logistic model was first trained on the RST corpus and tested on the SFU dataset (see Table 9.4), and then trained on the SFU corpus and tested on the RST corpus (see Table 9.5). The results are consistently lower for all of the three measures, confirming the superiority of our approach. Due to the lack of prior work on cue disambiguation in the context of RST, instead of using a baseline, this comparison is made to highlight the contribution of our work.

Based on the results of our proposed approach, it could be seen that the three lexical cues of *when*, *after*, and *before* have the lowest performance in the RST corpus (see Table 9.3). They are also among the four cues with lowest F-score in the SFU dataset (see Table 9.2). This finding could be attributed to the fact that, for these three cues, the corresponding datasets were among the smallest cue sets. Possibly more importantly, these three cues can function as temporal indicators, which may make it particularly difficult to disambiguate them. For example, consider the following instances extracted from the SFU

Table 9.4: Classification results on the SFU corpus when the syntactic features are used directly to train a model on the RST corpus

Cue	Precision	Recall	F-score
When	0.50	0.41	0.45
After	0.39	0.15	0.22
On	0.65	0.28	0.39
Before	0.52	0.49	0.51
With	0.53	0.34	0.41
Out	0.51	0.19	0.28
As	0.49	0.24	0.32
Average	0.51	0.30	0.37

corpus:

- Positive instance:
  When I have time to kill between flights, I like to wander through and browse

- Negative instance:
  I was surprised *when* he told me that all the equipment was standard even on the base model

The first sentence is an instance of the CIRCUMSTANCE relation signaled by *when*, while in the second one, *when* implies the temporal aspect of the sentence and is not signaling CIRCUMSTANCE. We expect that certain linguistic and contextual features associated with the text might be useful in disambiguation of such lexical cues, such as the verb tense and/or the attributes of the agents of different clauses in a sentence. Further studies are required to explore these features.

The two underlying corpora used in this study are from very different text genres. In addition, since RST places emphasis on the writer's intentions and the effect of the relation on the reader [24], RST annotations are inherently subjective and are based on the readers' understanding of the text [24]. Therefore, there can be differences across the two corpora due to potentially different knowledge possessed by each set of annotators regarding the culture, situation, and language that the text represents [24]. Despite the genre disparity and difference in annotations, we obtained encouraging results using the proposed model. However, the results are expected to improve when the models are built on corpora from similar genres and are annotated using ground truth rules.

Table 9.5: Classification results on the RST corpus when the syntactic features are used directly to train a model on the SFU corpus

Cue	Precision	Recall	F-score
When	0.53	0.65	0.58
After	0.31	0.14	0.20
On	0.36	0.17	0.23
Before	0.33	0.22	0.26
With	0.62	0.20	0.30
Out	0.57	0.53	0.55
As	0.52	0.25	0.34
Average	0.52	0.25	0.35

9.6 Conclusion

Study and analysis of rhetorical relations, as the building blocks of coherence in discourse, can contribute toward the development of sophisticated linguistic applications and algorithms. With the aim of facilitating automatic discovery of explicit rhetorical relations in text, we developed an algorithm to first detect potential lexical cues and to later disambiguate them by predicting the relation.

An altered version of TF-IDF was used to extract the potential cues, and a graph-based model built on syntactic features was used to address the cue disambiguation task. Overall, the evaluation results indicate the effectiveness of syntactic features in the disambiguation of lexical cues and prediction of explicit rhetorical relations across different genres. Our experiments revealed the superiority of our approach of encoding such syntactic features in a probabilistic graph compared to their direct usage.

As mentioned earlier, this study is our first attempt toward the identification of rationales in text. A rationale is an explanation of the reasons underlying decisions, conclusions, and interpretations. Prior studies on rationale articulation and sharing suggest that it contributes to quality control, knowledge management, and knowledge reuse [30, 29]. However, there have been only a few attempts in applying computational techniques to identify rationales from ill-structured text [1, 5, 2]. Our future research efforts are geared toward the development of algorithms to detect lightly-signaled and implicit relations and to further explore the potential and limitations of using such rhetorical relations in the identification of rationales in the text of discourse. In addition, we plan to evaluate the extensibility of our approach to PDTB and to directly compare it with a proper baseline from among earlier studies on PDTB.

Bibliography

[1] Or Biran and Owen Rambow. Identifying justifications in written dialogs. In *Proceedings of the IEEE International Conference on Semantic Computing*, pages 162–168, 2011.

[2] Filip Boltužić and Jan Šnajder. Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 49–58, 2014.

[3] Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. Building a discourse tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, pages 1–10, 2001.

[4] Iria da Cunha. A symbolic corpus-based approach to detect and solve the ambiguity of discourse markers. *Research in Computing Science*, 70:93–104, 2013.

[5] Debanjan Ghosh, Smaranda Muresan, Nina Wacholder, Mark Aakhus, and Matthew Mitsui. Analyzing argumentative discourse units in online interactions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 39–48, 2014.

[6] Ahmed Hassan, Vahed Qazvinian, and Dragomir Radev. What's with the attitude? Identifying sentences with attitude in online discussions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1245–1255, 2010.

[7] Julia Hirschberg and Diane Litman. Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*, 19(3):501–530, 1993.

[8] J.R. Hobbs. *Literature and Cognition*. Center for the Study of Language and Information - Lecture Notes. Cambridge University Press, 1990.

[9] Eduard H. Hovy. Automated discourse generation using discourse structure relations. *Artificial Intelligence*, 63(1-2):341–385, 1993.

[10] Syeed Ibn Faiz and Robert Mercer. Identifying explicit discourse connectives in text. In *Advances in Artificial Intelligence*, pages 64–76. 2013.

[11] Taraneh Khazaei and Lu Xiao. Corpus-based analysis of rhetorical relations: A study of lexical cues. In *Proceedings of the IEEE Conference on Semantic Computing*, pages 417–423, 2015.

[12] Dan Klein and Christopher D. Manning. Accurate unlexicalized parsing. In *Proceedings of the Annual Meeting on Association for Computational Linguistics*, pages 423–430, 2003.

[13] Alistair Knott and Ted Sanders. The classification of coherence relations and their linguistic markers: An exploration of two languages. *Journal of Pragmatics*, 30(2):135 – 175, 1998.

[14] Alex Lascarides and Nicholas Asher. Temporal interpretation, discourse relations and commonsense entailment. *Linguistics and Philosophy*, 16(5):437–493, 1993.

[15] Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, 20:151–184, 2014.

[16] William C. Mann and Sandra A. Thompson. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988.

[17] Daniel Marcu. The rhetorical parsing of unrestricted texts: A surface-based approach. *Computational Linguistics*, 26(3):395–448, 2000.

[18] Thomas Meyer, Andrei Popescu-Belis, Sandrine Zufferey, and Bruno Cartoni. Multilingual annotation and disambiguation of discourse connectives for machine translation. In *SIGdial Meeting on Discourse and Dialogue*, pages 194–203, 2011.

[19] Eleni Miltsakaki, Nikhil Dinesh, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. Experiments on sense annotations and sense disambiguation of discourse connectives. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, 2005.

[20] Emily Pitler and Ani Nenkova. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP Conference (Short Papers)*, pages 13–16, 2009.

[21] Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The Penn Discourse TreeBank 2.0. In *Proceedings of the Conference on Language Resources and Evaluation*, pages 2961–2968, 2008.

[22] Prasad Rashmi, Bonnie Webber, and Aravind Joshi. Reflections on the Penn Discourse TreeBank, comparable corpora, and complementary annotation. *Computational Linguistics*, 40(4):921–950, 2014.

[23] Radu Soricut and Daniel Marcu. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 149–156, 2003.

[24] Maite Taboada. Discourse markers as signals (or not) of rhetorical relations. *Journal of Pragmatics*, 38(4):567 – 592, 2006.

[25] Maite Taboada, Caroline Anthony, and Kimberly Voll. Methods for creating semantic orientation dictionaries. In *Proceedings of the Conference on Language Resources and Evaluation*, pages 427–432, 2006.

[26] Yannick Versley. Subgraph-based classification of explicit and implicit discourse relations. In *Proceedings of the International Conference on Computational Semantics*, pages 264–275, 2013.

[27] Ben Wellner, James Pustejovsky, Catherine Havasi, Anna Rumshisky, and Roser Saurí. Classification of discourse coherence relations: An exploratory study using multiple knowledge sources. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, pages 117–125, 2006.

[28] Lu Xiao. Do members converge to similar reasoning styles in teamwork? A study of shared rationales in small team activities. In *Proceedings of the iConference*, pages 524–530, 2013.

[29] Lu Xiao. The effects of a shared free form rationale space in collaborative learning activities. *Journal of Systems and Software*, 86(7):1727 – 1737, 2013.

[30] Lu Xiao. Effects of rationale awareness in online ideation crowdsourcing tasks. *Journal of the Association for Information Science and Technology*, 65(8):1707–1720, 2014.

Chapter 10

ProjectTales: A Proof of Concept

10.1 Introduction

In the last few decades, many software programs have been designed and developed to address various aspects of project management such as estimating a project's effort [4], managing collaborative activities [13], and monitoring projects' changes and risks [7]. In this study, we are interested in assisting project managers with making effective project-related change decisions. Projects are not conducted in vacuum and are normally affected by different dynamic factors such as availability of budgeted resources, level of priority in the organization, and project members. Hence, it is expected that project managers commonly need to make modifications to projects during the management process.

Researchers have explored computer supported means for helping project managers to make valid decisions when projects need to be adjusted to the new situation. For example, Karvonen (1998) presented a computer supported management process to illustrate how computer systems could support a project manager's decision-making in a change situation during a delivery project as well as in the continuous business process improvement of a company. Sauve et al. (2008) presented a method to evaluate the risk exposure associated with a change to be made to the infrastructure and services in IT service management. With the risk exposure metric, this method automatically assigns priorities to changes.

Our approach supports project managers in making change decisions by presenting interactive visual representations of the change history of the previous projects. Our assumption is that when facing a change situation, project managers may make more effective

A version of this chapter has been published in the *proceedings of the iConference* as a poster paper.

decisions if they are aware of rationales and decisions of the previous projects that correspond to the similar situations as well as the impact of those decisions on the projects. Therefore, our design of the interactive visualization tool has focused on presenting and making associations 1) among the history of previous projects' changes, their causes, and their decisions' rationales; and 2) between the decisions and the project status as an indicator of the decisions' impact on the project. In this poster paper, we describe our current prototype system, ProjectTales. In the remaining sections, we first discuss related work and then present the design rationales of the prototype. We conclude with our user evaluation plans and a summary of the research.

10.2 Related Work

After surveying project management tools such as TeamSpace and TeamSCOPE, Smith et al. (2005) proposed design guidelines of a project management tool for distributed design teams. They argued that to facilitate reuse of project management knowledge from previous projects, it is important to archive both projects' product knowledge and their process-related knowledge. To illustrate their configuration approach of project management information systems, Bērziša and Grabis (2011) presented how knowledge of previous change requests can be combined with a request in the current project. In their approach, a request has two configurations in the system: the attributes of the request (e.g., priority, description, remaining, due date) and the status workflows (e.g., closed, open, agreed). When a request is made in the current project, the system will provide suggested descriptions of the attributes and status workflows based on the previous projects, i.e., the project management knowledge repository.

Compared to the system presented by Bērziša and Grabis (2011), which facilitates the configuration of current changes based on the history of previous projects, our system focuses on facilitating project managers' decision-making processes using the change knowledge from the previous projects. In addition, we use interactive visualization techniques to let project managers browse and extract change information relevant to the current project situation and to reason about the association between the changes and their impact on the projects. We discuss our prototype in more detail in the next section.

10.3 ProjectTales: An interactive visualization tool

10.3.1 Database and Its Design Rationale

Our prototype, so called ProjectTales, is designed and built upon the open source software EGroupware's (http://www.egroupware.org/) history database. EGroupware is intended for businesses to manage contacts, appointments, projects, and to-do lists. It is an actively used software program with a reliable user reputation. For example, its user rating is 4.2 out of 5 on SourceForge.net (164 votes), and on its enterprise collaboration website, there are 9352 topics and 27071 posts for EGroupware users' discussion forum as of Sept. 13, 2013. We thus understand that EGroupware's history database has been widely accepted and used in real-world project management cases, and so we used it as the underlying database for ProjectTales.

EGroupware's history database provides diverse attributes for projects (e.g., title, description, priority, used budget) and several attributes for changes (e.g., project, changed project attribute, timestamp). We modified this history database to include the causes of the changes. With the causes available to project managers, they can easily retrieve the change decisions that correspond to the same causes as the current situation. When modifying a project, project managers are asked to assign one or more causes to the change. These causes can be selected from a pre-specified list of potential causes extracted from the literature [2, 8, 12], including the external causes (e.g., political, economic, customer needs, market force) as well as the internal causes (e.g., staff or budget shortage, strategic decisions, quality control). In order to allow project managers tailor the causes to their specific organization situation, they are allowed to add new ones to the list.

Arguing for the importance of providing the change rationales to the project manager, we also included this information in the database. In our research, we define the change rationale as the information that provides justification of the change decision. Social psychologists have investigated the kinds of information in communication that influence people's attitudes [1, 6]. These studies showed that the shared information that is plausible and logical and adds something new to the issue is more likely to cause attitude change [5, 3]. Fabrigar, Priester, Petty, and Wegener (1998) have found when people had higher access to different attitudes, the message's argument quality had a greater impact on persuasion. We believe that presenting not only the change decisions but also their rationales can benefit project managers when deciding the current situation. Xiao and Carroll's (2013) study indicates that individuals' reasoning skills are affected by sharing task rationale explicitly in group activities. By archiving and sharing the change rationales across the organization's projects, it is expected that the project managers will improve their reasoning skills

on making decisions in the long term. The improved reasoning skills can positively affect the quality of change decisions since decisions are made after better reasoning processes.

10.3.2 Interface Design

ProjectTales interface is divided into two separate components arranged vertically on the screen: the history overview component and the project detailed view component. Each component of the system then provides some coordinated views of different dimensions of the underlying dataset, supporting project managers in yielding deeper insight of the history data [10]. Figure 1 shows a screenshot of the system

History Overview Component

The history overview component consists of three different views, namely the history grid, the cause bar, and the rationale bar. The history grid is designed to provide a visual overview of the changes made in different attributes of different projects. In this view, each row represents a project in the database and each column represents an attribute of the

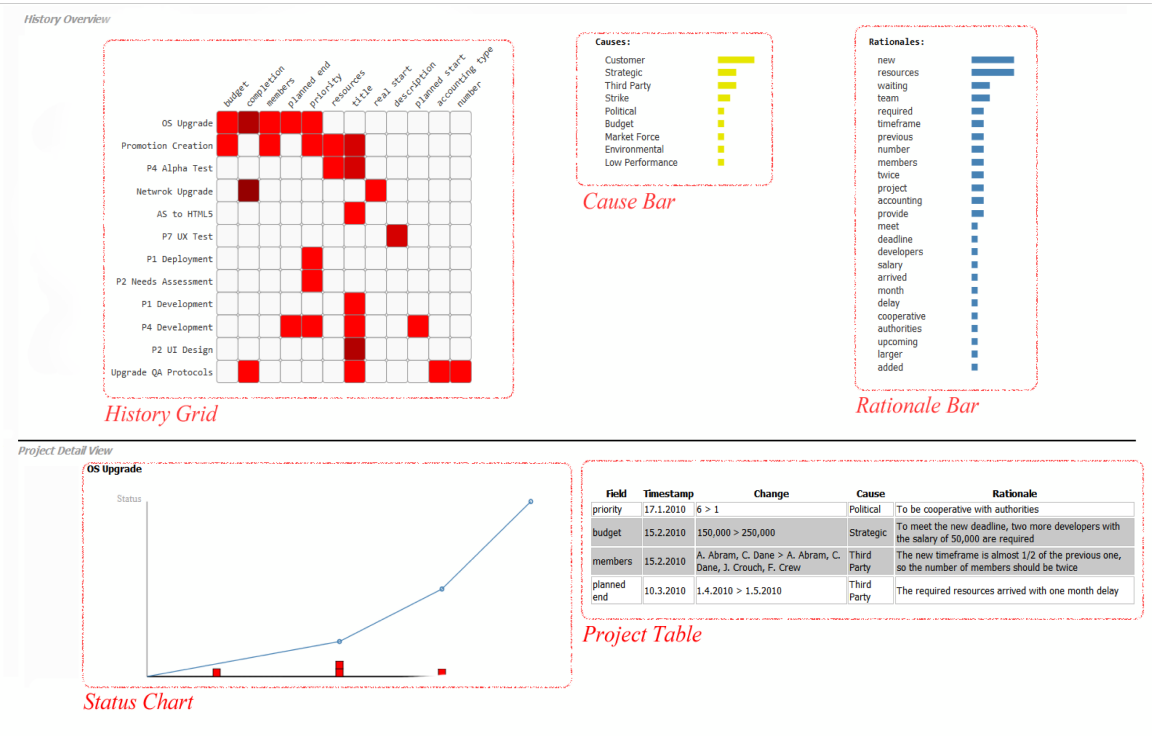


Figure 10.1: A screenshot of ProjectTales

project. Since colour is a pre-attentive visual feature and can be separately decoded from the spatial position by the human visual system [11], colour coding is used to represent the number of changes in different attributes of projects. When there has been no change, the cell is coloured with a light gray. In case of a change, the number of changes is encoded using the luminance channel of the red color; thus, cells with high change frequency appear darker and the ones with low change frequency are shown brighter.

The history grid is further enriched with an interaction mechanism, allowing project managers to sort the projects according to the number of changes in a specific attribute by clicking on the attribute label. Similarly, clicking on a project label causes the attribute list to be sorted according to the number of changes for that project. Such a visual representation allows project managers to quickly gain an overview of the change frequency across different projects and different attributes, and to further explore this information for projects and attributes of interest using the sort feature.

In addition to the history grid, the history overview component offers two vertical histograms to visually depict an overview of the causes and rationales for the changes. The cause bar represents a sorted view of the causes from the most frequent one to the least frequent, whereas the rationale bar shows the most frequent words used in the change rationales. Since stop words are of no value when exploring the prominent words in rationales, they are eliminated from the rationale bar. The history grid is highly linked with the cause and the rationale bar via a linking and brushing technique. Hovering the cursor over each cell in the history grid highlights the corresponding causes and rationales in both histograms. In addition, if project managers hover the cursor over a cause or a rationale, the corresponding cells in the history grid get highlighted.

Project Detailed View Component

When exploring the change history overview, project managers may be interested to focus on a particular project in detail. Double clicking on a project label loads the project detailed view component with two linked views of the status line chart and the project table. In the status line chart, the x-axis represents the duration of the project as a timeline. This axis is augmented with red glyphs, each representing a change that has happened to the project in the specific time. Presenting changes on a timeline may allow project managers to investigate the sequence of changes and see how a particular change has triggered the following changes. The y-axis of the chart represent the status of the project, allowing project managers to assess the effects of a particular change or a series of changes on the project status.

Finally, the project table provides project managers with the detailed information of

changes, including the changed attribute, time of the change, old and new values, causes of the change, and the full rationale, in a textual format. These two views are also linked with linking and brushing techniques, as hovering the mouse cursor over a glyph in the chart or a row in the table causes the corresponding representation of change to get highlighted in the other view. These two views allow a project manager to identify a potentially interesting change in one view and then detect the same change in the other view to perform further exploration.

10.4 Evaluation

We are currently conducting an evaluation study of the prototype in a controlled laboratory setting. This study is designed as a within-subject study, in which participants are asked to perform a different decision-making task with each interface (ProjectTales and a table-based baseline system). The two tasks are designed to have the same complexity level and they consist of three subtasks about retrieving and interpreting information about changes and change rationales. We have been recruiting projects managers to use ProjectTales and compare it with the baseline system.

10.5 Conclusion

Change management "is an integral process related to all project internal and external factors, influencing project changes; to possible change forecast; to identification of already occurred changes; to planning preventive impacts; to coordination of changes across the entire project" [9]. In this poster paper, we presented a software prototype with the main purpose of facilitating project managers to make more efficient and effective decisions when deciding a possible change. With our current design, ProjectTales provides project managers with the ability to browse, explore, and gain an insight into the history of causes, decisions, and rationales of the previous changes through interactive visualization techniques. Our underlying design rationale is that archiving the change causes and rationales are as important as the change decisions themselves for future reuse. We are currently conducting an evaluation study to examine the usability and our rationales of the design.

Bibliography

[1] W. A. Barnard, W.A. Mason, and M. L. Ceynar. Level of interaction and reciprocal influence in supportive and critical male discussion groups. *Journal of Social Psychology*, 133:833–838, 1993.

[2] Dov Dvir and Thomas Lechler. Plans are nothing, changing plans is everything: The impact of changes on project success. *Research Policy*, 33(1):1 – 15, 2004.

[3] M. R. Leippe and R. A Elkin. When motives clash: Issue involvement response and involvement as determinants of persuasion. *Journal of Personality and Social Psychology*, 52:269–278, 1987.

[4] B. Peischl, M. Nica, M. Zanker, and Wolfgang Schmid. Recommending effort estimation methods for software project management. In *Proceedings of the IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies*, volume 3, pages 77–80, 2009.

[5] R. E. Petty and J. T. Cacioppo. The effects of involvement on responses to argument quantity and quality: Central and peripheral routes to persuasion. *Journal of Personality and Social Psychology*, 46:69–81, 1984.

[6] Richard E Petty and Duane T. Wegener. Attitude change: Multiple roles for persuasion variables. *Handbook of Social Psychology*, 1-2:1998, 1998.

[7] J. L. Smith, S. A. Bohner, and S. D. McCrickard. Project management for the 21st century: Supporting collaborative design through risk analysis. In *Proceedings of the Annual Southeast Regional Conference*, volume 2, pages 300–305, 2005.

[8] Wolfgang Steffens, Miia Martinsuo, and Karlos Artto. Change decisions in product development projects. *International Journal of Project Management*, 25(7):702 – 713, 2007.

[9] Vladimir Voropajev. Change management-A key integrative function of PM in transition economies. *International Journal of Project Management*, 16(1):15 – 19, 1998.

[10] Michelle Q. Wang Baldonado, Allison Woodruff, and Allan Kuchinsky. Guidelines for using multiple views in information visualization. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, pages 110–119, 2000.

[11] C. Ware. *Information Visualization: Perception for Design.* Morgan Kaufmann, San Francisco, CA, second edition, 2004.

[12] Chao Wu, Ting Hsieh, and Wen Cheng. Statistical analysis of causes for design change in highway construction on taiwan. *International Journal of Project Management*, 23(7):554 – 563, 2005.

[13] Shaoke Zhang, Chen Zhao, Paul Moody, Qinying Liao, and Qiang Zhang. Lightweight collaborative activity patterns in project management. In *Proceedings of the International Conference on Engineering Psychology and Cognitive Ergonomics*, pages 465–473, 2007.

Chapter 11

Conclusions

During the last few decades, the Web has evolved from a platform to merely consume content to a place where people can create, share, modify, and discuss contents of various forms. This advancement has led to the phenomenon of the social Web that significantly impacts our day to day lives these days. The unprecedented levels of intelligence generated from such large-scale online interactions are one of the notable effects of the social Web. With the aim of contributing to the fields of web-based collective and business intelligence, we first identified the two social phenomena of privacy and persuasion to be of incredible influence on the intelligence paradigm (Chapter 1).

To further study the concept of social privacy, we first conducted a literature review (Chapter 2). Then we studied the relations of various social media data types with users' privacy preferences and proposed a model to detect the social media profiles for which the privacy dichotomy holds (Chapters 3, 4, and 5). We then focused on the phenomenon of online reasoning by surveying the relevant studies (Chapter 6). The mechanisms and strategies behind online persuasion were then analyzed and discussed 7. Given the value of rhetorical relations in studying persuasion in the text, a novel approach was proposed to identify and disambiguate explicit rhetorical relations from text (Chapters 8 and 9). Finally, an interactive visualization tool was offered to depict the values of the intelligence models and frameworks to further promote intelligence (Chapter 10). The summary of the dissertation, the primary contributions of this research, as well as the potential future directions are outlined in the remainder of this Chapter.

11.1 Summary of Research

The following highlights the summary of our low-level contributions in terms of the objectives declared in Chapter 1. These items mainly contribute to the fields of artificial

intelligence and information systems. In particular, these findings can be of value to the areas of machine learning, data mining, and natural language processing.

Identification of research gaps in the context of social privacy: Compared to the other solutions to manage the privacy dichotomy issue in social media, automatic identification of privacy preferences have shown modest success. Therefore, we reviewed the existing literature on the algorithms that take advantage of users' social footprints to characterize their privacy preferences. We categorized and reviewed the existing studies on privacy preference inference according to the data type of focus, namely demographics and profile features, social context and network features, as well as the shared content. The potential and limitations of the approaches were further discussed, where a set of gaps were identified. For instance, while the majority of the studies are focused on privacy configurations that are specific to a single social media platform, less attention has been paid to general user modeling and characterization. In addition, further research is required to study a wider range of data types available in social media. As well, the majority of the studies have utilized supervised methods. However, supervised techniques may not be feasible in the context of social media, where the labeled information often constitutes a very small portion of the available data. Unsupervised and semi-supervised techniques then can then be appealing alternatives in this context. Our research efforts to address the privacy dichotomy problem were informed by the results of the literature analysis. For instance, we incorporated a variety of data types in our approach and employed an unsupervised collaborative filtering approach.

Identification of research gaps in the broader context of discourse-centric collective intelligence: The proliferative body of literature analyzing dialogue-based collective intelligence is promising. Our analysis of this body of work revealed the lack of focus on task-oriented environments, the lack of sophisticated methods to analyze the impact of group interaction processes on each other, as well as the lack of focus on the study of the impact of group interaction processes on participants over time. Similar to our research direction on social privacy, our decisions to study reasoning traces were informed by the reviewed literature. For instance, we focused on a task-oriented environment that is primarily focused on persuasion and belief-change and analyzed the factors that can impact participants' beliefs.

Understanding the behaviours and characteristics of users with different privacy behaviours and attitudes: As an initial step toward understanding the behaviours of users with different privacy behaviours, we examined whether profile attributes of Twitter users with varying privacy settings are configured differently. As a result of the analysis, a set of features was identified and used to predict user privacy settings. For our best classifier,

we obtained an F-score of 0.71, which outperformed the baselines considerably. Then, we ran a series of experiments to examine if privacy preferences are localized in social media. As a result, we found a set of clues that users privacy preferences are similar to the privacy behaviour of their social contacts, signaling that privacy preferences are localized in social networks. Finally, the neighbourhood context was used to characterize the attributes of privacy-unaware people. In particular, we found privacy-unaware users to publish more personal and private information compared to those who intentionally follow the public setting.

Understanding the mechanisms behind online persuasion: By analyzing a set of human-annotated comments extracted from Reddit.com, we explored different dimensions of the language, the temporal aspects of the comments, as well as the attributes of the participating users and their relations to the persuasion process. Throughout the analysis, we found a large set of attributes that are specific to persuasive comments and examined their predictive power to detect the comments that will persuade. The majority of the features are associated with the persuasive impact of various components of the language including the psychological attributes of the language and writing quality and sophistication. These features resulted in a classification of persuasive and non-persuasive comments with a reasonable performance of 75%. A preliminary examination of the user explanations as to why and how their belief is changed allowed an analysis of the perceived reasons for persuasion and their comparison to the actual ones found in the earlier analysis of the comments. A comparison of our findings captured from an online platformed with the persuasion literature in traditional settings uncovered a set of similarities and differences between the two processes.

Facilitating identification and disambiguation of linguistic relations that are of importance in persuasion: Analysis and study of justifications and rationales can be very useful in the understanding of online argumentation and reasoning. Given the value of a subset of rhetorical relations in the identification of rationales, we proposed a novel approach for automatic detection of such rhetorical relations. We first conducted a corpus-based analysis to derive a set of n-grams as potential lexical cues that can signal the rhetorical relations of focus. These cues were then utilized in graph-based probabilistic models to determine the syntactic context in which the cue is signaling the presence of a particular relation. Evaluation results were reported for various cues of the CIRCUMSTANCE relation, confirming the value of syntactic features for the task of cue disambiguation. In addition, using a graph to encode syntactic information was shown to be a more generalizable and effective approach compared to the direct usage of syntactic features. Given that the underlying RST annotated corpora are not in the context of social media, we also conducted a cross-corpus study

to understand whether the identified lexical cues are genre-specific.

Development of a proof of concept: Besides the analytical benefits that studying persuasion can provide, the models and frameworks can be effectively harnessed to design and develop interfaces that can further promote intelligence. To demonstrate this benefit in a particular business intelligence context, we designed and developed an interactive visualization system called ProjectTales. This system provides project managers with the ability to browse, explore, and gain an insight into the rationales of the previous changes made in the projects. Our underlying design rationale was that archiving the change causes and rationales are as important as the change decisions themselves for future reuse. However, designing such a system would not be possible without an underlying model that can effectively and efficiently detect rationales from internal corporate communications.

11.2 Contributions

This section summarizes the high-level contributions of this dissertation by highlighting the fields it can contribute to and the benefits it can provide to these different research communities.

Studies on the concepts of intelligence, privacy, and persuasion within social interactions have long been conducted in a variety of disciplines such as sociology, psychology, and anthropology. This dissertation contributes toward these fields by re-examining some assumptions and findings regarding such social interactions in the context of online social media. In addition, new findings are presented regarding the structure and content of massive online social interactions that have been studied and observed in small-scale traditional studies. Besides, artificial intelligence and information system communities can benefit from this research. The gaps and patterns found in the literature, the proposed models, and the feature sets, along with their evaluation and analysis, can all provide new insights to these disciplines and suggest new paths for future research activities. Finally, our findings can also inform the design of the next-generation intelligence interfaces and platforms, contributing to the fields of information visualization and human-computer interaction.

The broad concern of this research has been to study and promote the intelligence that arises from the social Web. Tom Malone of the Massachuset Institute of Technology formulates the central research question of this area as follows: “How can people and computers be connected so that - collectively - they act more intelligently than any person, group, or computer has ever done before?” [4]. Our research efforts on the identification and analysis of social privacy and persuasion as the building blocks of the social Web intelligence, along with its subsequent contributions to the disciplines mentioned above, is a step forward to-

ward addressing this overarching research question.

11.3 Future Directions

Detailed suggestions to improve the proposed methods and models are provided in the future work section of the individual chapters. Whenever possible, these suggestions are taken into account in the subsequent chapters. Therefore, to get a more specific discussion of future work with regards to a particular aspect of the dissertation, the readers are referred to these sections. However, there exist several high-level research directions that can provide valuable future extensions to this thesis:

Verification of the findings across multiple platforms and datasets: In this dissertation, we focused on a few social data sets collected from specific platforms, such as Twitter and Reddit. For instance, our work on social privacy is mainly focused on two different sets of data from Twitter, while online persuasion is studied based on one set of data collected from a particular subreddit. However, other data collection strategies can be employed to retrieve other sets of representative samples from the same platforms, ensuring the generalizability of our approach across different parts of the network. In addition, the design and technical features of a given social platform can constrain, distort, and shape user behavior [3]. In a recent study by Malik and Pfeffer [3], the empirical evidence is provided across multiple social platforms, verifying theoretical concerns about the possible dramatics effects of platform design on user behaviour. Therefore, regardless of our valuable findings and contributions throughout this dissertation, the generalizability and external validity of the conclusions would not be sanctioned without comprehensive cross-platform studies. As such, conducting the same type of analysis and experimentation on other active social media platforms (e.g., Facebook and Instagram) is warranted.

Moving beyond a binary specification for privacy preferences: Characterizing privacy preferences based on social data is a difficult and challenging task. Therefore, as a starting point to fulfill this goal, we focused on a platform where privacy preferences are specified on a binary scale. Therefore, in our underlying dataset, users are either privacy-concerned or privacy-unconcerned. However, this simplifying assumption is not consistent with how people perceive privacy. In fact, depending on the context, users may choose to use different strategies to protect their privacy. For instance, in the study conducted by Wisniewski et al. [5], six different privacy protection strategies are identified in the context of Facebook. Studying such privacy protection strategies and their relations to users' characteristics in the network can provide further insight on how and when to use one's publicly available data for intelligence and analytics purposes.

Analyzing the effects of contextual factors in online persuasion: Persuasion studies have shown that there are various contextual factors that influence the persuasiveness of an act [1]. The importance of studying the platform effects as one of many contextual factors is discussed earlier in this section. As a starting point towards a comprehensive computational model of the detection of online persuasive comments, we have focused on the root comments provided in Reddit. However, the comments that occur in the form of threaded discussions can also lead to persuasion. The underlying dynamics of these conversational comments can indeed be influential in the persuasion process. Therefore, the analysis and study of the conversational content in which a belief change occur can shed light on a new set of strategies and mechanisms behind online persuasion.

Studying the predictive power of the rhetorical relations of focus to identify rationales: The three rhetorical relations of CIRCUMSTANCE, EVALUATION, and ELABORATION are shown to be commonly present in rationales [6]. By relying on this earlier work, we focused on research efforts to study, identify, and disambiguate these three rhetorical relations. However, to verify the power of rhetorical relations in the automatic identification of rationales, further experiments are required in the context of social media communications. This verification, though, hinges on another future research effort to develop novel methods for automatic identification of implicit and lightly signaled instances of these rhetorical relations.

Developing a set of design guidelines for intelligence tools: Studying the phenomenon of social Web intelligence can have implications for designing novel interfaces to further promote intelligence. Even though we designed and developed ProjectTales to showcase such benefits, user evaluations of the tool are included in our future research plans due to a variety of constraints. For instance, recruiting project managers to conduct the study proved to be a time consuming and challenging task. However, such a user study can validate the usefulness of rationale identification models in the context of business intelligence. In addition, it can validate interactive and visualization design choices. More importantly, it can provide insights and suggestions to develop a set of guidelines for the design of effective intelligence tools. Therefore, such design-oriented research [2] will be considered in our future research. In design-oriented research, the main contribution is the knowledge gained from studying and evaluating a designed artifact.

Bibliography

[1] Robert Cialdini. *Influence: The Psychology of Persuasion*. New York, NY: Collins, 2007.

[2] Daniel Fallman. research-oriented design versus design oriented research. In *Proceedings of Nordes: Nordic Design Research Conference*, pages 1–3, 2005.

[3] Momin Malik and Jürgen Pfeffer. Identifying platform effects in social media data. In *Proceedings of the AAAI Conference on Web and Social Media*, pages 241–249, 2016.

[4] MIT. MIT Center of Collective Intelligence. `http://cci.mit.edu/`.

[5] Pamela Wisniewski, Bart P Knijnenburg, and Heather Richter Lipford. Profiling Facebook users privacy behaviors. In *Symposium on Usable Privacy and Security*, 2014.

[6] Lu Xiao. Do members converge to similar reasoning styles in teamwork? A study of shared rationales in small team activities. In *Proceedings of the iConference*, pages 524–530, 2013.

Appendix A

Amazon Mechanical Turk Documentation

This appendix includes the formal approval received from the Human Research Ethics of Western University for the Amazon Mechanical Turk experiment. It also includes the experiment template that Amazon workers viewed and completed.

Western Research

Research Ethics

Western University Health Science Research Ethics Board
NMREB Delegated Initial Approval Notice

Principal Investigator: Dr. Lu Xiao
Department & Institution: Information and Media Studies\Faculty of Information & Media Studies,Western University

NMREB File Number: 106948
Study Title: Social Privacy - A Mechanical Turk Experiment
Sponsor:

NMREB Initial Approval Date: August 14, 2015
NMREB Expiry Date: August 14, 2016

Documents Approved and/or Received for Information:

Document Name	Comments	Version Date
Revised Western University Protocol		2015/07/22
Letter of Information & Consent	Information letter template - full version	2015/07/30
Other	Standalone document - short version of information letter	
Other	Tasks and survey - long version of information letter	2015/07/30
Other	Tasks and survey - short version of information letter	2015/07/30

The Western University Non-Medical Research Ethics Board (NMREB) has reviewed and approved the above named study, as of the NMREB Initial Approval Date noted above.

NMREB approval for this study remains valid until the NMREB Expiry Date noted above, conditional to timely submission and acceptance of NMREB Continuing Ethics Review.

The Western University NMREB operates in compliance with the Tri-Council Policy Statement Ethical Conduct for Research Involving Humans (TCPS2), the Ontario Personal Health Information Protection Act (PHIPA, 2004), and the applicable laws and regulations of Ontario.

Members of the NMREB who are named as Investigators in research studies do not participate in discussions related to, nor vote on such studies when they are presented to the REB.

The NMREB is registered with the U.S. Department of Health & Human Services under the IRB registration number IRB 00000941.

Ethics Officer to Contact for Further Information

____ Erika Basile ebasile@uwo.ca	___ Grace Kelly grace.kelly@uwo.ca	___Mina Mekhail mmekhail@uwo.ca	___Vikki Tran vikki.tran@uwo.ca

This is an official document. Please retain the original in your files.

Instructions

Please accept this hit **only you have an active Twitter account AND you have read the following paragraph about this research study and agree to participate. Please only complete one HIT from the group. Also, please answer all questions in the HIT before you submit the HIT to avoid any rejection**

In this study, you will answer two questions and complete two tasks. Just as any other Mechanical Turk task, you will earn the specified amount of money if your submission is approved. This task is part of an academic research project (Principle Investigator: Dr. Lu Xiao, Assistant Professor, Faculty of Information & Media Studies, the University of Western Ontario, Canada). Therefore, you need to be at least 18 years old to complete the task. Your submission will be collected and analyzed as research data if it is approved. Your confidentiality will be respected. If the results of the study are published, no information that discloses your identity will be released or published. If we find information we are required by law to disclose, we cannot guarantee confidentiality. We will strive to ensure the confidentiality of your research-related records. Absolute confidentiality cannot be guaranteed as we may have to disclose certain information under certain laws. If you have questions about your rights as a research subject you may contact: The Office of Research Ethics, The University of Western Ontario, 519-661-3036

The following study consists of three sections: A section with two questions about your privacy preferences and two categorization tasks.

Privacy Questions:

Privacy section asks questions regarding your privacy preference for sharing your Twitter data with organizations to recieve personalized offers.

Categorization Task 1:

For this task, you will be provided with some tweets. Given the tweets, you will then be asked to judge the privacy preference of the user who shared the tweet.

Categorization Task 2:

Similar to Task 1, you will be provided with some tweets. However, for this task, you will judge whether the tweets contain private and sensetive information by choosing one or more categories that are provided.

Desired Privacy:

1. What is you privacy preference in regard to your willingness to share your publicly available social media data with companies and organizations for personalization purposes?

- ○ I am not concerned about the use of my data at all
- ○ I am not concerned about the use of my data if certain conditions are held (e.g., data is being used anonymously)
- ○ I am extremely concerned about the use of my data

2. If you were to gain benefits (e.g., promotional offeres tailored to your current needs), what would be your privacy preference for sharing your publicly available social media data with companies?

- ○ I am not concerned about the use of my data at all

- ○ I am not concerned about the use of my data if certain conditions are held (e.g., data is being used anonymously)
- ○ I am extremely concerned about the use of my data

Categorization Task 1:

Given the tweet below, judge the privacy preference of the user who shared it:

Tweet: ${TWEET1_1}

- ○ Privacy Unconcerned: He/she is not concerned if his/her data is used by companies
- ○ Neutral/Objective: Neutral or objective tweets that reveal no information about the users' privacy preferences
- ○ Privacy Concerned: He/she is not willing to share his/her data with companies

Given the tweet below, judge the privacy preference of the user who shared it:

Tweet: ${TWEET1_2}

- ○ Privacy Unconcerned: He/she is not concerned if his/her data is used by companies
- ○ Neutral/Objective: Neutral or objective tweets that reveal no information about the users' privacy preferences
- ○ Privacy Concerned: He/she is not willing to share his/her data with companies

Given the tweet below, judge the privacy preference of the user who shared it:

Tweet: ${TWEET1_3}

- ○ Privacy Unconcerned: He/she is not concerned if his/her data is used by companies
- ○ Neutral/Objective: Neutral or objective tweets that reveal no information about the users' privacy preferences
- ○ Privacy Concerned: He/she is not willing to share his/her data with companies

Categorization Task 2:

Given the tweet below, please select the most appropriate category/categories:

Tweet: ${TWEET2_1}

- ☐ Location: Giving out someone's location information
- ☐ Medical: Revealing information about someone's medical condition
- ☐ Drug/Alcohol: Giving information about alcohol/drug use or revealing information under influence
- ☐ Emotion: Highly emotional content such as frustration
- ☐ Personal Attacks: Critical statements directed at a person
- ☐ Stereotyping: Ethical, racial, stereotypical references about a group
- ☐ Family/Association Details: Revealing information about family members
- ☐ Personal Details: Relationship status, sexual orientation, job/occupation, embarrassing or inappropriate content
- ☐ Personally Identifiable Information: e.g., SNN, credit card number, home address, birth date, etc.
- ☐ Neutral/Objective: Neutral or objective tweets that reveal no private or sensitive information

Given the tweet below, please select the most appropriate category/categories:

Tweet: ${TWEET2_2}

- ☐ Location: Giving out someone's location information
- ☐ Medical: Revealing information about someone's medical condition
- ☐ Drug/Alcohol: Giving information about alcohol/drug use or revealing information under influence
- ☐ Emotion: Highly emotional content such as frustration
- ☐ Personal Attacks: Critical statements directed at a person
- ☐ Stereotyping: Ethical, racial, stereotypical references about a group
- ☐ Family/Association Details: Revealing information about family members
- ☐ Personal Details: Relationship status, sexual orientation, job/occupation, embarrassing or inappropriate content
- ☐ Personally Identifiable Information: e.g., SNN, credit card number, home address, birth date, etc.
- ☐ Neutral/Objective: Neutral or objective tweets that reveal no private or sensitive information

Given the tweet below, please select the most appropriate category/categories:

Tweet: ${TWEET2_3}

- ☐ Location: Giving out someone's location information
- ☐ Medical: Revealing information about someone's medical condition
- ☐ Drug/Alcohol: Giving information about alcohol/drug use or revealing information under influence
- ☐ Emotion: Highly emotional content such as frustration
- ☐ Personal Attacks: Critical statements directed at a person
- ☐ Stereotyping: Ethical, racial, stereotypical references about a group
- ☐ Family/Association Details: Revealing information about family members
- ☐ Personal Details: Relationship status, sexual orientation, job/occupation, embarrassing or inappropriate content
- ☐ Personally Identifiable Information: e.g., SNN, credit card number, home address, birth date, etc.
- ☐ Neutral/Objective: Neutral or objective tweets that reveal no private or sensitive information

---

By submitting this HIT, you *confirm* that you are at least 18 years old and a registered user of Amazon's Mechanical Turk. Moreover, you *agree* that the work you submit here can be used in scientific publications that do not have any personal identifying information. You also *agree* that your Mturk ID will be replaced with an arbitrary alphanumeric code in the researchers' record, and then your work, along with other participants' work in this project, will be archived and available to share only for non-commercial purpose. If you are interested in the results of the study, you could check out our research lab's web site http://hii.fims.uwo.ca where we announce the new findings in the form of publications. Please note that typically about one to two years pass before an experiment is published.

Appendix B

Amazon Mechanical Turk Experiment and Results

The data collected from AMT was mainly intended for the evaluation of our hypothesis regarding the existence of homophily for privacy preferences (see Chapter 4). However, the collected data provides a valuable resource to further understand privacy attitudes and behaviours. This Appendix first describes the design of the study and then presents the statistics and the results of our preliminary analysis of the annotated data.

In AMT, workers can be asked to complete and submit Human Intelligence Tasks (HITs). For this experiment, each HIT provided Amazon Mechanical Turk workers with two questions regarding their own privacy preferences followed by two annotation tasks. For the first task, the workers were provided with three different tweets and were asked to judge the privacy preference of the author of the tweet. In particular, they were asked to judge whether the tweet implies that the author is privacy-concerned, privacy-unconcerned, or the content is neutral and objective. For the second task, the workers were given another set of tweets and were asked to judge whether the content is private and sensitive by choosing from among categories of sensitive information provided to them (e.g., medical information, personal detail, emotional content, etc.). These categories are adapted from the prior literature and are known to be what is considered as sensitive according to societal consensus [1]. One other category is provided, labeled as *neutral/objective*, to choose when the tweet does not contain any private or sensitive information. Each AMT worker was allowed to only submit one HIT. To see the HIT template, readers are referred to Appendix A.

The findings presented in this appendix are being drafted to be submitted to a relevant conference (e.g., *WWW* or *ICWSM*).

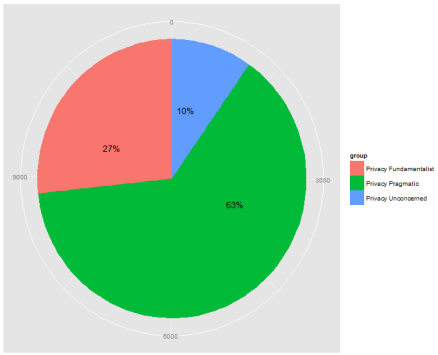


Figure B.1: The distrbution of AMT responses for the desired privacy.

We then published tweets to AMT to be categorized by the workers. These tweets are selected from our Twitter data described in Chapter 4. Our collected user set was first filtered only to include those users who tweet in English. Then, 1200 users with the lowest privacy ratio and 1200 users with the highest privacy ratio are extracted from the set. We then selected the five latest tweets published by these users that are not retweets or replies. We also ensured that these tweets include some text and are not generated automatically. This process resulted in the selection of 12K tweets published by 2400 different users that have either extremely low or extremely high privacy ratios. Each tweet posted to AMT is labeled by three different workers. Only when at least two workers agree on the same label, the annotation is considered qualified for further analysis. Otherwise, the data is excluded.

In the experiment, we first provided the workers with two questions to indicate their privacy preference in regards to their willingness to share their data with companies for personalization purposes. The workers could then choose one of the three possible given options (see Appendix A). These options are provided based on the three categories of privacy preferences determined through multiple privacy surveys, namely privacy fundamentalist, privacy pragmatic, and privacy unconcerned [2]. Figure B.1 shows the distribution of responses. According to the data, 27% of the workers mentioned that they are extremely concerned about their data being used, while 10% indicated no concern regarding their social media data being used by businesses. The majority of users, however, implied that they are not concerned if their data is used under certain privacy preserving conditions (e.g., anonymous use of data).

The workers were also asked to indicate their privacy preference for companies to use their data when they are given some benefits, such as promotional offers, in return. Figure B.2 shows the distribution of responses to this question. As can be seen, the number of workers who expressed extreme concern decreases, while the number of those who expressed less or no concern increases slightly. Our t-test analysis, however, shows no signif-

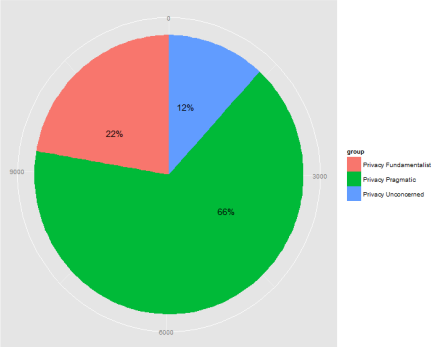


Figure B.2: The distrbution of AMT responses for the desired privacy when benefits are given in return.

icant differences between the two sets of responses (p-value > 0.05). This finding implies that privacy preferences and attitudes are not to be easily altered by external incentives.

The first annotation task asks users to judge the privacy preference of the tweet author. Figure B.3 shows the distribution of responses to this question. As can be seen, user judgments are extremely unbalanced. Only 1% of the users judged the tweet author to be privacy concerned. This issue might be due to a number of reasons. First, the majority of the selected tweets might not contain private and sensitive content, resulting in the selection of the first two labels most of the times. This hypothesis will be tested by the analysis of the second annotation task in the remainder of this document. Second, the study design, such as the placement of options and the content of the question, might have affected the results. Finally, non-expert annotators might not be suited for this task. In fact, our observation of the 153 tweets that are labeled under the *privacy concerned* category shows that instead of making judgments according to the semantics of the tweets, workers sought explicit indicators of privacy-related content. For instance, the following tweets are labeled categorized as tweets published by a privacy concerned user:

- He's mine i am his. Dont be calling him private and telling him to watch out who he is dating -___-t#ThatAnnoyingExHeHas.
- This just lays it out so simply. The #NSA has a weakness – it is up to us what we do now. #ResetTheNet
- You ask me personal questions but why should I even trust you with anything

The workers were then given three tweets and were asked to judge whether the content is private and sensitive by choosing from among nine categories of sensitive information and one *objective* category. The workers were allowed to choose as many labels as they

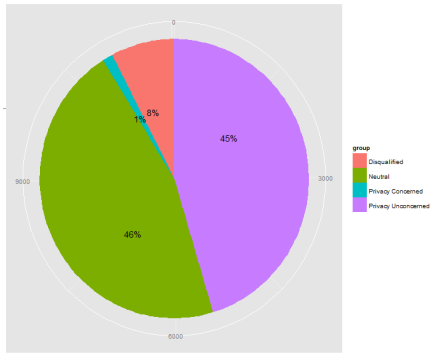


Figure B.3: The distrbution of responses for the first AMT annotation task.

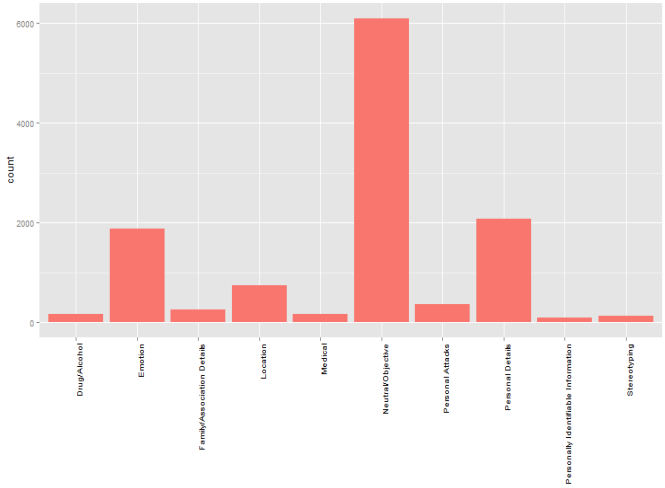


Figure B.4: The distrbution of responses for the second AMT annotation task.

found relevant. However, only if at least workers agree on the same category, the annotation is used for analysis. Even though the same tweet set are annotated in the first and the second task, we designed the study such that each worker is given a unique set of tweets across the two tasks. Figure B.4 shows how the responses are distributed among the categories. Even though a large number of results are associated with the objective category, almost half of the tweets are linked to at least one private and sensitive category. Among these categories, tweets that reveal personal details, emotions, and location information are more often seen compared to the other categories. Given that a considerable number of tweets contain private content, the unbalanced results of the first annotation task is unlikely due to the features of the underlying tweets and is probably the result of the study design. Providing instructions as of how to judge privacy preferences to AMT workers, as non-expert annotators, might result in a more balanced annotation set.

As mentioned earlier, from each user timeline, five tweets are annotated through AMT.

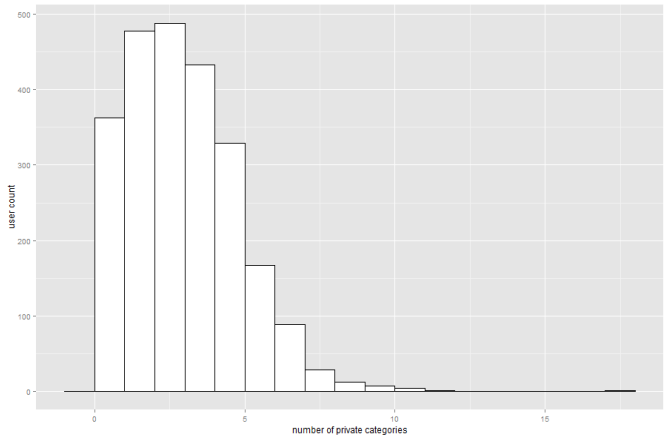


Figure B.5: shows the distribution of the number of private categories across different user timelines

We aggregated these annotations to characterize the amount of private information that is shared by each Twitter user. To do so, we counted the total number of private categories that are assigned to the tweets of the focal user. Figure B.5 shows the distribution of the number of private categories associated with different user timelines. On average, each user is associated with 2.5 sensitive categories. Finally, we compared the number of these categories across our two user groups with different privacy ratios. We found users with high privacy ratios to be associated with a larger number of private and sensitive categories, while those with a low privacy ratio are linked with fewer sensitive categories. This difference is statistically significant (T-test P-value < 0.005) and is consistent with our findings presented in Chapter 4.

Bibliography

[1] Aylin Caliskan Islam, Jonathan Walsh, and Rachel Greenstadt. Privacy detective: Detecting private information and collective privacy behavior in a large social network. In *Proceedings of the Workshop on Privacy in the Electronic Society*, pages 35–46, 2014.

[2] Ponnurangam Kumaraguru and Lorrie Faith Cranor. Privacy indexes: A survey of westin's studies. 2005.

Curriculum Vitae

Name:	Taraneh Khazaei
Post-Secondary Education	Amirkabir University of Technology, BSc Tehran, Iran, 2006 - 2010
	Memorial University, MSc St. John's, NL, 2010 - 2012
	University of Western Ontario, PhD London, ON, 2012 - 2016
Honours and Awards:	Fellow of the School of Graduate Studies 2011 - 2012
	Association of the Advancement on Artificial Intelligence Travel Award July 2014 and May 2016
	Ontario Graduate Scholarship 2014 - 2015 and 2015 - 2016
	Mitacs Accelerate Fellowship 2015 - 2016
Work Experience:	Teaching Assistant, Memorial and Western University 2010 - 2014
	Application Developer, SHARCNET London, ON, 2014 - 2015
	Data Science Intern, InfoTrellis Inc. Toronto, ON, 2015 - 2016

Publications:

- T. Khazaei L. Xiao, R. Mercer, and A. Khan, Homophily and Privacy: Addressing Privacy Dichotomy in Twitter, *ACM Transactions on Social Computing* (to be submitted).
- T. Khazaei L. Xiao, and R. Mercer, How to Change Beliefs? Determinants of Persuasion in Online Comments, *ACM Transactions on Social Computing* (submitted).
- T. Khazaei, L. Xiao, R. Mercer, and A. Khan. Privacy Preference Inference via Collaborative Filtering, *In Proceedings of the AAAI Conference on Web and Social Media*, pp. 611-615, 2016.
- T. Khazaei, L. Xiao, R. Mercer, and A. Khan. Privacy Behaviour and Profile Configuration in Twitter, *In Proceedings of the Conference on World Wide Web - WWW Companion Volume*, pp. 575-581, 2016.
- T. Khazaei, L. Xiao, R. Mercer, and A. Khan. Detecting Privacy Preferences from Online Social Footprints: A Literature Review, *In Proceedings of the iConference*, 2016.
- T. Khazaei and O. Hoeber, Supporting Academic Search Tasks through Citation Visualization and Exploration, *International Journal on Digital Libraries (in press)*, 2016 .
- T. Khazaei, L. Xiao, R. Mercer, and A. Khan, Social Computing and Intelligence: Exploring Opportunities for the Public and the Enterprise, *IBM CASCON (Emerging Technologies Track)*, 2015.
- T. Khazaei, L. Xiao, and R. Mercer, Identification and Disambiguation of Lexical Cues of Rhetorical Relations across Different Text Genres, *In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) - Workshops*, pp. 54-63, 2015.
- O. Hoeber and T. Khazaei, Evaluating Citation Visualization and Exploration Methods for Supporting Academic Search Tasks, *Online Information Review*, 39(2): 229-254, 2015.
- T. Khazaei and L. Xiao, Computational Analysis of Collective Intelligence - Towards Automatic Detection of Rationales in Online Deliberations, *In Proceedings of the Collective Intelligence Conference*, 2015.
- T. Khazaei and L. Xiao, Corpus-based Analysis of Rhetorical Relations: A Study of Lexical Cues, *In Proceedings of the IEEE Semantic Computing Conference*, pp. 417-423, 2015.
- T. Khazaei and L. Xiao, Computational Analysis of Collective Intelligence in Conversational Text, *In Proceedings of the Hawaii International Conference on System Sciences (HICSS)*, pp. 1596-1605, 2015.
- T. Khazaei, Modeling Argumentation and Explanation in the Social Web, *In Proceedings of the AAAI Conference on Artificial Intelligence - Doctoral Consortium*, pp. 3057-3059, 2014.
- T. Khazaei and L. Xiao. Collective Intelligence in Massive Online Dialogues, *In Proceed-*

ings of the Collective Intelligence Conference, 2014.
- L. Xiao and T. Khazaei. ProjectTales: Reusing Change Decisions and Rationales in Project Management, *In Proceedings of the iConference*, pp. 895-900, 2014.
- T. Khazaei and O. Hoeber, Metadata Visualization of Scholarly Search Results: Supporting Exploration and Discovery, *In Proceedings of the Conference on Knowledge Management and Knowledge Technologies*, pp. 1-8, 2012.

Talks and Posters:

- ICWSM Conference, Cologne, Germany, 2016.
- WWW Conference [MSM'6], Montreal, QC, Canada, 2016.
- iConference, Philadelphia, PA, United States, 2016.
- IBM CASCON, Toronto, ON, Canada, 2015.
- EMNLP Conference [LSDSem'5], Lisbon, Portugal, 2015.
- Social Media and Society Conference, Toronto, ON, Canada, 2015.
- IEEE Semantic Computing Conference, Anaheim, CA, United States, 2015.
- CANARIE - Research Software Developer's Workshop, Toronto, ON, Canada, 2014.
- AAAI Conference, Quebec City, QC, Canada, 2014.
- Collective Intelligence Conference, Boston, MA, United States, 2014.