Influence diffusion detection between linked bloggers

Tan, Luke Kien Weng

2014

Tan, L. K. W. (2014). Influence diffusion detection between linked bloggers. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/62264>

<https://doi.org/10.32657/10356/62264>

Nanyang Technological University

Downloaded on 26 Apr 2023 18:48:23 SGT

**NANYANG
TECHNOLOGICAL
UNIVERSITY**

**INFLUENCE DIFFUSION DETECTION
BETWEEN LINKED BLOGGERS**

TAN KIEN WENG LUKE

**WEE KIM WEE SCHOOL OF COMMUNICATION AND
INFORMATION**

2014

INFLUENCE DIFFUSION DETECTION BETWEEN LINKED BLOGGERS

TAN KIEN WENG LUKE

Wee Kim Wee School of Communication and Information

A thesis submitted to the Nanyang Technological University
in partial fulfilment of the requirement for the degree of
Doctor of Philosophy

2014

ACKNOWLEDGEMENTS

During the course of my study, I was blessed with people who have made this journey possible. I would like to express my deepest gratitude to my supervisors, Dr. Na Jin Cheon and Dr. Theng Yin Leng for their supervision, guidance, and generosity in sharing their invaluable knowledge. Their support given and willingness to accommodate my tight work schedule allowed me to do this program on a part-time basis. My deep gratitude to Dr. Na for his valuable advice and guidance given throughout my graduate studies. His dedication towards mentoring and willingness to share his knowledge and experiences has steered me to improve the quality of my study. I am grateful to Dr. Theng for her advice on envisioning this dissertation with a broader perspective and her support in the research studies.

I would like to thank Dr. Chang Kuiyu for accepting me into his research program which allowed the expansion of my research studies. To Dr. Christopher Khoo, the words of encouragement and advice during our meetings are deeply appreciated. I am thankful to the GSO staff for their administrative support.

A personal note of gratitude to Mr. Kang, who was instrumental in getting me started on this PhD program and has always encouraged and supported my study. I am grateful to my management for their continuous support and understanding during the whole duration of this study.

Without the love, support, and sacrifices of my parents, wife Angela, and children Josiah and Jenell, this challenging endeavor would not have been possible.

Above all, I thank God for His abundant grace and manifold blessings along this learning journey.

TABLE OF CONTENTS

TABLE OF CONTENTS	ii
List of Figures.....	v
List of Tables	vii
ABSTRACT.....	ix
CHAPTER 1	1
INTRODUCTION	1
1.1 Background and Motivation.....	1
1.2 Research Gaps and Objectives	4
1.3 Research Methodology Overview	8
1.4 Importance of Research	14
1.5 Outline of Thesis	16
CHAPTER 2	19
LITERATURE REVIEW	19
2.1 Introduction	19
2.2 Influence Definitions.....	19
2.3 Influence Detection in the Blogosphere.....	23
2.4 Sentiment Analysis	27
2.5 Influence Diffusion Detection Models.....	51
2.6 Summary	61
CHAPTER 3	64
ANALYSIS OF BLOG FEATURES FOR INFLUENCE DETECTION	64
3.1 Introduction	64
3.2 Blog Features Analysis Model	65
3.3 Dataset Used	71
3.4 Experimental Results	74
3.5 Evaluation of Product Review Domain Blogs	88
3.6 Discussion	99
3.7 Summary	102
CHAPTER 4	105
SENTIMENT ANALYSIS OF BLOG POSTINGS	105
4.1 Introduction	105
4.2 Background	106
4.3 Heuristic Method.....	109
4.4 Class Sequential Rules Method.....	111

4.5	Typed Dependencies Rules Evaluation.....	114
4.6	Handling Complex Relationships	128
4.7	Sentiment Analysis Model Evaluation.....	135
4.8	Summary	144
CHAPTER 5	.....	146
	BLOGGERS' INFLUENCE STYLE ANALYSIS.....	146
5.1	Introduction	146
5.2	Background	147
5.3	Influence Style Model	149
5.4	Influence Style Analysis Framework.....	154
5.5	Dataset Used	163
5.6	Evaluation of Influence Style Model	165
5.7	Discussion	177
5.8	Summary	179
CHAPTER 6	.....	181
	INFLUENCE DIFFUSION DETECTION BETWEEN LINKED BLOGGERS ...	181
6.1	Introduction	181
6.2	Background	182
6.3	Influence Diffusion Detection Model	184
6.4	Dataset Used	198
6.5	Model Evaluation and Results	198
6.6	Discussion	207
6.7	Summary	210
CHAPTER 7	.....	212
	CONCLUSION.....	212
7.1	Introduction	212
7.2	Review of Research Objectives	212
7.3	Comparison with Related Work.....	221
7.4	Contributions.....	223
7.5	Limitations of Study.....	225
7.6	Future Work	228
7.7	Concluding Remarks.....	230
REFERENCES	.....	232
APPENDIX A: Additional samples of terms found in lexicon	.....	244
APPENDIX B: Software Specifications	.....	247
APPENDIX C: Sample coding for typed dependencies rules.	.....	248
APPENDIX D: Bloggers Influence Network	.....	249

APPENDIX E: Samples of coded data for influence diffusion detection.	250
APPENDIX F: Instructions given to coders for inter-coder reliability testing...	252
APPENDIX G: Samples of Clause-Level Typed Dependencies	255
APPENDIX H: Experiments to determine the IDD models' threshold values. ..	258
APPENDIX I: List of Publications.....	259

List of Figures

Figure 1.1 Research methodology overview.	10
Figure 1.2. Overview of phase one.	11
Figure 1.3 Overview of phase two.	12
Figure 1.4 Overview of phase three.	14
Figure 1.5 Structure of thesis.	17
Figure 2.1 Output of typed dependency parsing.	31
Figure 2.2 Dependency representation of words.	31
Figure 2.3 Top 10 most frequent cascades discovered.	52
Figure 2.4 Typical Bass curve.	53
Figure 2.5 Diffusion process in a Linear Threshold model.	55
Figure 2.6 Diffusion process in an Independent Cascade model.	56
Figure 3.1 Blogs network of the collected data.	72
Figure 3.2 Time-difference-value versus frequency of influence.	82
Figure 3.3 Average time difference versus influence frequency occurrence.	86
Figure 3.4 Time-difference-value versus frequency of influence (product review domain)	93
Figure 3.5 Average time difference versus influence frequency occurrence.	98
Figure 4.1 Typed dependency bigrams.	106
Figure 4.2 Typed dependency tree.	107
Figure 4.3 Generated typed dependencies.	108
Figure 4.4 Overview of the CSR and HR approach.	109
Figure 4.5 Sample generated output of coded data.	115
Figure 4.6 Sample generated output for POS-polarity evaluation.	116
Figure 4.7 Sample generated output for polarity evaluation.	116
Figure 4.8 Subjective phrases across typed dependencies.	130
Figure 4.9 Generated typed dependencies.	131
Figure 4.10 Generated phrase structure tree.	132
Figure 4.11 Phrase entities with the corresponding typed dependencies.	133
Figure 4.12 Generated phrase structure tree.	134
Figure 4.13 Typed dependency tree.	135
Figure 4.14 Overview of the proposed Sentiment Analysis model.	135
Figure 5.1 Bloggers of multi-author blog site MacRumors.	147
Figure 5.2 Description of blog site MacRumors.	148
Figure 5.3 Klout's Influence Matrix.	149
Figure 5.4 Influence style analysis framework.	155
Figure 5.5 Example of subjective phrases.	157
Figure 5.6 Example of objective phrases.	157
Figure 5.7 Phrase structure tree of clause.	158
Figure 5.8 Typed dependencies of clause.	159
Figure 5.9 Example of compliance between linked blog posts.	161
Figure 5.10 Example of non-compliance between linked blog posts.	161
Figure 5.11 Near-term linked posts count of top 3 MacRumors bloggers.	169
Figure 5.12 Mid-term linked posts count of top 3 MacRumors bloggers.	169
Figure 5.13 Near-term linked posts count of 9to5Mac bloggers.	175
Figure 6.1 Link frequency approach.	183
Figure 6.2 Sentiment links approach.	183
Figure 6.3 Sentiment links and influence style approach.	183

Figure 6.4 Influence style analysis framework.	185
Figure 6.5 Independent Cascade Scoring for linked bloggers.	194
Figure 6.6 Linear Threshold Scoring for linked bloggers.....	194
Figure 6.7 Example of compliance between linked blog posts	200
Figure 6.8 Example of non-compliance between linked blog posts.	201
Figure 6.9 Graph of Influence Style Scores.....	201
Figure 6.10 Influence diffusion path.....	208
Figure 7.1 Overview of the research's first phase approach.....	213
Figure 7.2 Overview of the research's second phase approach.	215
Figure 7.3 Overview of the research's third phase approach.....	217
Figure 7.4 Overview of the IDD model.	219
Figure D1. Bloggers Influence Network.....	249
Figure H1. Scores for the IDD-IS-Sent (IC) and IDD-IS-Sent (LT) model	258

List of Tables

Table 2.1	Contextual sentiment for subject or object.	34
Table 2.2	Derivation of sentiment scores.	46
Table 2.3	Examples of syntactic rules for finding targets of opinions.	49
Table 3.1	Blog post level and blog level features analysed.	66
Table 3.2	Inter-coder agreement between two coders.	73
Table 3.3	Sentiment value versus influence cross-tabulation.	75
Table 3.4	Agreement value versus influence cross-tabulation.	77
Table 3.5	Descriptive statistics for total sentences count	81
Table 3.6	Similar sentiment ratio versus influence cross-tabulation	83
Table 3.7	Agreement ratio versus influence cross-tabulation.	84
Table 3.8	Agreement value versus sentiment value.	85
Table 3.9	Auth(linked blog) > Auth(linking blog) versus influence cross-tabulation.	87
Table 3.10	Community value versus influence.	87
Table 3.11	Sentiment value versus influence (product review domain).	89
Table 3.12	Agreement value versus influence (product review domain).	90
Table 3.13	Agreement value versus sentiment value (product review domain).	91
Table 3.14	Descriptive statistics for total sentences count (product review domain)	92
Table 3.15	Similar sentiment ratio versus influence (product review domain)	94
Table 3.16	Agreement ratio versus influence	96
Table 4.1	Heuristic pattern rules (HR).	110
Table 4.2	Results for respective confidence threshold values	112
Table 4.3	Results for respective support threshold values.	113
Table 4.4	Polarity distribution of the subjectivity lexicon terms.	114
Table 4.5	Polarity distribution of bigrams in coded dataset.	115
Table 4.6	CSR rules for AMOD typed dependency.	117
Table 4.7	Top 10 AMOD CSR rules.	118
Table 4.8	Results for AMOD POS-polarity and polarity rules.	119
Table 4.9	AMOD polarity rules.	119
Table 4.10	CSR generated pattern rules for ADVMOD typed dependency.	121
Table 4.11	Top 10 ADVMOD CSR rules.	122
Table 4.12	Results for ADVMOD POS-polarity and polarity rules	122
Table 4.13	ADVMOD polarity rules.	123
Table 4.14	CSR generated pattern rules DOBJ typed dependency	123
Table 4.15	Top 10 DOBJ CSR rules.	124
Table 4.16	Results for DOBJ POS-polarity and polarity rules.	124
Table 4.17	DOBJ polarity rules.	125
Table 4.18	List of sample subjective phrases.	133
Table 4.19	AMOD typed dependency rules	136
Table 4.20	ADVMOD typed dependency rules.	136
Table 4.21	DOBJ typed dependency rules.	137
Table 4.22	NSUBJ typed dependency rules	137
Table 4.23	List of prior sentiment scores assigned to word terms.	138
Table 4.24	Lexicon based rules.	138
Table 4.25	Baseline test results.	141
Table 4.26	Typed dependency (TD)-only test results.	142
Table 4.27	Proposed Sentiment Analysis model test results.	142
Table 5.1	Influence style model.	150

Table 5.2 Syntactic rules for finding targets of opinions.....	160
Table 5.3 Examples of target topics and their overall aspects and features.	160
Table 5.4 Threshold values of corresponding influence style scores.	162
Table 5.5 MacRumors bloggers and their respective blog post count.	163
Table 5.6 9to5Mac bloggers and their respective blog post count.	164
Table 5.7 Participating-Listening values.	166
Table 5.8 Engagement style (Jaccard coefficient) scores.	168
Table 5.9 Subjectivity values	170
Table 5.10 Similar sentiment and dissimilar sentiment values.....	171
Table 5.11 Profiles of MacRumors bloggers influence.	172
Table 5.12 Participating-Listening values.	173
Table 5.13 Creating-Sharing (Jaccard coefficient) scores.	174
Table 5.14 Profiles of 9to5Mac bloggers influence.....	176
Table 6.1 Influence style analysis methods.	185
Table 6.2 Examples of target topics, aspects and features.....	189
Table 6.3 Summary of influence styles score measurement methods.	189
Table 6.4 Influence style scores.....	199
Table 6.5 Influence Style Profiles.....	202
Table 6.6 Influence diffusion detection evaluation test results.....	204
Table 6.7 T-test results.....	206
Table G1. Categorization of Clause-Level Typed Dependencies.....	255

ABSTRACT

The easy access and availability of blogs have encouraged web-users to change from consumers to providers of information. Providers of such content exert a certain level of influence on the receivers, giving rise to the interest in influence detection within the blogosphere. Previous works have focused on simple blog features to detect blogosphere influence which may not yield accurate results as influence itself is a complex concept that requires more in-depth analysis, where other intrinsic blog features, such as sentiments and agreement expressed in the blog posts or bloggers' influence styles, that provide further knowledge on the blogs' influence could be used to improve influence diffusion detection performance. This study aims to develop a model that uses sentiments on common topics, agreement expressions, and influence styles as possible features to detect influence diffusion between linked bloggers. Influence is defined and limited to the scope in the capacity of the linked blogger to exert on the linking blogger to agree with or have similar sentiments on the discussed topics.

The objectives addressed in this dissertation study are: 1) Identify the relevant blog features that are useful in detecting influence diffusion; 2) Establish an automatic sentiment analysis approach that takes into consideration the complex relationships between words in the blog posts for influence diffusion detection; and 3) Develop an influence diffusion detection model using similar sentiments between the linking and linked bloggers, agreement towards the linked bloggers, and bloggers' influence styles as features to improve performance. The first phase of the study involved statistical analysis on the blog dataset to identify the various blog features that could indicate influence. Results from the initial study show that similar sentiments on common

concepts between linked blog posts give the clearest indication of influence as compared with other blog features. Recent sentiment analysis studies focused on the functional relationships between words using typed dependency parsing, which provides a refined analysis on the grammar and semantics of textual data. Consequently, the second phase of the study evaluated and established an automatic sentiment analysis process based on a linguistic and semantic analysis approach, which further considered the complex relationships between words in the context of the target terms. Context in this case refers to the neighboring terms of a current term within the blog post. In addition, influence as a concept could be correlated with other factors such as the different influence styles. Bloggers are not restricted to a monolithic description of influence based on link existence, and may differ in the manner in which they exert influence. Knowing the bloggers' influence styles can better describe how influence is propagated in the network. Based on the findings in phase one and two, the third phase analyzed bloggers' influence styles and developed the *Influence Diffusion Detection* (IDD) model to automatically detect influence diffusion between linked bloggers using the sentiments on common topics between the linking and linked bloggers, agreement towards the linked bloggers, and bloggers' influence styles as features. Influence style refers to the manner in which a blogger exerts influence through the blog postings, and can be described through the engagement, persuasion, and persona styles of the blogger. The results show that The IDD model performed well (average 76% F-score) compared to the in-degree and sentiment values baseline approaches as the model closely maps the influence characteristics of the bloggers to the links between them through providing a clearer description of the manner in which influence was exerted. Though this dissertation study focused on the analysis of influence diffusion between bloggers and their immediate neighbors, the influence

diffusion across a blogosphere network could be inferred through analyzing the influence diffusion through subsequent linked bloggers.

The main contributions of this study are the use of sentiment analysis to extract the opinions of the linking and linked blog posts on the topics discussed to detect influence diffusion, and the introduction of influence styles as features in an influence diffusion detection model approach. As a practical implication, the IDD model could automatically detect influence diffusion between linked bloggers with higher accuracy and further describe the detected influence in relation to the influence styles of the linked bloggers. In addition, the influence style of each blogger along the influence path could be plotted to visualize the chain of influence styles pattern in the bloggers' network.

CHAPTER 1

INTRODUCTION

1.1 Background and Motivation

The growth and popularity of opinion-rich resources such as blogs and online-reviews have presented new opportunities and challenges for researchers to actively use technologies to extract opinionated information and measure the effects of influence propagation from these resources. The aim of this study is to develop an influence diffusion detection model to automatically detect the influence flow between linked bloggers. In this study, influence diffusion detection is focused at the blogger level through aggregating the behaviors of the bloggers at blog posts level, where the scope of analysis is confined to the diffusion of influence between a blogger and his/her immediate neighboring linked bloggers.

A Weblog or Blog is a specialized web site that allows an individual or group of individuals to express their thoughts, voice their opinions, and share their experiences and ideas. Blogs have certain characteristics. The entries are called blog posts, while the individuals who authored the blog posts are referred to as bloggers. The websites that publish these blog posts are termed blog sites or blogs. Blogs are regularly updated web pages where dated articles (i.e. blog posts) are listed in reverse chronological order. Bloggers post individual postings whereby readers could read and possibly link via their own postings (Herring et al., 2005b; Herring, 2009). The easy access and availability of blog sites (e.g., www.blogspot.com) have encouraged web-users to change from consumers to providers of information. Providers of such content

exert a certain level of influence on the receivers and this is evident from blog sites having an effect on their readers' purchase decisions (e.g., www.engadget.com), attitudes and approaches to life (e.g., www.lifehack.org), financial decisions (e.g., www.goodfinancialcents.com), and others.

Merriam-Webster dictionary (www.merriam-webster.com) defines influence as “*the power or capacity of causing an effect in indirect or intangible ways*”. Influence is a characteristic of an individual that defines the capacity of exerting some effect on other individuals (Agarwal and Liu, 2009). This study confines the scope of influence to the capacity of the linked blogger to exert on the linking blogger to agree with and have similar sentiment polarity on the topics discussed in the linked blog post content based on the concept of conformity and influence observed in Kelman (1958), Cialdini (2001), and Cialdini and Goldstein (2004). Influence is often related to conformity, which is the act of matching attitudes, beliefs, and behaviors to group norms (Hogg and Vaughan, 2005). Kelman (1958) identified conformity as a form of compliance, describing it as the expression of agreement towards the people or groups who are influential to the individual, while possibly keeping one's own original beliefs. Several human characteristics: authority, reciprocity, commitment and consistency, scarcity, and liking were defined in Cialdini (2001) with respect to influence. *Authority* refers to the expert bloggers who hold a certain level of influence in the subject matter. *Reciprocity* is behavior whereby people tend to return a favor often given from an influential blogger. People are usually *committed* and *consistent*, and do not like to be self-contradictory. Once they commit to an idea or behavior, they are averse to changing their minds without a good reason. This means influential bloggers tend to maintain their status unless something adverse happens. In addition, *scarcity* of resources will generate demand, increasing the influence of bloggers with valued

resources. Cialdini (2001) further postulated that *liking* or a high level of agreement received from others is an indication of influence, inferring that a high degree of agreement towards the linked blogger expressed in the blog content would indicate influence. Cialdini and Goldstein (2004) observed that how people react to beliefs held by others is often contingent on their perceptions of the level of consensus for those beliefs. The objective consensus approach (Mackie, 1987) states that individuals are more likely to systematically process a majority-endorsed message because people assume that the majority view reflects reality and people believe they share similar attitudes to the members of the majority. This means that bloggers would tend to share blog post content of those who had influenced them and whom they agree with. Brehm (1966) observed that emotion and disposition may affect likelihood of conformity or anti-conformity. Persuasion is defined as a process of influence through appeals to reason or appeals to emotion. The effect of discrete emotions affects a target's cognition as well as the eventual outcome of the influence attempt (Cialdini and Goldstein, 2004). Therefore, influence may be determined by a blogger's persuasiveness.

Blogs are readily available sources whereby opinions and sentiments could be propagated to the readers. There could be a deeper relation whereby the authority of the information source determines the speed and intensity of the influence the opinion could have on the blogs readers. The ability to detect influence diffusion in the blogosphere could be used to identify influential blogger and the chain of information diffusion. For example, influential bloggers are often market-movers who can influence the buying decisions of fellow bloggers, and identifying the influence diffusion from them can help companies better understand the spread of new trends and concerns about the products, and predict possible potential customers through

analyzing the influence diffusion path. A further stimulus could be added to aid the flow of positive information, or pre-emptive and preventive actions taken to minimize any negative impact.

1.2 Research Gaps and Objectives

Previous studies linked information and influence propagation to blog features which are mainly graph-based, such as the number of in and out-links (Adar and Adamic, 2005; Agarwal and Liu, 2008). However, the use of blog features alone to detect influence diffusion would not give highly accurate results as influence is a subjective concept and often depends on the context, that is, the circumstances surrounding the posting (Herring et al., 2005b; Cai et al., 2011; Li, Bhowmick, and Sun, 2011). Content analysis is also used to detect influence within the blogosphere (Gruhl et al., 2004; Kempe et al., 2003; Matsumura et al., 2008; Song et al., 2007). Most of the content analysis approaches measured document similarity between linked blog posts to detect influence. These approaches mainly checked whether the linked blog posts are discussing the same topic and did not consider the circumstances surrounding the blog post content for influence detection. Agarwal (2008), Adar and Adamic (2005), and Ghosh and Lerman (2008) studied community identity in relation to influence. The assumption is that blogs within a community would tend to influence each other as close affiliates would. However, community identity alone would not be able to accurately detect influence, or detect all cases of influence. This is because being in the same community may not necessary mean that bloggers will influence each other, and influence could also exist in cases where bloggers are from different communities. The first research gap identified is (RG1: *The use of simple blog features*

to detect influence diffusion may not yield highly accurate results as influence itself is a complex concept that requires more in-depth analysis, where other intrinsic blog features that provide further details on the influence diffusion process could be used to improve influence diffusion detection performance). This leads to the question on *what are the relevant blog features that could be used to detect influence diffusion between linked blog posts (RQ1).*

An exploratory study to analyze the possible useful blog features is described in Chapter 3. The study identified the sentiment expressed on common topics between the linked blog posts as a useful blog feature for detecting influence diffusion between blog posts. Recent sentiment analysis studies focused on the functional relations of words using typed dependency parsing, which provides a refined analysis on the grammar and semantics of textual data. Heuristics are typically used (Shaikh et al., 2008; Thet et al., 2010) to determine typed dependency polarity patterns, which may not comprehensively identify all possible rules. This leads to the second research gap (RG2: *The existing heuristic approach to sentiment analysis is based on available knowledge base where rules generation is not automatic, and may not provide a complete set of rules*). It was observed in Agarwal et al. (2008) that acquiring the ability to detect sentiments in the content could identify the chain of influence flow. For example, the negative sentiment expressed by the phrase “extensive damage” in the sentence “The oil spill caused extensive damage to marine and wildlife habitat.” would cause readers to have a negative opinion of the oil spill. Wiebe (1994) defined sentiment analysis as a process to identify opinions, emotions, and evaluations expressed in natural language. Sentiment analysis was studied by Turney (2002) and Riloff and Wiebe (2003) to detect sentiments at both document and sentence level. However, Pang and Lee (2004) highlighted that sentiment often could be expressed in

a subtle manner, making it difficult to identify by a sentence or document's terms considered in isolation. Wilson et al. (2009) and Polanyi and Zaenen (2006) further observed that there exist more complex relationships between words where deeper analysis is required to decipher the expressed sentiments. Hence, there is a need to explore solutions for a sentiment analysis method that can take into account the complex relationships between words in the context of the target terms, where context refers to the neighboring terms of a target term. The second research question posed in this study is: *How can the sentiment analysis process be enhanced to better detect the sentiments expressed within blog posts through considering the complex relationships between words (RQ2)?*

While more recent studies also considered the influence between blog posts by using the sentiments between linked blog posts (Leskovec et al, 2010; Cai et al., 2011; Li et al., 2011), these studies focused on a single notion that influence exists between blogs with hypertext links, and ignored the details in the other aspects of influence. Influence as a concept could be correlated with other factors such as the different influence styles, and could not be described using a single detached feature. Bloggers are not restricted to a monolithic description of influence and may differ in influence style. Other studies (Guadagno et al., 2008; Yarkoni, 2010) studied bloggers' personalities with regards to their propensity to blog, but did not relate the personalities of bloggers to influence. Bloggers may differ in their way or manner of exerting influence. For example, bloggers could be active participants or just passive sharers, or may express their express ideas in a rational or subjective manner, or may be received positively or negatively by the readers. The third research gap identified is (RG3: *Influence diffusion could not be described using a single detached feature such as the existence of links or similar sentiment polarity*), where the influence styles of

bloggers, for example engagement style, persuasion style, and persona, could be further used as features in the proposed influence style model to improve influence diffusion detection performance.

The aim of this study is to develop an influence diffusion detection model to automatically detect the influence flow between linked bloggers. A review of the literature shows previous studies focused on a single notion that influence exists between linked bloggers (Adar and Adamic, 2005; Agarwal and Liu, 2008) or bloggers with similar sentiment polarity (Leskovec et al, 2010; Cai et al., 2011; Li et al., 2011), but ignored the other aspects of influence. Influence as a concept could be correlated with factors such as the different influence styles, where knowing the bloggers' influence styles can better describe how influence is propagated in the network. The third research question in this study is: *Will automatic influence style analysis help detect influence diffusion between linked bloggers (RQ3)?*

To answer the research questions, the following research objectives will be carried out:

- *Identify the relevant blog features that could be used to detect influence diffusion between linked blog posts (RO1).*
- *Establish a sentiment analysis approach that takes into consideration the complex relationships between words in blog posts to automatically detect influence diffusion between linking and linked blog posts (RO2).*
- *Develop an influence diffusion detection model using sentiment values between linking and linked bloggers, agreement towards the linked bloggers, and bloggers' influence styles as possible features to improve influence diffusion detection between linked bloggers (RO3).*

The research explores areas in the identification of blog features that could reveal influence, including the influence styles of bloggers. The research uses

sentiment analysis and influence style features in the proposed model to detect influence between linked bloggers. Unlike previous studies which focused only on influence detection, but did not describe the manner in which influence is exerted, the use of sentiment values, agreement expressions, and influence style features in this research could improve performance and provide a clearer explanation for influence diffusion between the linked bloggers.

1.3 Research Methodology Overview

In this study, influence is defined as the capacity of the linked blogger to have an effect on the linking blogger to agree with or have similar sentiment polarity on the topics discussed in the linked blog post content. Influence diffusion from the linked blogger to the linking blogger is deemed to have occurred when the linking blogger expressed agreement with the linked blogger or expressed similar sentiments on common topics found in both linked blog posts as a result of the linking blogger having read the post of the linked blogger and creating a link to it.

A blogosphere could have numerous posts resulting in a complex network of influence between the bloggers. In this study, influence diffusion detection is focused at the blogger level through aggregating the behaviors of the bloggers at blog posts level, where the scope of analysis is confined to the diffusion of influence between direct-linked bloggers, as compared to previous studies (Gruhl et al., 2004; Lim et al., 2011) which focused on influence initiated from selected nodes and propagated throughout the network. Nevertheless, the results from the analysis in this study between direct-linked bloggers could be applied iteratively across subsequent linked bloggers to determine the influence diffusion across a blog network.

Blogs can exist in different domains, ranging from product reviews, sports, health care, finance, automobile, etc. Blogs from different domains could contain domain specific jargon (e.g., "google" could refer to the action of searching for a topic online) and terms that have different meanings in different domains (e.g., "stocks" refers to equities in a finance blog, while it could refer to the physical product (e.g., "available stock" in a product review blog site). The focus of this study is on product review blog sites that contain blog posts discussing the review of products. To understand the influence propagation characteristics of bloggers, archives from product review blog posts of bloggers are used as the main dataset in the study.

The overview of the research methodology is shown in Figure 1.1. To meet the research goal of developing an influence detection model for automatic detection of influence diffusion between linked bloggers, the research study is done in three phases. Phase one of the research used statistical methods to determine blog features that are suitable for influence detection between linked blogs. The results showed that similar sentiment polarity expressed on common topics could significantly detect the diffusion of influence between linked blog posts. Based on this finding, the CSR algorithm was applied in phase two of the research to improve and automate the typed dependencies rules generation. Coupled with a semantic analysis approach that considers the complex relationships between words, the sentiments expressed in the blog posts were detected. While previous studies ((Leskovec et al, 2010; Cai et al., 2011; Li et al., 2011)) analyzed the relation between sentiments and influence in the blogosphere, this study introduced the novel concept of using bloggers' influence styles to detect influence diffusion. Influence styles were defined through the engagement level, persuasiveness, and persona of the bloggers. Methods to detect sentiments between bloggers and agreement towards the linked blogger studied in the previous phases were used to

analyze the persona influence style of the bloggers. The Independent Cascade and Linear Threshold approaches were applied in phase three to model the Influence Diffusion Detection Model using the sentiment and agreement values between the linked bloggers and their influence styles as features.

Figure 1.1 Research methodology overview.

An overview of phase one is shown in Figure 1.2. An exploratory study is done to determine relevant blog features that could determine influence diffusion in response to the research gap (RG1), where blog features used in previous studies might

not detect influence accurately as influence itself is a complex concept that requires more in-depth analysis, where other intrinsic blog features, such as sentiments and agreement values or the manner in which influence is exerted by the bloggers, that provide further details on the blogs' influence could be used to improve influence diffusion detection performance. Statistical analysis will be performed on the blog dataset to determine blog features that are useful in detecting influence diffusion between the linked blogs as part of the first research method (RM1).

Figure 1.2. Overview of phase one.

Though the focus of this study is on the product review domain, the sample dataset for the blog features analysis study was collected from blog postings within the U.S. political domain blogosphere where it is common to find communities of bloggers with their respective political inclination which is part of the study's consideration for

community identification evaluation. Further testing was conducted on a product review domain dataset to validate the findings of the initial evaluation and the results from the product review dataset similarly showed that similar sentiments and agreement values between bloggers are clear indication of influence diffusion. The quantitative analysis of features found in the blogs together with manual sentiment analysis and community identification are performed on blogs postings and their content. Correlation analysis of the selected variables with influence is performed to determine the effectiveness of each variable.

Figure 1.3 Overview of phase two.

An overview of phase two of the research is shown in Figure 1.3. In phase two, sentiments expressed within the blog content are analyzed using an approach based on typed dependencies rules and semantic analysis on subjective phrases to address the

second research question (RQ2). Sentences with ten words or longer were extracted from the product review blog dataset available from Zafarani and Liu (2009) to handle the complex relationships between words cases in the context of the target terms. Context in this case refers to the neighboring terms of a current term. A linguistic method that leveraged typed dependency rules derived from Class Sequential Rules (CSR) (Liu, 2006) was used, and the complex phrase relationships found in a sentence are considered through analyzing the semantic relationships of words as part of the proposed research method (RM2) to establish an automatic sentiment analysis approach in consideration of the complex relationship between words in the blog posts. The CSR method was adopted as it is a machine learning approach that could automatically derive the typed dependencies rules based on the sequential ordering of the bigrams found in the word phrases.

An overview of phase three of the research is shown in Figure 1.4. For the third research objective (RO3) to address the research gap (RG3) and to answer the research question (RQ3), relevant features, for example degree and scope of topics discussed, content sharing ratio, and count of subjective words as well as the predicted sentiment polarity on topics were used to describe the influence style of the linked bloggers based on their engagement, persuasion, and persona style in the research method (RM3-i). The dataset consists of product review blog posts extracted through Technorati.com where the influence styles of bloggers were analyzed. Eventually, to meet the third research objective (RO3), the derived sentiment polarities between the linking and linked bloggers, agreement expressions towards the linked bloggers, and bloggers' influence styles were used as features in the Independent Cascade model (Goldenberg et al., 2001) and Linear Threshold model (Granovetter, 1978) as part of the research method (RM3-ii) to build the proposed Influence Diffusion Detection

(IDD) model. The software used in the analysis of data as well as for the development of the models is described in Appendix B.

Figure 1.4 Overview of phase three.

1.4 Importance of Research

The automatic detection of influence diffusion within the blogosphere has possible applications in enabling financial institutions to evaluate the market by measuring the impact of negative or positive influence flow for the listed companies, helping governmental institutions find the sphere of influence of terrorist organizations

using the blogosphere as a communication tool, or the identification of negative or positive influence flow for products where further marketing campaign could be employed to further spread positive influence or through measures taken to minimize or stop negative influence.

By performing quantitative analysis of features found in the blogs, as well as the sentiment and agreement analysis on the blog postings content, and community identification, the research study could determine the usefulness of each feature in detecting influence diffusion in the blogosphere. Previous studies (Adar and Adamic, 2005; Agarwal and Liu, 2008) focused on the use of blog features and did not fully consider the context, that is, the circumstances surrounding the blog post content in detecting blogosphere influence. Knowing the important blog features and the effect of content analysis in detecting influence among blog posts allows development of a more efficient influence diffusion model for determining the main chain of influence propagation in the blogosphere. Results from the initial study showed that sentiments on common topics between links are highly correlated with influence diffusion and could be used to detect influence propagation between linked blog posts.

Unlike most of the previous content analysis studies (Matsumura et al., 2008; Song et al., 2007) that measure document similarity (i.e. common terms) between linked blog posts, this research study explores the use of a linguistic and semantic analysis approach to take into consideration the sentiments expressed within the whole blog post content based on the sentiment analysis model. As compared to previous studies (Shaikh et al., 2008; Thet et al., 2010) which were based on a manual heuristic approach, this research study adopts a machine learning approach to automatically derive a comprehensive list of polarity pattern rules for detecting phrase polarity output. The study further introduces semantic rules into the sentiment analysis process

to handle the complex relationships between words in a phrase to improve performance.

This research study attempts to implement the novel approach of analyzing bloggers' influence styles and using the influence styles as features in the proposed Influence Diffusion Detection (IDD) model to improve the performance of influence diffusion detection performance between linked bloggers. Identification of bloggers' influence style could better describe the manner in which influence is propagated between linked bloggers. On the other hand, previous studies (Cai et al., 2011; Li et al., 2011) which assumed a monolithic definition of influence existence between linked blog posts would not be able to differentiate the varying influence styles of the individual bloggers. The proposed IDD model can be used to describe the influence styles of subsequent linked bloggers, thus providing a more comprehensive understanding of the bloggers' influence style pattern in the influence diffusion path. For example, influential bloggers are often market-movers who can influence the buying decisions of fellow bloggers, and identifying the influence diffusion from them can help companies better understand the spread of new trends and concerns about the products, and predict possible potential customers through analyzing the influence diffusion path.

1.5 Outline of Thesis

The thesis is organized into seven chapters as depicted in Figure 1.5. Chapter 1 presents the background and motivation of the research, the research objectives and overview of the research's methodology and contributions. Chapter 2 provides the literature review for the respective research areas, which encompasses previous studies

done on blogosphere influence detection, sentiment analysis, as well as influence styles profiling and diffusion models used to detect influence diffusion. The research methodology is described specifically in each respective chapter, with Chapter 3 providing details of the exploratory study on the analysis and identification of blog features for influence detection within the blogosphere. Chapter 4 focused on the study to analyze the sentiments polarity output between linked blog posts in consideration of the complex relationships found between words. Details of the bloggers' influence style analysis framework are provided in Chapter 5.

Figure 1.5 Structure of thesis

Chapter 6 describes the proposed Influence Diffusion Detection model that uses similar sentiment values between the linking and linked bloggers, agreement towards

the linked bloggers, and bloggers' influence styles as features to improve performance. The conclusion is given in Chapter 7 with a review of the research objectives and further discussion on the major findings, with a discussion on the contributions of the research, its limitations, and recommendations for future work.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

The literature review chapter covers previous studies done in the areas of influence definitions, blogosphere influence detection, sentiment analysis, influence style profiling, and influence diffusion detection models. The definition of influence was explored to provide a better understanding of the influence types in the blogosphere context in Section 2.2. The various blog features that were used in other studies to detect influence were analyzed in Section 2.3. This is followed by reviews of previous sentiment analysis works based on the linguistic and semantic approaches in Section 2.4. Section 2.5 explores the various diffusion models that could be applied for detecting influence diffusion, with the summary for the chapter given in Section 2.6.

2.2 Influence Definitions

Previous studies (Guadagno et al., 2008; Yarkoni, 2010) measured the personality of bloggers in relation to predicting blogging. Costa and McCrae (1992) evaluated personality based on clinical studies through five key traits: *neuroticism*, *extraversion*, *agreeableness*, *openness to experience*, and *conscientiousness*. Neuroticism represents an individual's tendency to experience distress. Extraversion is the dimension underlying a broad group of traits, including sociability, activity, and the tendency to experience positive emotions. Agreeableness is a dimension of

interpersonal behaviour, where high agreeableness individuals are trusting, sympathetic, and cooperative. Openness to Experience refers to traits that describe individuals as imaginative and sensitive to art and beauty and having a rich emotional life. Conscientiousness is a dimension that contrasts scrupulous, well-organized, and diligent people with lax and disorganized individuals. Guadagno et al. (2008) measured the personality of bloggers based on the five key traits used by Costa and McCrae (1992) to predict blogging. Their studies indicated that people who are high in openness to new experience and high in neuroticism are likely to be bloggers. The results indicated that personality factors impacted the likelihood of being a blogger and have implications for understanding who blogs. Similarly, Yarkoni (2010) analyzed blogger personality and word use in blog postings to study the association between personality and language of bloggers. However, these studies manually measured bloggers' characteristic and personality with regards to their propensity to blog, and did not relate the characteristic and personality of bloggers to their influence styles. Moreover, the five key personality traits used by Costa and McCrae (1992) are not directly applicable for measuring influence style. In this study, bloggers' traits in relation to their influence style, which could determine their propensity to influence bloggers, were studied.

Influence is often related to conformity, which is the act of matching attitudes, beliefs, and behaviors to group norms (Hogg and Vaughan, 2005). Norms are implicit rules shared by a group of individuals that guide their interactions with others and among society or social group. This tendency to conform occurs in small groups and society as a whole, and may result from subtle unconscious influences, or direct and overt social pressure. Kelman (1958) identified three major types of conformity. *Compliance*, a form of public conformity, is the expression of agreement towards the

people or groups who are influential to the individual, while possibly keeping his or her own original beliefs. Compliance is motivated by the need for approval and the fear of being rejected. *Identification* is the changing of attitudes or behaviors due to the effects of influence. *Internalization* is the process of both public and private acceptance of a set of norms established by the influential, and is the deepest level of influence. This study attempts to measure the compliance between blog posts through analyzing the sentiments on target topics in the textual content. The scope is limited to measuring only compliance in the absence of further contextual information to validate the other conformity types.

Cialdini (2001) identified certain characteristics that contribute to an individual's propensity to be influenced. People have the propensity to conform and agree to *authority*, which in this context refers to the subject experts. *Reciprocity* is the behavior whereby people tend to return a favor leading to mutual conformity. People are usually *committed* and *consistent*, and do not like to be self-contradictory. Once they commit to an idea or behavior, they are averse to changing their minds without good reason. From a social proof perspective, people will be more open to people who also agree with them, tend to obey authority figures, and be more easily swayed by people they like. Cialdini (2001) further proposed that a perceived limitation or *scarcity* of resources will generate demand. For example, bloggers deemed to be an authority in the subject matter are usually few in number and held in high esteem. Characteristics used in this study to evaluate bloggers' engagement style include authority, reciprocity, commitment and consistency, and scarcity. The behavior of sharing influential blog posts' content in an engagement style could be explained by Cialdini and Goldstein (2004)'s observation that how people react to beliefs held by others is often contingent on their perceptions of the level of consensus for those

beliefs. That is, bloggers would tend to share content of blog posts which they agree with. The objective consensus approach (Mackie, 1987) states that individuals are more likely to systematically process a majority-endorsed message because people assume that the majority view reflects reality and believe they share similar attitudes to the members of the majority. This means that bloggers would also tend to share blog post content of blogs with many in-links. This study considers the behavior of sharing blog posts' content as a *Creating- Sharing* type in the engagement style evaluation.

Brehm (1966) observed that emotion and disposition may affect likelihood of conformity or anti-conformity, while Cialdini and Goldstein (2004) concluded that the effects of discrete emotions could affect a target's cognition as well as the eventual outcome of the influence attempt. Persuasion is a process of influence through appeals to reason or appeals to emotion. Subjectivity, that is, opinionated or emotional phrases expressed in the blog content could provide clear evidence of the blogger's persuasion style. This study considers persuasion as a component of influence and evaluates it through analyzing the subjectivity expressed in the blog posts content.

Social influence could be defined through positive and negative social relationships based on the idea of persona in sociology studies where influence is the approvals gained from other reliable people (Cai et al., 2011). Granovetter (1973) and Krackhardt (1992) observed that the effects of the social influence by users of varying personae are different, where the three general types of persona identified were positive, negative, and controversial. This study defines and limits the scope of influence to the capacity of the linked blogger to exert on the linking blogger to agree with and have similar sentiment polarity on the topics discussed in the linked blog post content. Influence is further described through the engagement style, persuasion style, and persona of the bloggers based on the study in Chapter 5. The bloggers' influence

styles were then used as features to detect influence flow between bloggers through the Independent Cascade (IC) and the Linear Threshold (LT) diffusion models studied in Chapter 6.

2.3 Influence Detection in the Blogosphere

Social influence describes the phenomenon by which the behavior of an individual is directly or indirectly affected by the thoughts, feelings, and actions of others in a community (Cialdini, 2001). Similarly, social influence is found within the blogosphere social network community. Blogs play a vital role in spreading new ideas and information on the web. Typically, blogs' content represents the opinion of bloggers about topics ranging from technology and art to business marketing and public relations. The ability to detect influence in the blogosphere could be used effectively to identify the influential blogger and chain of information flow.

Previous studies that used blog features to detect influence were mainly graph-based. Herring et al. (2005b) investigated the extent to which blogs are interconnected using quantitative social network analysis, visualization of link patterns, qualitative analysis of references, and comments of the blogs. Their results showed that the blogosphere is partially interconnected and sporadically conversational. However, the interconnectivity of the blogosphere does not indicate the influence of the bloggers as there could be disagreement between linked blog posts. In the study by Agarwal and Liu (2008), an influential blogger was defined based on the number of in-links, comments for the posts, and the number of out-links. A high authority value based on a large number of in-links to the blog postings indicates high number of followers. Similarly, having many in-links may not necessary indicate influence as the blog posts

could contain opposing opinions. Hence, number of in-links alone could not accurately describe the influence exerted by a blogger. In addition, according to the study by Agarwal and Liu (2008), length of posts was equated to eloquence of the blogger where an eloquent blogger was deemed influential. Length of posts may be subjective in detecting influence as it is dependent on the writing style of the blogger. Instead the quality of the blog content would have provided clearer evidence of influence.

Analyzing the blog features alone may not be sufficient to yield highly accurate results in detecting influence within the blogosphere, and other studies have included content analysis in their approach. For instance, Adar and Adamic (2005) considered text similarity, in addition to the blog features such as the number of common blogs explicitly linked to by both blogs, the number of non-blog links shared by the two, and in-link and out-link count for the two blogs. Content analysis was used by Agarwal and Liu (2009), Matsumura et al. (2008), and Song et al. (2007) to detect influence within the blogosphere. Most of the content analysis approaches for influence detection measured document similarity between linked blog posts using information retrieval techniques, such as cosine similarity. These approaches mainly checked whether the linked blog posts were discussing the same topic for influence detection. An InfluenceRank algorithm was proposed by Song et al. (2007), which uses information novelty to rank blogs according to how influential and informative they are in the network. Using cosine similarity distance, information novelty was measured based on how dissimilar an entry blog was different from the entries it linked to. Agarwal and Liu (2009) analyzed the blog content and also used cosine similarity to find influential blogs that were representative and diverse. The Influence Diffusion Model (IDM) was applied in Matsumura et al. (2008) to identify the influencer and the influential terms in consumer package goods. IDM assumes if a term appears in all the

content of a chain of blog postings, the last blog post is influenced by the first blog post in the chain. However, the detection of influence through propagation of similar terms is subjected to variability where terms could propagate through different topics being discussed. Appearance of terms downstream could be interpreted as copying or referencing of quotations from the linked post, and may not always be interpreted as influence as the linking post may be quoting copied terms as reference while expressing opposite opinions on the subject. In the above studies, measuring text similarity between linked blog posts alone may not accurately measure influence in the postings as text similarity does not reveal the contextual information of the postings which affects whether influence is present or not.

Community identity could also correlate with influence within the blogosphere as shown in Adar and Adamic (2005), Agarwal (2008), and Ghosh and Lerman (2008). Ghosh and Lerman (2008) generalized the notion of network connectivity to be the number of paths, of any length, that exist between two nodes. They proposed that this network connectivity, which relates to influence in the blogosphere could be used to identify blogs within a community. Adar and Adamic (2005) used the number of common blogs explicitly linked to by both blogs to indicate the two blogs were in the same community. This community identity was then used as part of the technique to detect influence in the blogosphere. As employed in previous studies, community identification was used to detect influence in the blogosphere. This was usually based on the assumption that blogs within the same community would tend to influence each other as close affiliates would. However, community identity alone would not be able to accurately detect influence, or detect all cases of influence. This is because blogs outside of the community could also exert influence on blogs within the community and vice versa. Blogs of different communities could also share similar opinions on

certain topics, and blogs within a community may not always influence each other on all topics. For example, community detected through number of common links might not necessary imply influence as a high number of common links does not equate to influence. Thus, although community identity could be used to detect influence in the blogosphere, other measures have to be used in conjunction with it to produce more accurate and comprehensive results.

The above studies that used the feature-based, content similarity and community identity approaches utilized positive interactions (e.g., agreement and trust) between individuals, and ignored the negative relationships and conformity of people. Milgram (1963) observed that social influence is the strongest when the group perpetrating it is consistent and committed. Even a single instance of dissent can greatly weaken the strength of an influence. This means that the detection of both agreement (positive sentiments) and disagreement (negative sentiments) expressed in the content is equally important in influence analysis. In addition, content analysis could provide the blog's contextual information through the sentiments expressed on specific topics. Recent studies included consideration of both positive and negative links in the analysis of influence. Leskovec et al. (2010) adapted a framework of trust and distrust in an attempt to infer the attitude directed towards each blogger using the observed positive and negative relations. Tan, Na, and Theng (2011) observed that analyzing the positive, neutral, and negative sentiments of linked blog posts on common topics could improve influence propagation detection. Li et al. (2011) similarly considered the positive and negative edges of the nodes, and computed the influence index and conformity index in their attempt to detect influence in the social network. However, these studies focused on a single notion of the influence of

bloggers in detecting influence, that is, whether influence exists between linked blog posts, and did not study in details the influence styles of the blog sites or bloggers.

2.4 Sentiment Analysis

Sentiment analysis is a type of content analysis which aims to identify opinions, emotions, and evaluations expressed in natural language (Wiebe, 1994). With the recent proliferation of Web 2.0 sites and applications and the rapid growth of user generated content on the Internet, it has been studied by many researchers in recent years. Sentiment analysis usually involves detecting subjectivities in sentiment within the textual content. This research study aims to do fine-grained analysis at the phrase and sentence level to derive polarity pattern rules to predict with reliability the polarity of subjectivities expressed in the posts.

Wilson et al. (2009) and Polanyi and Zaenen (2006) observed that more complex relationships exist between words, such that the sentiment output of the words is affected through negation, shifting, intensification, mitigation, and domain related effect. This study further used a semantic analysis approach to resolve the complex word relations problem. Specific lexicons were created for the respective complex relations to identify the semantic words in the textual content. A common approach in sentiment analysis is the document or sentence level classification of sentiment polarity using sentiment or subjective words found in the text content. Polarity could be positive, negative, or at times neutral, and is often derived from a predefined sentiment word dictionary. Osman et al. (2009) proposed a document level opinion detection system which can find more relevant opinion documents (i.e. blog posts) using five system fusion methods, i.e. a voting method, an inverse rank method,

a linear-normalized score method, and two weighted methods. The resulting document list can then be used as a list for a search engine or as input into an opinion analysis system for further analysis. Hu and Liu (2004) mined and summarized customer reviews of electronic products, such as digital cameras, cellular phones, and mp3 players. They extracted the features or aspects (such as picture quality and screen size) of the product on which the customers have expressed their opinions, and predicted whether each opinion sentence is positive or negative. If positive or negative opinion words prevail, the opinion sentence is predicted as a positive or negative one. The approach used by Kim and Hovy (2006) identified an opinion bearing word, labeled semantic roles related to the word in the sentence, and then found a holder and a topic of the opinion word among labeled semantic roles. Zhang et al. (2009) studied sentiment analysis of Chinese content rather than English. They used a rule-based approach to determine a sentence's sentiment based on word dependency, and then predicted a document sentiment by aggregating the analysis results of individual sentences. Sentiment analysis is not without challenges. Pang and Lee (2008) pointed out that sentiment often could be expressed in a more subtle manner, making it difficult to identify by any sentence or document terms when considered in isolation. Sentiment and subjectivity are context-sensitive and, at a coarser granularity, domain or document genre dependent (Na et al., 2010). Further to that, the order in which different opinions are presented can result in a completely opposite overall sentiment polarity. These considerations have complicated sentiment analysis on the blog content, especially for initiatives to automate the process.

2.4.1 Linguistic Approach

Earlier studies in sentiment analysis at the sentence or phrase level used the notion that an opinion word associated with its aspect or features would appear in its vicinity (Hu and Liu, 2004; Kim and Hovy, 2004). However, opinionated text can be written in elaborate styles where sentences have nested clauses with the related opinion and subject words hidden in separate clauses. Methods using word distances (Hu and Liu, 2004) or part-of-speech patterns (Yi et al., 2003) will not be able to detect the relationship of the opinionated words as these methods worked only if both the opinion and aspect words appear within a certain window of one another. Moreover, the grammatical relationship between the words had largely been ignored.

More recent sentiment analysis studies used a linguistic approach which leveraged the semantic dependencies between words to predict sentiments. Wilson et al. (2006) observed that specific word patterns alongside its modifiers could determine the intensity of private states including sentiments. Modifiers are words that determine the grammatical relationships between the bigram words in a dependency representation. In their study, bigrams (termed as bilex) that appears in a specific pattern involving a word and just one of its modifiers were included as features in the evaluation. A common approach to derive the semantic and syntactic dependencies between words is through typed dependency parsing (Thet et al., 2010; Wilson et al., 2009; Hassan et al., 2010). Typed dependencies are otherwise known as grammatical relations.

2.4.2 Typed Dependency Level Analysis

The fundamental notion of dependency is based on the idea that the syntactic structure of a sentence consists of binary asymmetrical relations between the words of the sentence (Nivre, 2005). A natural language parser is a program that works out the grammatical structure of sentences, for example, the grouping of words as phrases, and the identification of words that are the subject or object of a verb. Probabilistic parsers use a knowledge base of language gained from manual parsing of sentences to predict the parsing of new sentences. An example is the Stanford Parser¹, which provides Stanford typed dependencies output as well as phrase structure trees. The parser implements a probabilistic context-free grammar (PCFG) method, that is, a context-free grammar in which each output is augmented with a probability. The probability of a parse derivation is the product of the probabilities of the productions used in that derivation. A simple 2-rule PCFG grammar is given below, where each rule is preceded by a probability that reflects the relative frequency with which it occurs.

Rule (1): 0.7 Verb_Phrase $\rightarrow$ Verb Noun_Phrase

Rule (2): 0.3 Verb_Phrase $\rightarrow$ Verb Noun_Phrase Noun_Phrase

For example, the number of noun phrases expected while deriving verb phrases is computed as $0.7 \times 1 + 0.3 \times 2 = 1.3$. The Stanford dependencies provide a representation of grammatical relations between words in a sentence. The dependencies are triplets consisting of: name of the relation, governor, and dependent. The standard dependencies and the graphical representation of the typed dependency parsing output for the sentence “*I like the iPhone as it has a very bright screen*” are given in Figure 2.1 and Figure 2.2.

---

¹ <http://nlp.stanford.edu/software/lex-parser.shtml>

Tagging

I/PRP like/VBP the/DT iPhone/NN as/IN it/PRP has/VBZ a/DT very/RB bright/JJ screen/NN

Parse

(ROOT

(S

(NP (PRP I)

(VP (VBP like)

(NP (DT the) (NN iPhone))

(SBAR (IN as)

(S

(NP (PRP it)

(VP (VBZ has)

(NP (DT a)

(ADJP (RB very) (JJ bright))

(NN screen))))))

(. .)))

Typed dependencies

nsubj(like-2, I-1)

det(iPhone-4, the-3)

dobj(like-2, iPhone-4)

mark(has-7, as-5)

nsubj(has-7, it-6)

advcl(like-2, has-7)

det(screen-11, a-8)

advmod(bright-10, very-9)

amod(screen-11, bright-10)

dobj(has-7, screen-11)

Figure 2.1 Output of typed dependency parsing.

Figure 2.2 Dependency representation of words.

Parsing generated the Stanford typed dependencies output showing the semantic relation between the bigrams. Governor words are the nodes that branch out to the dependent word in the phrase tree structure. The generated phrase tree structure provides an overview of the grammatical dependencies between words and clauses within a sentence. The use of typed dependency relationship between words provided the means to analyze the semantic orientation of words by considering the grammatical relations of words and the syntactic structure of the sentence. By doing so would allow longer distance words to be analyzed without neglecting the semantic and syntactic significance that could impact sentiment prediction performance.

Cer et al. (2010) reported the performance of the Stanford parser in generating the typed dependencies has a F1-score of up to 87%. Errors are attributed to the incorrect phrase structures which led to difficulty in labeling a relationship within a phrase tree.

2.4.3 Typed dependency Pattern Rules

A common approach to extract patterns within word dependencies for sentiment detection was through heuristic means (Moilanen and Pulman, 2007; Shaikh et al., 2008; Thet et al., 2010). Moilanen and Pulman (2007) proposed a sentiment composition model based on the concept that the global polarity of a sentence is a function of the polarities of its parts. Specifically, the model combined two input constituents of any dependency type or size and calculated a global polarity for the resultant composite output constituent. For example, a rule ($OUT^{aij} \rightarrow SPR^{ai} + SUB^{aj}$) includes the consequent (OUT), the superordinate (SPR), which is the stronger of the input constituents and the dominated constituent subordinate (SUB). The polarity (α) of the OUT constituent was determined by the SPR constituent and the compositional

processes executed by the SPR constituent on the SUB constituent. The superscripts i and j were tags that specified how the SPR constituent could alter the polarity of the accompanying SUB constituent. Possible values for the tags include: default, positive, negative, and reverse. The model assumed non-neutral sentiment polarity over neutral polarity and handled negation through polarity reversal. Polarity resolution was achieved by ranking the input constituents based on their assigned weights. Shaikh et al. (2008) considered the semantic relationship between textual components in a sentence and the computation of contextual valence of the used words. Semantic processing of the input text was based on the dependency analysis of each semantic verb frame, which was composed of the frame-invoking verb with its corresponding subject and object. Cognitive and common sense knowledge resources were used in the valence scoring of words. Rules were then used to calculate contextual valence to support word sense disambiguation, assess the valence of the semantic verb frame, and to assign overall valence to the whole sentence. Examples of rules used were $(\text{ADJ}_{\text{pos}} + (\text{CON}_{\text{neg}} \text{ or } \text{NE}_{\text{neg}}) \rightarrow \text{negative valence})$ and $(\text{ADJ}_{\text{pos}} + (\text{CON}_{\text{pos}} \text{ or } \text{NE}_{\text{pos}}) \rightarrow \text{positive valence})$ where the adjectives (ADJ), concepts (CON), and named entities (NE) were assigned with their valence groupings.

Thet et al. (2010) proposed the use of heuristic rules to compute the sentiment of a clause from the prior sentiment score assigned to individual words with consideration of the clause grammatical relations. The method adopts a linguistic approach of computing the sentiment of dependency structure of the clause. The subject or object of clause can be a noun phrase which consists of a noun and an adjective. When inputs are of the same sentiment orientation, they tend to intensify each other. Hence, the absolute value of the output should be larger than the absolute values of the inputs. For example, the adjective “*great*” in the phrase “*great art*”

intensifies the positive noun “*art*”. Thus, the phrase “*great art*” becomes more positive than the noun “*art*” itself. The rules used to compute the contextual sentiment for subject or object are given in Table 2.1. Thet et al. (2010) further analyzed the rules for verb phrase, predicate, and clause to handle the contextual sentiments of words.

Table 2.1 Contextual sentiment for subject or object (Thet et al., 2010).

ID	(A)djective	(N)oun	Output	Examples
F1.1	+ / 0*	+ / 0	+	great art
F1.2	+ / 0	-	-	Big failure
F1.3	-	+ / 0	-	Lousy star
F1.4	-	-	-	Worst disaster
*: 0 indicates neutral				
Formula				
F1.1	Positive A and Positive N (positive includes neutral) $\Rightarrow + (N + (1- N) * A)$			
F1.2	Positive A and Negative N $\Rightarrow - (N + (1- N) * A)$			
F1.3	Negative A and Positive N $\Rightarrow A$			
F1.4	Negative A and Negative N $\Rightarrow - (N + (1- N) * A)$			

The manual approach used by the previous studies to derive pattern rules to predict sentiments has its limitations. Prior knowledge on the possible rules is required in the manual approach as compared to an automatic approach that learns the pattern rules through the annotated training data. A manual approach may not provide substantial rule coverage as the comprehensiveness of the pattern rules is highly dependent on the knowledge base used to derive the rules. An automatic approach should generate a more comprehensive set of rules to provide better sentiment classification performance. The automatic sentiment analysis approach used in this study is further explained in Chapter 4.

2.4.4 Semantic Analysis

It is observed in the research study that there were many words whose semantic orientations depended on contexts in which they appear. Wilson et al. (2005) recognized there are more complex relationships between words in identifying the contextual polarity of phrases. Positive words could be used in phrases expressing negative sentiments, and vice versa. Words that are positive or negative out of context could be neutral in context, e.g., negative “*fiction*” in neutral “*science fiction*”. Conversely, words that are neutral out of context could be subjective in context, e.g., neutral “*red*” in positive “*red carpet*”. A simple pattern rules approach based on the prior sentiment lexicon approach may not sufficiently handle all the complex relationships between words. Often these complex relationships are the cause of polarity conflict situations where the polarity outcome is undecided. Further fine-grained analysis into these more subtle word relationships is required to better understand the context and resolve the polarity conflict situations. This study identified negation, domain terms, and shifters as possible complex word relationships. Negation, which may be local (e.g., “*not good*”) or involve longer distance dependencies such as negation of the proposition (e.g., “*does not look very good*”) or the negation of subject (e.g., “*no one thinks that it’s good*”). There are also negation words that intensify rather than change polarity (e.g., “*not only good but amazing*”). Domain terms could contain subjectivity when out of context but are neutral when considered within the domain vocabulary, e.g., positive “*star*” in neutral “*star trek*”. Shifters words which have inherent semantic properties to impact contextual polarity were discussed in details in Polanyi and Zaenen (2006). They proposed that the organization of text could contribute information relevant to assessing attitude. This study observed that

how lexical items reflect attitudes is more complex than simply counting the valences of terms. For example in the following text:

"Of course, that would not stop deregulation of the power industry altogether. The blunderbuss(-) of state initiatives will see to that. However, by prolonging uncertainty(-) it would needlessly(-) delay(-) the arrival of the bonanza(+) of benefits(+) that consumers deserve(+) and give them legitimate(+) grounds for their cynicism(-)."

While there are six negative(-) lexical items and only four positive(+) items in the text, readers do not conclude that the author is negative about the topic “deregulation”. The basic valence of individual lexical terms may be strengthened or weakened by context provided by the presence of other lexical items, the genre type and discourse structure of the text, and cultural factors. While some terms in a text are inherently positive or negative, others change base valence according to context from the domain of discourse or from other lexical terms nearby in the document. This study defined shifters as lexical items that exhibit the phenomena that can cause the valence of their modifying lexical item to shift from one polarity to the other or, less forcefully, to modify the valence towards a more neutral position. Examples of identified shifters are negatives, intensifiers, modals, presuppositional items, and connectors. Negative modifiers reverse the valence of a term, and examples include “never”, “none”, “nowhere”, etc. There are other modifiers that intensify valence rather than negate, for example “rather” in “rather efficient” and “deeply” in “deeply suspicious” weaken and strengthen the valence of the modified term respectively.

Wilson et al. (2005) considered a word to be an intensifier if it appears in a list of intensifiers and if it precedes a word of the appropriate part of speech (e.g., an

intensifier adjective must come before a noun). This research study compiled the intensifier words from those listed in Quirk et al. (1985), intensifiers identified from existing entries in the subjectivity lexicon, and intensifiers identified during exploration of the developmental data. Quirk et al. (1985) described the effects of intensifiers as scaling upwards from an assumed norm, and identified two intensifier types: maximizers (e.g., absolutely, completely, and perfectly) and boosters (e.g., very much, a lot, deeply). Maximizers denote the upper extreme of the intensity scale, and boosters denote a high point on the intensity scale. It is noted the subtypes provide nothing more than a rough guide to semantic distinction, because the varying effects of intensifiers represent a semantic gradient, which is obscured by a clear-cut division into classes. It was also observed that intensification is realized for the most part by adverbs, but occasionally also by noun phrases and prepositional phrases. Mitigators, also termed as downtoners, have a generally lowering effect on the force of the verb or predication and many of them apply a scale to gradable verbs. Mitigators as described by Quirk et al. (1985) can be categorized into four groups: approximator, compromisers, diminishers, and minimizers. Approximators (e.g., almost, nearly, as good as) serve to express an approximation to the force of the verb, while indicating that the verb concerned expresses more than is relevant. Compromisers (e.g., kind of, sort of, quite, rather) have only a slight lowering effect and tend to call in question the appropriateness of the verb concerned. Diminishers (e.g., mildly, partly, somewhat) scale downwards and roughly mean to “*to a small extent*”. Minimizers (e.g., barely, hardly, little) are negative maximizers.

Polanyi and Zaenen (2006) further described modal operators such as “*if*”, “*might*”, “*could*”, “*possibly*”, etc., set up a context of possibility (non-real events) and necessity (real events) where the valence expressed by terms may not necessary

reflect the author's opinion on the target. For example, "*Mary is a terrible person. She is mean to her dogs.*" is a realis event. However, the sentence "If Mary was a terrible person, she would be mean to her dogs" is a non-realis event which asserts neither that Mary is a terrible person nor that she is mean to her dogs. Rather the modal operators "if" and "would" neutralize the base valence of the terms "terrible" and "mean". Often words shift the valence of evaluative terms through their presuppositions, implicitly assuming the background belief of an expression to be true. For example, the expression "*barely sufficient*" presupposes that something more sufficient is expected. Words like "*barely*" and "*even*" were marked in the lexicon as shifter words. Other shifter words include nouns such as "*failure*" in "*failure to succeed*" and verbs like "*fail*" in "*failed to pass*". Connectors such as "*although*", "*however*", "*but*", etc., could mitigate the force of the valence found in sentences. For example, "*Although Boris is brilliant at math, he is a terrible teacher*". The force of "*although*" effectively negates the positive polarity of "brilliant".

Ding et al. (2008) proposed an approach to opinion mining by exploiting external evidences and linguistic conventions of natural language expressions. Several context handling rules were given. In a sentence where the opinion is unsure, the intra-sentence conjunction rule tries to detect the opinion expressed by the conjoining clause of the sentence by another reviewer. For example, the sentence "*the battery life is long*", and the opinion expressed by "*long*" is unsure to be positive or negative. The opinion could be detected from the conjoining expression "*which is great*" in "*The camera has a long battery life, which is great*" sentence generated by another reviewer. Additionally, the context of the previous or next sentence (or clause) could be used to decide the opinion of the sentence. Normally, people express the same opinion across sentences unless there is an opinion change using words such as "*but*" and "*however*".

Hence, the unknown orientation of a current sentence could be detected from the last clause of the previous sentence and if not, then from the first clause of the next sentence, provided the opinion changing words are not found. Further study on specific word ordering and semantic of words in consideration of these complex relationships will improve performance in predicting sentiment polarity. This research study further analyzed the semantic of words to build specialized lexicons based on added knowledge of the complex word relationships with the details given in Chapter 4.

2.4.5 Discourse Level Analysis

Beyond the pattern rules at the typed dependency level, it is observed there are discourse relations between phrases in a sentence. An understanding and analysis on the discourse relations between phrases could reveal polarity pattern rules at discourse level. A discourse relation is a description of how two segments of discourse are logically connected to one another. The Penn Discourse Treebank (PDTB) (Prasad et al., 2006) is a large-scale corpus annotated with information related to discourse structure and discourse semantics. It contains annotations of explicit and implicit discourse connectives and their arguments on the one million word Wall Street Journal corpus. The four classes of explicit connectives annotated in the PDTB are:

- (1) subordinating conjunctions, both bare (e.g., because, when, since, although) and with a modifier (e.g., only because, particularly since, even after),
- (2) subordinators (e.g., in order that, expect that),
- (3) coordinating conjunctions (e.g., and, or, nor), and
- (4) discourse adverbials (e.g., however, otherwise, then)

Implicit connectives are annotated in the PDTB between adjacent sentences when no connective appears explicitly to relate the second sentence to the first. For example, in the following sentence,

"Also unlike Mr. Ruder, Mr. Breeden appears to be in a position to get somewhere with his agenda. IMPLICIT = BECAUSE As a former White House aide who worked closely with Congress, he is savvy in the ways of Washington."

The PDTB includes sense distinctions for polysemous connectives (e.g., since, while, if, when, because). Possible senses for the polysemous connectives include TEMPORAL (relating to time), CAUSAL (relating to cause), and both CAUSAL and TEMPORAL. Attribution, which has to do with ascribing beliefs and assertions expressed in text to the agent(s) holding or making them, is annotated in the PDTB to primarily distinguish between two different sources of attribution, the writer of the text or some other agent mentioned by the writer. The PDTB was used by Somasundaran and Wiebe (2009) to analyze discourse level opinion relations and by Ruppenhofer et al. (2008) for finding sources and targets of subjective expressions. Though a clause-level discourse relationship analysis might have provided better performance, this study limits its scope to using the connective terms identified through the typed dependencies to identify the discourse relations in a sentence.

2.4.6 Lexicon Building

Automatic sentiment polarity detection often relies on a semantic orientation lexicon of positive and negative words (Turney, 2002; Wilson et al., 2005; Pang and Lee, 2008). Examples of words of positive polarity are “good”, “excellent”, “impressed”, etc., while words of negative polarity include “sad”, “bad”, “lousy”, etc.

Creating a semantic orientation lexicon usually tends to be computationally expensive and to rely on a significant amount of manual annotation and large corpora.

There are various references and dictionaries built for natural language processing applications, e.g., General Inquirer (Stone et al., 1966), WordNet (Fellbaum, 1998), FrameNet (Baker et al., 1998), and the OpinionFinder subjectivity lexicon (Wilson et al., 2005). The General Inquirer is essentially a mapping tool that maps words to their identified categories. The database contains more than 4000 unique words, most of which are adjectives. For each word, it defines the categories the word is in; some categories indicate whether the word is usually used in the positive or negative sense. Types of category include emotions (sub-categories: arousal, urge, affection, etc.), evaluation (sub-categories: good, bad, and ought), positive and negative adjectives, and others. WordNet is a lexical reference system in which English words are grouped into sets of synonyms called synsets, and the semantic relations between the synsets are provided. The database contains approximately 150,000 words organized in 117,000 synsets for a total of 206,000 word-sense pairs. FrameNet is a lexical resource based on frame semantics, which is a conceptual structure that describes a particular type of situation, object, or event along with its participant and props. For example, the Apply_heat frame describes a common situation involving a COOK, some FOOD and a HEATING_INSTRUMENT, and is evoked by words such as bake, boil, simmer, steam, etc. These roles are called frame element (Fes) and the frame-evoking words are lexical units (LU) in the Apply_heat frame. For example, in the following frame, [Cook Matilde] fried [Food the catfish] [Heating_instrument in a heavy iron skillet] the frame-evoking LU is “fried” and the Fes are its syntactic dependents. OpinionFinder is a system that processes documents and automatically identifies subjective sentences as well as various aspects of subjectivity

within sentences, including agents who are sources of opinion, direct subjective expressions and speech events, and sentiment expressions. The subjectivity lexicon that is part of OpinionFinder contains a list of subjectivity clues which were mainly compiled from the work reported by Riloff and Wiebe (2003). The MPQA Opinion Corpus contains news articles from a wide variety of news sources manually annotated for opinions and other private states (i.e., beliefs, emotions, sentiments, speculations, etc.).

Certain semantic orientation lexicons are manually annotated and compiled. But often these manual lexicons are limited in their number of words. Examples of manually compiled lexicons are the General Inquirer (Stone et al., 1966), which has orientation for only about 4000 entries, and the Pittsburgh subjectivity lexicon (Wilson et al., 2005) with about 8,000 words with semantic orientation labels. Reliance on manually created lexical resources and the use of large corpora are two common approaches to the automatic creation of semantic orientation lexicons. Esuli and Sebastiani (2006) created the SentiWordNet lexicon by using a supervised algorithm to attach sentiment scores to the WordNet dataset. They trained a set of classifiers on the WordNet dataset for the classification and assignment of the semantic polarity word scores. Turney and Littman (2003) used a corpus of 350 million web-pages and applied a supervised algorithm based on the probability of words to co-occur with a set of positive or negative words. Hatzivassiloglou and McKeown (1997) used a 21 million word corpus and a supervised algorithm to determine the semantic orientation of adjectives. They generated a graph with adjectives as nodes. An edge between two nodes indicates whether the two nodes (adjectives) have a similar or different semantic orientation. The graph is then partitioned into two sub-graphs using a clustering algorithm. The nodes within a sub-graph have the same semantic orientation. The sub-

graph with adjectives that occur more often in text is tagged positive while the other sub-graph is tagged negative. Qian et al. (2009) built a semantic lexicon of English nouns via bootstrapping. The bootstrapping algorithm discovers words with semantic properties similar to a small set of labeled seed words, which could be manually selected from an existing lexicon. The bootstrapping procedure iteratively trains a perceptron-based linear classifier. A perceptron algorithm determines whether the active features of a test case are similar to those learned from given categories of examples. In an iterative training process, the classifier first learns from a small seed set, which contains examples of all categories (in binary classification, both positive and negative examples) manually selected to reflect human knowledge of semantic categories. The classifier then discovers new instances of each category. Based on activation values, these newly discovered instances are selectively admitted into the original training set, which increases the size of training examples for the next iteration. However, the process has two apparent disadvantages: new examples may be wrongly classified by the model, and it is difficult to evaluate the discriminative models produced in successive iterations.

Due to the limited scope of manually created lexicons, other studies have tried to expand the lexicon by discovering through a thesaurus, words similar in semantic orientation with the seed words. Cai et al. (2008) used the approach of augmenting the Inquirer database and WordNet to derive polarity words. By associating the polarity of the words found in Inquirer² with their synonyms found in WordNet, the database of the polarity word lexicon could be expanded. Mohammad et al. (2009) proposed an approach to generate a semantic orientation lexicon, which includes both individual words and multi-word expressions using a Roget-like thesaurus and a list of affixes.

---

² <http://www.wjh.harvard.edu/~inquirer/>

The study automatically identified seed words from a list of affix patterns. For example, words such as dishonest, unhappy, and impure, tend to have negative semantic orientation, whereas others like honest, happy, and pure tend to have positive semantic orientation. Subsequently, Roget's thesaurus is used to further generalize the seeds to expand the words with similar semantic orientation.

Another thesaurus-based work on sentiment analysis is by Aman and Szpakowicz (2007) on detecting emotions such as happiness, sadness, and anger. The General Inquirer (Stone et al., 1966) and WordNet-Affect (Strapparava and Valitutti, 2004) lexical resources were used in the work to automatically recognize emotion words. Yang et al. (2007) mined the relationships between words and emotions using weblog corpora and constructed an emotion lexicon consisting of words that collocate with a special set of graphic icons called emoticons. Different kinds of emoticons could be introduced into text expressions to convey bloggers' emotions. They adopted a variation of pointwise mutual information (Manning and Schutze, 1999) to measure the collocation strength $co(e, w)$ between an emotion e and a word w :

$$co(e, w) = c(e, w) \times \log \frac{P(e, w)}{P(e)P(w)} \quad (\text{Eq. 2.1})$$

where $P(e, w) = c(e, w)/N$, $P(e) = c(e)/N$, $P(w) = c(w)/N$, $c(e)$ and $c(w)$ are the total occurrences of emotion e and word w in a tagged corpus, respectively, $c(e, w)$ is total co-occurrences of e and w , and N denotes the total word occurrences. A word entry of a lexicon may contain several emotion senses, which are ordered by the collocation strength (co). When all collocations (i.e. word-emotion pairs) are listed in descending order of the collocation strength, the top n threshold collocations are used to build the emotion lexicon. For a sentence S to be classified consisting of n emotion words, the

emotion of S is derived by mapping from a set of n emotion words to m emotion categories. Their best performing method considered all senses of the emotion word as votes and each emotion word votes only one sense that has the maximum collocation strength. The sentence S is then labeled with the emotion that receives the most votes. Though this study limits its scope to the analysis of bloggers' engagement, persuasion, and persona influence styles, future work could involve the study of bloggers' emotions to provide further details on the bloggers' influence styles.

In the study by Thet et al. (2010), a domain specific lexicon and a generic opinion lexicon are created to store the prior sentiment scores of opinion words. The domain specific lexicon contains movie domain specific opinion words, and the generic opinion lexicon holds general opinion words derived from SentiWordNet and the subjectivity lexicon. Their method prepared a set of domain specific opinion words for the movie review domain by semi-automatically analyzing the message posts on discussion boards and user movie reviews. Words and their information gain measure are extracted and calculated from the training dataset. Information gain is frequently used as a term goodness measure in the areas of machine learning and information retrieval (Mitchell, 1997; Robertson and Jones, 1976). Opinion words with high information gain are manually examined and added to the domain specific lexicon if they are found to be domain specific. In the same study, SentiWordNet is used to collect general opinion words and derive their prior sentiment scores. Then the subjectivity lexicon of Wilson et al. (2005) to determine the overall orientation of the words in the generic opinion lexicon while the prior scores are calculated based on the scores of their senses in SentiWordNet. Table 2.2 summarizes how the prior sentiment scores of the words in the generic opinion lexicon are derived from the subjectivity lexicon and SentiWordNet in Thet et al. (2010).

Table 2.2 Derivation of sentiment scores (Thet et al., 2010).

		Subjectivity				
		Strongly Positive	Weakly Positive	Strongly Negative	Weakly Negative	NA
Generic Opinion Lexicon	Positive	A	B	C	D	E
	Negative	F	G	H	I	J
	NA	K	L	M	N	
Formula						
A, F = Max (positive scores of multiple senses)						
B, G, E = Avg (positive scores of multiple senses)						
C, H = Min (negative scores of multiple senses)						
D, I, J = Avg (negative scores of multiple senses)						
$K = (\text{Avg (scores of all instances of A)} + \text{Avg (scores of all instances of F)}) / 2$						
$L = (\text{Avg (scores of all instances of B)} + \text{Avg (scores of all instances of G)}) / 2$						
$M = (\text{Avg (scores of all instances of C)} + \text{Avg (scores of all instances of H)}) / 2$						
$N = (\text{Avg (scores of all instances of D)} + \text{Avg (scores of all instances of I)}) / 2$						

2.4.7 Topic Detection

Li and Croft (2006) assumed that finding information patterns such as combinations of named entities, phrases and other sentence patterns may contain more important and relevant information than single words. Their study used sentence level information patterns to detect novel and relevant topics from the textual content, and applied the information patterns features as weights in an information retrieval query based on the Term Frequency Inverse Document Frequency (TFIDF) formula. Information pattern analysis includes sentence lengths, opinion patterns, and named entities. It was found that relevant topic-related sentences have more words (average 15 words) than non-relevant topic-related sentences (average 9 words). An opinion sentence is defined as a sentence that contains opinion patterns, and the assumption is that opinion sentences are more likely to be relevant topic sentences compared with non-relevant topic sentences. The opinion patterns were extracted from scanning a few documents in the data collection, examples of which are “*said*”, “*according to*”, “*finds that*”, etc. It was observed that there are relatively more opinion sentences in relevant

topic related sentences than non-relevant topic-related sentences. However, the detection of patterns at sentence level would not be able to consider the context such as the specific subject or topic that is referred to in the sentence as compared to a typed dependency approach. For example, in the sentence “*I bought the iPhone and Blackberry and they are both great phones*”, a general sentence analysis approach may not detect “*both*” as referring to “*iPhone*” and “*Blackberry*”. In addition, the use of the TFIDF approach is not applicable to a fine-grained analysis of topic detection due to the loss of context and term ordering information. Moreover, the assumption of term independence for TFIDF is also not valid.

Cai et al. (2008) leveraged sentiment classification on web content to detect topics that are highly correlated with positive and negative opinions. The study detailed a topic-detection method using point-wise mutual information and term frequency distribution. Their approach focused on snippets, which are small text segments around a specified keyword in a given document. The sentiment topic recognition is evaluated from the Pointwise Mutual Information (PMI) value and the word support in category. PMI measures the association between two discrete random variables, and the value reflects their dependence. A zero value indicates the two variables are independent, a greater than zero value shows association, and the variables are complementary if the value is less than zero. The algorithm to calculate the PMI value of word w against a sentiment category s is as follow.

$$PMI(w,s) = \log(p(w,s)/(p(s)*(p(w)+0.05))) \quad (\text{Eq. 2.2})$$

where the $p(w,s)$ describes the co-occurrence between w and s , $p(s)$ represents the distribution of category s and $p(w)$ evaluates the distribution of word w in the whole

snippet collection. Word support evaluates the importance of word in category and is calculated as follow.

$$Freq(w, s) = N(w, s) / \sum_{s \in +, -, 0} N(w, s) \quad (\text{Eq. 2.3})$$

where $N(w, s)$ denotes the number of word w in category s . By evaluating PMI and word support simultaneously, the relevant sentiment topic words can be detected, as compared to evaluating them individually. For example, suppose that words A and B have equal PMI value for the positive category, but their frequencies in the positive category are 1 and 100 respectively. Without the word support metric, the importance of A and B to the positive category may be considered equal, which could be misleading. Though Cai et al. (2008)'s approach evaluates the dominant sentiment class of the identified topic word, it does not consider the context of the textual data content. This study further explores other works that considered textual context by using the dependency structure approach.

Lee et al. (2007) proposed the use of a dependency structure language model for topic detection and tracking (TDT) to compensate for weaknesses in the unigram and bigram language models. A dependency structure language model was able to capture the long-distance dependencies between words, hence improving the performance of the model. Somasundaran and Wiebe (2009) used the subjectivity lexicon of Wilson et al. (2005), which contains approximately 8000 words, to tag the prior priority of words. To pair opinions with targets, they used a rule-based system based on dependency parse information. Examples of syntactic rules for finding targets of opinions are shown in Table 2.3. Once these opinion-target pairs are created, the identity of the opinion word is masked by replacing the word with its polarity with the

opinion-target pair converted to a polarity-target pair. The abstraction is essential due to the sparseness of the data.

Table 2.3 Examples of syntactic rules for finding targets of opinions
(Somasundaran and Wiebe, 2009).

DIRECT OBJECT Rule: <code>dobj(opinion, target)</code> In words: The target is the direct object of the opinion Example: I love_{opinion1} Firefox_{target1} and defended_{opinion2} it_{target2}
NOMINAL SUBJECT Rule: <code>nsubj(opinion, target)</code> In words: The target is the subject of the opinion Example: IE_{target} breaks_{opinion} with everything.
ADJECTIVAL MODIFIER Rule: <code>amod(target, opinion)</code> In words: The opinion is an adjectival modifier of the target Example: The annoying_{opinion} popup_{target}
PREPOSITIONAL OBJECT Rule: <code>if prep(target1, IN))pobj(IN, target2)</code> In words: The prepositional object of a known target is also a target of the same opinion Example: The annoying_{opinion} popup_{target1} in IE_{target2} (“popup” and “IE” are targets of “annoying”)
RECURSIVE MODIFIERS Rule: <code>if conj(adj2, opinion_{adj1}))amod(target, adj2)</code> In words: If the opinion is an adjective (adj1) and it is conjoined with another adjective (adj2), then the opinion is tied to what adj2 modifies Example: It is a powerful_{opinion(adj1)} and easy_{opinion(adj2)} application_{target} (“powerful” is attached to the target “application” via the adjective “easy”)

In online blogs, bloggers express their preference and argue why their viewpoints are correct. It is common for these opinions to be expressed on associated features or aspects of the topic discussed. For example, “*I like the large screen size of the iPhone*” and “*The keyboard of the Blackberry is too small*”. Thus, it is important to consider not only the semantic orientation of the expressions, but also what the opinions are targeting. Product review mining work has explored finding opinions with respect to aspects (Hu and Liu, 2004; Popescu et al., 2005). Ruppenderhofer et al., (2008) also observed a distinction between global targets and local targets, where in product reviews, opinions are expressed both towards the product as a whole, the global target, and towards some of its features, the local targets. For example, in the text below:

"The computer is very good and very easy to use. It has a built in camera and bluetooth; the all singing and dancing machine. Love it. The only glitch is the scrolling pad is not as smooth as my last Toshiba notebook. One thing is that Vista is a nightmare... "

“the computer” is the global target and the local targets are the features, “the scrolling pad”, “the Vista OS”, “the camera”, and “the Bluetooth interface”. A problem highlighted is the relating of sentiments about local targets to the global targets of a discourse. A number of sentiments towards local targets may not add up in a simple linear fashion as an overriding sentiment could sway the attitude towards the global target from an expected outcome based simply on counting opinions. The study proposed finding targets, specifically in arguing, through the use of conditional (*if-clause*) and causal connective (*because-clause*) constructions. Depending on how the opinion expressions are presented, the target may be found in the conditional and causal clause or in the main clause. It was also observed in the study that the target of an opinion expression could be a proposition or an entity, which determines the span of the opinion expressed. For example, in the sentence “*In his view, Kao said [the cross-strait balance of military power] is not critical to the ROC’s national security*”, the target is the argued-for proposition “*the cross-strait balance of military power*”. Another observation was that the prominence of entities could be used to infer targets of the opinion expressed. The study further proposed a study in co-reference relations between unexpressed semantic phrases with definite reference and the overt phrases they co-refer with using the Penn Discourse Treebank as a possible resource. Further to the findings gathered from previous studies on opinion target detection, this research study leverages the use of typed dependencies to analyze the relationships between the

opinions and targets and improve performance through providing a fine-level of syntactic and semantic analysis of the words in a sentence.

2.5 Influence Diffusion Detection Models

Processes of influence propagation through a network have been studied in various applications such as viral marketing (Domingos, 2005), contagion models (Dodds and Watts, 2004) and blog networks (Leskovec et al., 2007a). Domingos (2005) attempted to build a model that maximize the influence of a product based on the concept of viral marketing where the potential of each customer (node in the network) to spread the influence is measured to determine the best possible marketing strategy for the product. Factors identified to affect the model include the connectivity of the customer, the customer's sentiments towards the product, opinion leaders or proximity to opinion leaders in the network. Previous studies (Bailey, 1975; Girvan et al., 2002) investigating the flow of information through networks using the contagion models used the analogy between the spread of disease and the spread of information in networks. These models were based on the susceptible person being infected with a disease with some probability. Analogously, the propagation of information through the blogosphere could be viewed as the spread of disease through the network where a blogger is exposed to the topic by reading a blog post and decided to write about the topic, becoming infected. The topic may then spread to other readers. Leskovec et al. (2007a) presented a model that mimics the spread of information on the blogosphere based on the information cascade properties and cascading pattern found in the network. The results from the study showed that the simulated cascades generated from the model had a very good agreement with the real cascades found in the blog

network, where the model in turn could be used to predict the influence diffusion pattern in a network. The top 10 most frequent cascades discovered by the study in Leskovec et al. (2007a) are shown in Figure 2.3.

Figure 2.3 Top 10 most frequent cascades discovered.

The existing diffusion studies can be broadly divided into the macro-level studies (Meade and Islam, 2006; Iribarren and Moro, 2011) and micro-level studies (Granovetter, 1978; Chen et al., 2010b; Goldenberg et al., 2001; Kempe et al., 2003). The macro-level studies model the aggregated number of adoptions in the population throughout the diffusion process, where features commonly used in the macro-level studies include rate of adoption, number of adopters at any time point, and peak of adoptions. One common assumption in the macro-level models is that each node in the network is related to every other node, i.e., a complete network. Based on this assumption, the Bass Model (Bass, 1969) has been applied in previous studies (Meade and Islam 2006; Iribarren and Moro 2011). The Bass Model presents a rationale for how current adopters and potential adopters of a new concept interact and is represented by the following equation:

$$\frac{f(t)}{1-F(t)} = p + qF(t) \quad (\text{Eq. 2.4})$$

where $f(t)$ is the rate of change of the installed base fraction, $F(t)$ is the installed base fraction, p is the coefficient of innovation, and q is the coefficient of imitation.

Coefficient p represents external influence or advertising effect, while coefficient q gives the effect of internal influence or word-of-mouth effect. Typical values of p and q determined through experiments are 0.03 and 0.38 respectively (Mahajan et al., 1995). The typical shape of the Bass curve is shown in Figure 2.4, where the number of new adopters is the sum of the innovators and imitators.

Figure 2.4 Typical Bass curve.

It is seen that the number of innovators will diminish and the number of new adopters would increase then fall over time, which is synonymous to the influence diffusion pattern of new influence in a social network.

Studies (Shaikh et al., 2010; Cowan and Jonard, 2004) also assumed that the nodes are connected in a non-complete network, which allows influence to be diffused to nodes only through neighboring adopters. Shaikh et al. (2010) nests the Bass model and Non-Uniform Influence model (Easingwood et al., 1983) to assess the incremental value of social network structures for explaining the patterns of diffusion. Similarly, Cowan and Jonard (2004) explores how the topology of nodes' (agents) interactions

influence aggregate regularities to capture the effects of incremental innovations and their diffusion among a network of heterogeneous agents.

In contrast, the micro-level study focuses on modelling each node as an agent making local adoption decisions based on the neighbors' adoptions. Micro-level diffusion models require complete knowledge of node-to-node relationships, and the knowledge of influence probability between two neighboring nodes. Examples of these models include the Linear Threshold model (Granovetter 1978; Chen et al., 2010b) and In-dependent Cascade model (Goldenberg et al., 2001; Kempe et al., 2003; Saito et al., 2010; Gomez et al., 2010), the stochastic model for opinion formation (Stauffer, 2002), and the model on the impact of the network characteristics on innovation diffusion (Liu et al., 2005). While the macro-level models are used in situations where the detailed network structure is not given or when the influence probabilities are missing due to privacy or data non-availability reasons, this research study adopted a micro-level approach through providing a detailed analysis on the influence relationships between the linked bloggers.

Directed graphs were used to model the spread of an idea or innovation through a network. Each individual node in the graph is either active (indicating adaptation of the innovation) or inactive. The motivation is that each node's tendency to become active increases monotonically as more of its neighbors become active. Granovetter (1978) proposed the Linear Threshold (LT) model to capture the diffusion process based on the use of node-specific thresholds. In the Linear Threshold model, a node v is influenced by each neighbor w according to a weight $b_{v,w}$ such that $\sum_{w \text{ neighbor of } v} b_{v,w} \leq 1$ as shown in Figure 2.5. Each node would choose a threshold θ_v uniformly at random from the interval $[0, 1]$.

Figure 2.5 Diffusion process in a Linear Threshold model.

The threshold represents the weighted fraction of v 's neighbors that must become active in order for v to become active. Given a random choice of thresholds and an initial set of active nodes, the diffusion process unfolds deterministically in discrete steps. In step t , all nodes that were active in step $t - 1$ remain active, and any node v for which the total weight of its active neighbors is at least θ_v , as shown in Equation 2.1 would be activated.

$$\sum_{w \text{ neighbor of } v} b_{v,w} \geq \theta_v \quad (\text{Eq. 2.5})$$

The threshold θ_v intuitively represents the different latent tendencies of nodes to adopt the innovation (to be influenced) when their neighbors do. Random selection of threshold values is in effect averaging over possible threshold values for all nodes. Other approaches hard-wired the thresholds at a known value like $\frac{1}{2}$ (Berger, 2001; Morris, 2000)

Other studies used probability theory to consider dynamic cascade models for the diffusion process. Goldenberg et al. (2001) investigated the Independent Cascade (IC) model, where the network is defined as a directed network (or equivalently graph) $G = (V, E)$, where V is the set of nodes and E a set of links, where each link is denoted by $e = (v, w) \in E$ and $v \neq w$, meaning there exists a directed link from a node v to a

node w . For each directed link, a real value $\kappa_{v,w}$ with $0 < \kappa_{v,w} < 1$ is specified as the diffusion probability through link (v, w) . In the Independent Cascade model, the diffusion process proceeds from a given initial active set $D(0)$ in the following way. When a node $v(\in D(t))$ first becomes active at time-step t , it is given a single chance to activate each currently inactive child node w , and the attempt succeeds with probability $\kappa_{v,w}$. If v succeeds, then w becomes active at time-step $t+1$, i.e., $w(\in D(t+1))$. If multiple parent nodes of w first become active at time-step t , then their activation attempts are sequenced in an arbitrary order, but all the attempts are performed at time-step t as shown in Figure 2.6.

Figure 2.6 Diffusion process in an Independent Cascade model.

Whether or not v succeeds, v will not make any further attempts to activate w in subsequent rounds. The diffusion process starts with an initial activated set of nodes, and then continues until no further activation is possible. Information diffusion in the blogosphere is captured correctly only when the diffusion probability is assigned reflecting well the strength of a relationship between the linked pair of bloggers.

The Independent Cascade (IC) model and Linear Threshold (LT) model are two main diffusion models used in the study of influence propagation. In the Independent Cascade model, each edge could independently activate their neighboring nodes based on an activation probability. While in the Linear Threshold model, each edge has a

weight and each vertex has a randomly assigned threshold value such that the vertex is activated if the weighted sum of its active neighbors exceeds its threshold. The two models characterize two different aspects of social interaction. The Independent Cascade approach models individual and independent interaction and influence among linked nodes. On the other hand, the Linear Threshold model focuses on the threshold behavior in influence propagation, where consideration is given to all linked neighbors in influencing the node.

Other variations of models based on the two fundamental Independent Cascade and Linear Threshold approaches were proposed. Kempe et al. (2003) proposed the General Cascade model which generalizes the Independent Cascade model and Threshold models by discharging the independence assumption. Yang and Leskovec (2010) developed the Linear Influence Model where the number of newly infected nodes was modeled as a function of which other nodes were infected in the past. For each node, an influence function that quantifies how many subsequent infections can be attributed to the influence of that node over time was derived, that is, influence is modeled as the volume of infected nodes along time. Kempe et al. (2003) further proposed the generalized cascading model through removing the independency assumption between the network nodes for both the IC and LT approaches in an attempt to maximize the spread of influence through the social network. Garg et al. (2012) further extend the IC model and LT model through considering the positive and negative opinions expressed in the content to improve propagation detection performance. Lim et al. (2011) used the independent cascade model to explain information diffusion in blog network, where the diffusion probability assigned to each edge between a pair of bloggers in the blog network was derived from the number of posts, comments, trackbacks, and scraps. Previous studies used sentiment analysis to

determine the influencers within a network. For instance, Kale et al. (2007) used the link structure of blog graph to associate sentiments with the links connecting two blogs. This sentiment was termed link polarity and the sign and magnitude of this value was based on the sentiment of text surrounding the link. These polar edges indicated the bias/trust/distrust between the respective blogs. A trust propagation model (Guha et al., 2004) was then used to spread this sentiment from a subset of connected blogs to other blogs to generate the fully connected polar blog graph. Analyzing just the sentiment of text surrounding the link might not give a comprehensive perspective of the information or message to be conveyed by the linked blog post. By considering the polarity of the whole post content, and not just that specific to text surrounding the links would give a more complete sentiment outlay between the two bloggers. Polarity could be further segregated based on similar topics discussed and agreement towards the linked blogger expressed by the linking blogger. The above studies used the existence of influence between the linked bloggers to detect influence propagation, and did not provide details on the influence types and styles of the bloggers in the influence diffusion process, which is considered in this research study.

Earlier works (Kimura and Saito, 2006; Leskovec et al., 2007b; Chen et al., 2010a) did not consider how to obtain the diffusion probability values which were often randomly assigned an identical value to all edges within a network for analysis. More recent studies (Xiang et al., 2010; Saito et al., 2008; Goyal et al., 2010) attempted to derive the values of diffusion probabilities to improve the performance for information diffusion prediction. Xiang et al. (2010) observed that both strong and weak ties exist between nodes in a network, and treating all relationships equally will increase the level of noise in the learned models and likely lead to degradation in performance. The study used the ancillary interaction information among the nodes

and proposed a latent variable model to infer (hidden) relationship strengths and develop a coordinate ascent optimization procedure for inference based on the assumption of homophily which postulates that people tend to form ties with other people who have similar characteristics (McPherson et al., 2001). Saito et al. (2008) presented a method for predicting diffusion probabilities for the Independent Cascade model by using the Expectation Maximization algorithm to obtain the set of information diffusion probabilities which maximizes the objective function that maps the probability of a node becoming active. Goyal et al. (2010) proposed a model that directly leverages available propagation traces (databases) to learn how influence flows in the network and used this to estimate expected influence spread under the General Threshold Model. The study extended the model by introducing temporal decay, as well as factors such as the influenceability of a specific user and influence-proneness of a certain action. While each of the above studies considered influence probabilities in their own aspect, none considered the influence styles of bloggers in computing the influence probabilities, which would have provided a clearer understanding of how influence is propagated.

Studies focused on influence maximization proposed methods to determine the optimal set of initial nodes. Kempe et al. (2003) proved the properties Monotonicity and Submodularity for choosing the initial nodes. Monotonicity states that as the set of activated nodes grows, the likelihood of a node getting activated should not decrease. Submodularity shows that the probability for an active node to activate some inactive nodes does not increase if more nodes have already attempted to activate the inactive nodes. Based on these two properties, the greedy algorithm could be used to find the set of initial nodes that could maximize influence in the network. Other studies (Java, 2006) alternatively used PageRank based heuristics to select the influential set of

bloggers that could maximize the spread of information on the blogosphere. In this study, the scope is limited to detecting influence diffusion between linked pairs of bloggers and the identification of optimal initial nodes for maximizing influence is not considered.

Previous influence diffusion studies (Liu et al., 2010; Lin et al., 2011; Weng et al., 2010) had adopted a topic-aware approach. Tang et al. (2009) proposed the Topical Affinity Propagation (TAP) to model the topic-level social influence on large networks, taking into account the predefined topic ontology or topic clusters based on pLSI (Hofmann, 1999) and LDA (Blei et al., 2003) and the existing network structure to perform topic-level influence propagation. Gruhl et al. (2004) measured the blog features metrics and further applied the metrics on diffusion models based on the Independent Cascade model and Linear Threshold model to detect influence within social networks. The influence detection was topic-based where topics were identified through the chatters and spikes in the blogosphere. Chatter refers to new thoughts on topics or reactions to topical postings by other bloggers, while spikes are the occasional event occurrence in the real world that induces a reaction from the real world that induces a reaction from the bloggers. The probability of a linking blogger writing on the article that a linked blogger has written (referred to as the copy probability) and the probability of the linking blogger reading the topic from the linked blogger (referred to as the reading probability) were computed to provide the propagation patterns of the network graph. Using these probabilities and the given initial influenced nodes, the study attempted to predict the propagation pattern. Barbieri et al. (2012) similarly proposed topic-aware extensions of the Independent Cascade and Linear Threshold models. In the topic-aware IC model, the diffusion probability derived represents the strength of the influence of the linked blogger on the

linking blogger with respect to the topics discussed, while in the topic-aware LT model, each arc is assigned a weight that is based on the respective topics. Results from the study showed that performance improved when topics are considered in the influence models. Kang et al. (2012) proposed a topic-oriented model to predict information diffusion. The Naïve Bayes approach was used to calculate the diffusion probability based on factors related to the influential effect of the bloggers on the topics. Regression analysis was then performed to determine the probability function.

Both the Linear Threshold and Independent Cascade approaches are evaluated in this study's proposed influence diffusion detection model. The Linear Threshold approach is used to study the use of influence styles in detecting influence diffusion that is initiated from all the active neighbors of a blogger versus the Independent Cascade approach, which only considers influence diffusion from a single neighbor at each step of the evaluation process. The Independent Cascade model provides specific influence analysis between two linked bloggers to provide a more precise evaluation, while the Linear Threshold model conceptualizes a more realistic situation in a blogosphere network as bloggers are usually linked to multiple bloggers. Hence, they could be influenced by their neighboring bloggers instead of only one blogger. This study aims to evaluate the performance of both models through the use of influence styles and sentiment links in detecting influence diffusion between linked bloggers.

2.6 Summary

It was observed that previous studies used various blog features to detect influence, such as number of in-links and out-links. However, the use of these features could not provide an in-depth explanation on the context to why influence is present.

This leads to the first research gap of this study (RG1: *The use of simple blog features to detect influence diffusion may not yield highly accurate results as influence itself is a complex concept that requires more in-depth analysis, where other intrinsic blog features that provide further details on the influence diffusion process could be used to improve influence diffusion detection performance*) where the relevant blog features for detecting influence diffusion are explored in Chapter 3.

The blog feature analysis study in Chapter 3 revealed that analyzing the blog post content could provide the expressed opinions and the stance of the bloggers which provided a better indication of influence. Consequently, this study adopted a linguistic approach to sentiment analysis in response to the second research question (RQ2: *How can sentiment analysis be enhanced to better detect the sentiments expressed within blog posts through considering the complex relationships between words?*). Sentiment analysis was studied in Chapter 4 and a model based on typed dependencies rules generation and subject phrases detection was developed to analyze the sentiments expressed in the linked and linking blog posts for subsequent studies in Chapter 5 and 6.

In addition to the display of sentiments on topics, bloggers could differ in the influence style, which refers to the manner the blogger exert influence through the blog postings. Previous studies that used simple blog features to detect influence in the blogosphere focused on a single notion that influence exists between blogs with hypertext links leading to the research gap (RG3: *Influence diffusion could not be described using a single detached feature based on the existence of links or similar sentiment polarity*). This literature review further studied previous works in the area of influence styles and influence diffusion detection modeling in relation to the research objective (RO3: *Develop an influence diffusion detection model using sentiment values*

between linking and linked bloggers, agreement towards the linked bloggers, and bloggers' influence styles as possible features to improve influence diffusion detection between linked bloggers). The novel concept of using bloggers' influence styles as possible features to detect influence diffusion between the linked and linking bloggers was studied in Chapter 5 and 6 in an attempt to improve influence diffusion detection between bloggers.

CHAPTER 3

ANALYSIS OF BLOG FEATURES FOR INFLUENCE DETECTION

3.1 Introduction

Chapter 3 provides details of the study to analyze the various blog features to determine correlations between the features and influence within linked blogs in relation to research gap (RG1: *The use of simple blog features to detect influence diffusion may not yield highly accurate results as influence itself is a complex concept that requires more in-depth analysis, where other intrinsic blog features that provide further details on the influence diffusion process could be used to improve influence diffusion detection performance*). This study investigated three common approaches (quantitative blog features analysis, content analysis, and community identification) and performed statistical analysis (RM1: *Perform statistical analysis to determine blog features suitable for detecting influence diffusion*) to evaluate the effectiveness of each method in detecting influence in the blogosphere. The aim is to achieve the research objective (RO1: *Identify the relevant blog features that could be used to detect influence diffusion between linked blog posts*).

Previous studies studied blog features (Adar and Adamic, 2005; Agarwal and Liu, 2008), content analysis (Matsumura et al., 2008; Song et al., 2007), and community identity (Adar and Adamic, 2005; Ghosh and Lerman, 2008) as features to detect influence. Blog features, both at the post and blog levels, analyzed in the study are *authority rank*, *text similarity*, *length of posting*, and *time difference* between postings of the linked and linking blog postings. For content analysis, sentiment and

agreement analysis are applied since sentiments expressed on common topics by two linked posts and agreement expressed by the linking blogger on the linked blog within the blog post could be indicators of influence between the blog posts.

The next section describes the blog features and model used in the analysis. Subsequently, the dataset used in this study and the experimental results are provided in section 3.3 and section 3.4 respectively. Section 3.5 provides details on additional experiments conducted on a sample product review domain dataset to validate the findings of the product review domain blogs with the political domain blogs. This is followed by a discussion on the results in section 3.6.5. Lastly, the chapter concludes with a summary of the findings in section 3.7.

3.2 Blog Features Analysis Model

Previous studies used blog features to detect influence through analyzing the in and out-links (Adar and Adamic, 2005; Agarwal and Liu, 2008) to measure the degree of influence flow. Another approach used to detect influence was content analysis (Matsumura et al., 2008; Song et al., 2007) where the common method entailed computing document similarity between the linked and linking blog posts as a measurement of influence flow. Adar and Adamic (2005) and Ghosh and Lerman (2008) studied community identity as a possible feature to detect influence diffusion. The common assumption was that blogs within the same community tend to influence each other as close affiliates would. The model for this chapter's study is based on the inference that certain blog features, including sentiments and community identity could be used to determine influence between two linked blog posts.

Table 3.1 Blog post level and blog level features analysed.

No.	Features	Definition
Blog post level data between the linked and linking blog posts		
1.	<i>Sentiment values between the two blog posts</i>	Tally of the positive or negative sentiments expressed on the common topics discussed by each linked and linking posts. A same resultant polarity would indicate same sentiment and conversely, different polarity indicates different sentiment. A balanced polarity means neutral or unknown sentiment value.
2.	<i>Agreement on linked blogger by linking blogger</i>	Agreement expressed on the linked blogger by linking blogger in the linking posts content.
3.	<i>Copied Sentences count</i>	
	<i>(a) Number of sentences copied from linked post to linking post</i>	The number of sentences explicitly copied from the linked post by the linking blogger on the linking post.
	<i>(b) Copied sentence ratio in linked blog post / linking blog post</i>	Number of copied sentences divided by number of sentences found in each post.
	<i>(c) Total sentences count in linked blog post and linking blog post</i>	Number of sentences found in each post.
4.	<i>Time difference between postings of linked and linking blog</i>	The difference in timing between the linked and linking blog.
Blog level data between the linked and linking blogs		
1.	<i>Similar sentiment ratio between linked and linking blog</i>	Total number of similar sentiment posts divided by the total number of posts excluding the neutral posts.
2.	<i>Agreement ratio of linking blog on linked blog</i>	Total number of agreement posts divided by the total number of posts excluding the neutral cases.
3.	<i>Average copied sentences over total number of sentences.</i>	Total copied sentences divided by the total number of sentences found between the two linked blogs.
4.	<i>Average time difference over all the cases</i>	Total time difference in postings between the two linked blogs divided by the total number of cases.
5.	<i>Number of cases where authority of linked blog is greater than linking blog</i>	Number of posts where $\text{Auth}(\text{Linked Blog}) > \text{Auth}(\text{Linking Blog})$ divided by the number of cases.
6.	<i>Community value between the two blog posts</i>	If the two blog posts are in the same community, the community value is tagged “same”, otherwise “different”.

Four blog post level features and six blog level features were analyzed and evaluated for detecting influence of the linked blog posts on the linking blog posts. The features analyzed are summarized in Table 3.1. The blog posts features are analyzed at the *post level* and *blog level* to reinforce the findings between results at both levels. The linked blog posts were analyzed individually at the post level and then aggregated to the blog level, specifically at the blog site-to-blog site pair level. At the blog level, other than the aggregated values of sentiments, agreements, copied sentences, and time differences values, specific blog level features include authority ranking and community identity. It is postulated that influence is present when two bloggers express similar sentiments on the common topics discussed. The likelihood of the linked blog post passing on its intended message across the linking posts would be higher if similar sentiments are expressed. Sentiment analysis aims to determine the attitude of a writer with respect to some topics, and whether the expressed opinion in a document or sentence is positive, negative or neutral. It has been studied by many researchers (Kwon et al., 2007; Pang and Lee, 2008; Polanyi and Zaenen, 2006; Wilson et al., 2005). From the literature review, previous studies (Leskovec et al, 2010; Cai et al., 2011; Li et al., 2011) that attempted to analyze sentiments of blog content to detect influence are relatively recent. Sentiment analysis on the whole blog content takes into consideration the overall opinion of the blog postings on the topics discussed. Following is an example of how influence can be inferred from two linked blog posts.

Linking post by RiehlWorld

Mark Levin and **Glenn Reynolds** touched on some critical points for everyone interested in politics in this video interview. **I wanted to follow them up** with a brief item on politics, the GOP and the Tea Party Movement **because the concepts are so important...**

(Source: http://www.riehlworldview.com/carnivorous_conservative/2009/05/politics-and-the-tea-party-movement.html)

Linked post by Glenn Reynolds

PJTV: I talk with Mark Levin about his new book, *Liberty and Tyranny*, about the Tea Parties, and about what people should be doing in politics today...

(Source: www.pajamasmedia.com/instapundit/77663/)

Phrases such as “*Glenn Reynolds touched on some critical points*”, “*I wanted to follow them up*”, and “*because the concepts are so important*” extracted from the blog content through reading could lead readers to infer influence of the linked blog post (Glenn Reynolds) on the linking blog post (RiehlWorld). There are also cases where blogs are linked but influence is not present. An example of non-influence between linked blogs is as follow:

Linking post by RedState

...***HuffPo*** and other places tripping all of themselves proclaiming Legum as the next big thing...

(Source : www.redstate.com/briangriffiths/2009/04/09/do-as-i-say-not-as-i-do/)

Linked post by Huffington Post

Judd Legum, who founded ThinkProgress, served as Hillary Clinton's national research director during the presidential campaign, and has been a longtime friend of AMERICAblog, is running for state delegate in the Annapolis area of Maryland. Joe and I hope you won't mind us asking if you can help Judd's campaign with a donation...

(Source:http://www.huffingtonpost.com/2009/04/06/first-major-national-oli_n_183710.html)

Non-influence could be inferred from the negative phrase “***HuffPo*** and other places tripping all of themselves” posted by the linking blog on the linked blog. This was further reinforced by taking into consideration the different sentiments expressed by both bloggers on the common topic “Legum” discussed in both postings.

Sentiment analysis is used to collect the sentiment values at blog post level data, and the aggregated sentiment values between the linked blogs are evaluated at the blog level data. The task in sentiment analysis is to classify the polarity of the posting towards common topics discussed within the two linked posts. Linked blogs may not necessarily write on similar topics, or different multiple topics could be discussed in the postings. The scope of analysis is restricted to determining the sentiments expressed on similar topics within the linked postings. Topics and the sentiments on the topics expressed by the bloggers that appeared in both postings are identified through manual analysis of the post content. A count of positive and negative sentiments on the topics is tallied separately for each blog post. If the polarity of the sentiments expressed after tally is the same, the sentiment between linked and linking variable is tagged with “same”. If the sentiments are different, the variable is tagged “different”, and if the sentiments expressed could not be determined because no common topics are found, or no opinionated words are detected, an “unknown” tag would be assigned. The similar sentiment ratio was calculated by summing up the number of blog posts with similar sentiments across all common topics for each unique blog site-to-blog site pair and dividing it by the total number of blog posts, excluding the neutral cases, between the blog site-to-blog site pair. The similar sentiment ratio measures the degree of similar sentiments expressed between the blog site-to-blog site pair with respect to the non-similar sentiments expressed.

Similarly, influence tends to be present if the linking blogger expressed agreement to the linked blogger in the posting. In the study, *positive agreement* is defined as expressions of concordance by the linking blogger with the linked blogger found in the blog posts. The whole content of the blog posts is analyzed for agreement. Possible values for the agreement parameter could be “agree” where agreement is

present, “neutral” if no agreement is detected, and “disagree” if there is disagreement expressed by the linking blogger on the linked blogger.

Other blog features analyzed include the number of sentences copied from the linked blog by the linking blog. It is postulated that a linking blogger would tend to make references by copying sentences from an influential blogger onto their own blog postings. A high occurrence of copied sentences in the linking post would have indicated an influential linked blogger. It is assumed that influential bloggers would have readers closely monitoring their posts. Thus, a shorter time difference between the posting of the linked and linking blog could be inferred as influence by the linked blogger on the linking blogger. Collection of posting time is subject to differences in time zone which are resolved by standardizing the time collected.

Authority values of the blogs are collected from the Technorati site (www.technorati.com) where authority is defined as the number of unique links to the blog site. A high authority value means a high number of other blogs are linking to the blog site. The assumption is that a high authority blog attracts more attention, and the study evaluates whether high authority necessarily do exert influence or if it is possible that readers of high authority blogs have different viewpoints and are not influenced by the high authority bloggers.

The underlying assumption for community discovery is that similar or like-minded individuals interact more frequently. In the community identity analysis, blogs in the dataset are categorized into the Democrat or Republican community with the taggings gathered from Technorati (www.technorati.com).

3.3 Dataset Used

A sample dataset of 196 blog postings pairs was collected from blog postings within the US political domain blogosphere. The political blog dataset is suitable because it is common to find communities of bloggers with their respective political inclination, in consideration for the community identification evaluation. Blog postings were collected for the period from June 2008 to June 2009 from two groups of a balanced number of bloggers based on their political inclination towards the Democrats or Republicans. The blog contents were accessed and retrieved in the month of December 2009. Bloggers with high authority values were extracted from Technorati (www.technorati.com) for each group. In Technorati, authority value is defined as the number of unique links between other blog sites and the specified site. Data collected consists of postings within a group as well as postings across the groups. Adar and Adamic (2005) showed that different community blogs link primarily within their separate communities, with far fewer cross-links exchanged between them. This can be seen in Figure 3.1 where there are more links between blogs within the same community and fewer links between blogs from different communities. For example, there are 29 postings on Instapundit (pagamasmedia.com/instapundit/) linking to RiehlWorld (www.riehlworldviews.com) in the same community, compared to 5 postings on Instapundit (pagamasmedia.com/instapundit/) linking to Huffington (www.Huffingtonpost.com/) in the other community. In Figure 3.1, the round (red) shapes represent Republican blog sites while the rectangular (blue) boxes represent Democratic blog sites. The number on an arrow refers to the number of blog postings on a source blog site which have a link (or links) to a destination blog site.

Figure 3.1 Blogs network of the collected data.

Inter-coder reliability

Manual evaluation of the blog content for detection of influence may be subjective as the results depend on the criteria used by the evaluator. In order to ensure reliability in the evaluated presence of influence, inter-coder reliability is measured to validate the influence evaluation process. It is noted that though inter-coder reliability testing checks the reliability of manually coded influence value, it does not check the reliability of sentiment analysis and agreement values. The use of inter-coder reliability testing is to test whether two coders can reliably code the influence; if agreement is high, then coding is reliable. Two coders were asked to perform the inter-coder reliability testing, where the agreement between each coder were evaluated. Instructions given to the coders are shown in Appendix F. In addition, the coders were asked to rank the blog features by their ability to indicate influence flow, and were given details on how the blog features were measured to provide a better understanding

of the scope and concept of influence flow used in the study. In the event, when the two coders disagree about an influence value, the coordinator would discuss with the coders and decide on the final influence value to provide the answer key. Table 3.2 shows the total numbers of agreements and disagreements between the two coders after the two coders independently tagged the influence value (Yes or No) of the 196 postings pairs. The calculated Cohen Kappa value between the two coders is 0.65, and it is within an acceptable range considered valid and significant (Cohen, 1960; Landis and Koch, 1977).

Table 3.2 Inter-coder agreement between two coders.

			Tester A		Total
			Influence		
			Yes	No	
Tester B	Influence	Yes	117	4	121
		No	27	48	75
		Total	144	52	196

Disagreement between the two coders usually arose due to different sentiments expressed between the blog posts and where neutrality was expressed in one of the blog posts. For example, as seen in the example shown below, the linking post expressed negative sentiments on the topic “Newt Gingrich”, while the linked post expressed neutral sentiments.

Linking post by Riehlworldview

...With no disrespect, I bet you think Newt Gingrich is a leader, one who might help us navigate difficult times. But who is leading whom here? Credit Gingrich for having the intelligence to *jump on a bandwagon*, but don't fool yourself into believing he is leading anything right now. Our current leadership has failed and continues to fail across the board. And that's why Americans are stepping up - and out...

(Source : <http://riehlworldview.com/2009/03/embracing-the-coming-polarity-shift-in-american-politics.html>)

Linked post by Instapundit

NEWT GINGRICH jumps on the Tea Party protest bandwagon.

(Source : <http://pjmedia.com/instapundit/73673/>)

One of the coders tagged the case as influence based on the reference to the linked post by the linking blogger. However, as the sentiment expressed between the two posts were different, the linking blogger was deemed to form his opinion independently from the linked blog post showing no influence between the two blog posts.

3.4 Experimental Results

The correlations between the measured blog features and influence values were evaluated to determine the effect of the features on influence between the linked and linking blog posts. Unexpected results and outliers were analyzed and possible explanations were given to rationalize the outcome of the results.

3.4.1 Blog Post-Level Data versus Influence of Linked Blog Posts

Sentiment values between the two blog posts

A contingency table analysis was conducted to evaluate the effect of sentiment value on influence. Table 3.3 shows the relationship in the sentiment values between the two blog posts versus the influence of the linked blog post. The sentiment values could either be same, neutral/unknown, or different with or without influence. A Chi-Square test confirmed that the relationship between sentiment values and influence was statistically significant ($\chi^2 = 132.59$, $df = 2$, $p < 0.001$) with a Lambda value of 0.72

and $p < 0.001$ indicating strong association between the sentiment values and influence. Hence, there is statistical evidence to relate sentiment values expressed on the blog topics with influence of the linked postings.

Table 3.3 Sentiment value versus influence cross-tabulation.

		Influence		Total
		Yes	No	
Sentiment Value	Same	94	0	94
	Neutral / Unknown	44	6	50
	Different	8	44	52
Total		146	50	196

As shown in Table 3.3, postings of the same sentiment values had high influence (94 cases). There was absence of non-influence in the linked blog post while sentiments expressed by both postings were of the same polarity. This indicated a strong relationship between influence and postings of same sentiment values, whereby if the posts expressed same sentiments, tendency was for influence to be present. Similar opinions expressed on blog topics would tend to lead towards the original blog's influence.

Neutral or unknown sentiment-value shows high frequency of linked blog postings with influence (44 cases). This could be due to blog links having information flow through the links even though no sentiments were expressed in the content. An analysis on the smaller number of blog posts with different sentiments but with influence (8 cases) revealed that though different opinions could be expressed on the topics discussed, agreement with the linked blogger could still be expressed.

Following is an example of the sentiments expressed by the linked paired blog posts on the common topic of “*Obama*”. RiehlWorld expressed negative sentiments on what “*Obama*” had done, while Instapundit was positive in sentiment about “*Obama*”.

Though both blog posts expressed different sentiments on the topic discussed, RiehlWorld was influenced by Instapundit as inferred by the phrase “*Add this below to the*” expressed before the hyperlink (“*I won*” *comment*) to Instapundit’s post.

Linking post by RiehlWorld

Add this below to the “I won” comment and I think it’s safe to say Obama is getting some serious on the job training which, hopefully will serve him well down the line. My chief concern at this point is that *a certain arrogance won't allow that*. Time will tell...

Also in his first closed door meeting which included Republicans, Obama either elevated a radio pundit / entertainer, or diminished his Presidency *to round up a rather amateurish and ego driven first few days*. This is *a man who has not yet learned the import of his words*, or come to appreciate that even a glowing media filter won't prevent them from getting out...

(Source: www.riehlworldview.com/carnivorous_conservative/2009/01/another-orookie-mistake.html)

Linked post by Instapundit

“I WON.” In Bush, this would have been troubling hubris. *In Obama, fortunately, it's just manly self-assurance...*

(Source : www.pajamasmedia.com/instapundit/67303/)

Agreement of linking blogger in linked post on linked blogger

As shown in Table 3.4, agreement of the linking post towards the linked post indicates a strong correlation with influence. The relationship between agreement values and influence was statistically significant based on the Chi-Square test conducted ($\chi^2 = 160.96$, $df = 2$, $p < 0.001$) with a Lambda value of 0.86 ($p < 0.001$) indicating very strong association between agreement and values and influence. In the collected data sample, there were no cases of non-influence if there was agreement expressed by the linking blog on the linked blog (zero case in Table 3.4). Likewise,

expressed disagreement indicated a low count of influence cases (3 cases) as bloggers in disagreement tend not to spread the influence of the other linked blogger through their writings.

Table 3.4 Agreement value versus influence cross-tabulation.

		Influence		Total
		Yes	No	
Agreement Value	Agree	46	0	46
	Neutral / Unknown	97	4	101
	Disagree	3	46	49
Total		146	50	196

Following is an example of agreement expressed by the linking blogger on the linked blogger.

Linking post by Gateway Pundit agreeing with linked post by RedState

...It looks like **RedState was right**-- This was just a typical liberal media distraction from the real news...

(Linking post source : www.gatewaypundit.firstthings.com/2009/07/figures-secret-cia-al-qaeda-plan-was-approved-by-congress/; Linked post source: www.redstate.com/erick/2009/07/12/distractions/)

A clear indication of agreement expressed by the linking blogger (Gateway Pundit) on the linked blogger (RedState) was inferred from the expression “*RedState was right*” found in the linking post. Additionally, the context of the post was verified by reading the whole content of both the linked and linking post which focused mainly on media reporting. This was an expected outcome since presence of influence usually meant there was agreement between the bloggers. Conversely, disagreement by the linking blogger towards the linked blogger could lead to non-influence of the linked posts.

Following is an example of disagreement expressed by the linking blogger on the linked blogger.

Linking post by RedState disagreeing with linked post by Huffington Post

I don't know whether Jason Linkins at the Huffington Post is a lawyer, but from this post I have to assume not - and that he really should have talked to a lawyer before publishing it...

...Let's look at the quotes Linkins chooses and why they are - assuming Linkins was writing in good faith - **so hilariously misguided**...

...But just as pundits and bloggers sometimes need to talk to military people before making mistakes about matters within the sphere of military expertise, it's a good idea to ask a lawyer before you go quoting judicial opinions. You might, like Jason Linkins, **end up making a fool of yourself**.

(Linking post source: www.redstate.com/dan_mclaughlin/2009/05/28/jason-linkins-encounters-the-hazards-of-blogging-a-subject-you-do-not-understand/; Linked post source: www.huffingtonpost.com/2009/05/28/antonin-scalia-judges-makn208531.html)

From the initial sentences in the example, disagreement might not be explicitly stated and could not be inferred. However, as the reader continues to read the blog post, the opinions expressed in the blog post slowly becomes apparent. Expressions such as “*so hilariously misguided*”, and “*end up making a fool of yourself*” would then indicate the disagreement of the linking blogger on the linked blogger.

Even if the linking blogger is neutral to the linked blogger, there was influence flow between the bloggers as shown in the below example with the neutral expression “*are for other people*” found in the linking post.

Linking post by Atrios on the linked post by Huffington Post

Permission Not Sought Or Granted

Posted by Atrios

Copyright laws **are for other people**.

-Atrios 11:57

(Linking post source : www.eschatonblog.com/2008/08/permission-not-sought-or-granted.html; Linked post source : www.huffingtonpost.com/2008/08/29/van-halen-mad-at-mccain-o_n_122464.html)

This was due to the presence of the links which created a natural outward flow of information. Further to that, in the absence of sentiment or agreement expressions, other blog features such as the same community identity could be used as possible considerations to infer influence.

The importance of considering the sentiments of the whole blog posting content and not specific to expressions or terms within the hyperlink is shown in the following example:

Credit where credit is due: **this is a nice piece of reporting on Dodd by the Huffington Post**. The title (“Dodd Dinner With Online Payday Lenders Transforms Into Fundraiser”) explains a bit, but the background is this: there’s a lobbying group called the Online Lenders Alliance. Fair enough. They’re throwing a conference this week. So far, so good. There’s a bunch of people from Congress involved or speaking, on both sides of the aisle. Fine*. Senator Dodd was one of the scheduled Senators for the event, except when asked about it first his staff, then Dodd himself flatly denied that he was there on OLA’s behalf at all; it was an independent fundraising dinner. Nothing unusual... wait, what?...

...Needless to say, the Huffington Post has a **contradictory screenshot**, presumably from the OLA website...

(Linking post source: http://www.redstate.com/moe_lane/2009/05/21/dodd-lying-to-left-over-online-lender-lobbyist-ties/; Linked post source: http://www.huffingtonpost.com/2009/05/20/dodd-dinner-with-online-_n_205629.html)

The phrase “*this is a nice piece of reporting on Dodd by the Huffington Post*” could initially be taken as agreement towards the linked blogger. However, upon further reading, it is revealed that the linking post has a negative overall sentiment about the common topics posted, and subsequent negative comments on the linked blogger were made. If sentiment is analyzed only at text surrounding the hyperlink, an erroneous positive influence conclusion would have been made. Therefore, sentiment analysis on the whole blog content is necessary to detect the correct influence indicator.

Sentences count

(a) Number of sentences copied from linked post to linking post

For both influence and non-influence cases, copying of sentences was not prevalent. Copying of sentences were found in non-influence blog posts as copied sentences could be used to reference specific disagreement over the discussed topics. However, it was observed that there was no specific trend in the number of sentences copied from linked post to linking post between influenced and non-influenced blog posts as copying of sentences was a subjective action dependent on bloggers' writing style and was often used impromptu on an ad-hoc basis. Due to the non-existent of a definite trend in the use of sentence copying, its use in the detection of influence might not be effective.

(b) Copied sentence ratio in linked blog post and linking blog post

The copied sentence ratio in linked blog post and linking blog post relates the number of copied sentences to the number of sentences in the linked post at blog level. A high ratio value in the linked blog post would have indicated a high proportionality of the linked blog content was copied across to the linking blog. That is, references to specific content found in the original linked blog could be a general trend. However, the test yielded low mean values of 0.05 and 0.14 for both copied sentence ratio in the linked and linking post respectively. It was observed that copying of sentences in linked blogs by the linking blogs is not a prevalent practice. Possible explanation is that references to the linked content could be done via the link itself and there was no need for sentences to be copied, or copying of content is limited due to copyright concerns. This was the case for both linked and linking blog posts. The study results

revealed that the correlation between sentence copying and influence was not significant for use to detect blogosphere influence.

(c) Total sentences count in linked blog post and linking blog post

As shown in Table 3.5, linked blogs posts had mean values of over twenty sentences for both influence and non-influence cases and are higher than the mean values for linking blog posts.

Table 3.5 Descriptive statistics for total sentences count

	Total Sentences in linked blog posts		Total Sentences in linking blog posts	
	Influence cases	Non-influence cases	Influence cases	Non-influence cases
Mean	27	29	10	20
Median	18	26	4	16
Mode	9	7	1	8
Max	326	129	127	69
Min	2	2	1	2

This could be because referenced posts would have mentioned most relevant content, and further elaboration of perspective or opinions would generally be shorter. There was no specific trend in the number of sentences between influential and non-influential blog posts. It was observed from the collected sample that length of blog post was subjective in nature and was dependent on factors such as writing style, interest level, the number of sub-topics discussed, etc. The total number of sentences in linking blog without influence showed a higher value than total sentences in linking blog with influence. This could be because bloggers might need to describe more to explain the rationale for not being influenced, as further details or arguments had to be provided to justify alternative viewpoints.

Time difference values between postings of linked blog and linking blog

Time difference between the posting of the linked and linking postings were considered in Figure 3.2.

Figure 3.2 Time-difference-value versus frequency of influence

It was observed that influence cases tend to have a shorter posting interval. This could be because linking bloggers were readers of influential blogs and they monitored them closely for updates. The results would have to consider the many variables, such as unpredictable reading and writing habits of bloggers, engagement in other activities, etc. that could affect the time of postings. Hence, the time of postings was subjective in nature. It was noted that there were issues on collecting time differences in blog postings due to the time zone differences. Corrective action to account for the differences in time zones was made to standardize the time comparison.

3.4.2 Blog-Level Data versus Influence of Linked Blogs

Similar sentiment ratio between linked blog and linking blog

As shown in Table 3.6, the similar sentiment ratio was computed from dividing the number of cases with the same sentiments over the number of cases that are not neutral at the blog level. A Pearson correlation test confirmed that there was a high correlation between similar sentiment ratio and influence in a positive direction ($r = 0.804$, $p < 0.001$, $N = 196$). A Pearson correlation $r \geq 0.5$ value shows correlation, while a correlation value of $r \geq 0.7$ would indicate a strong relationship between each feature and influence of the linked post (Cohen, 1988).

Table 3.6 Similar sentiment ratio versus influence cross-tabulation

		Influence		Total
		1 (Yes)	-1 (No)	
Similar Sentiment Ratio	0	7	13	20
	≤ 0.0952	6	19	25
	≤ 0.111	4	18	22
	≤ 0.941	29	0	29
	1	100	0	100
Total		146	50	196

Conversely, blogs with low similar sentiment ratio tend to inhibit influence flow. Thus the correlation between different sentiment ratio and influence propagation is inversely proportional. However, there were cases where different sentiments between blogs led to influence which reflects that though blogs could have different sentiments on the discussed topics but still emanate influence of the linked blog on the linking blog.

Agreement ratio of linking blog on linked blog

The agreement ratio was computed from dividing the number of cases with agreement over the number of cases that are not neutral at the blog level. Table 3.7 shows that blogs with a high number of agreement postings tend to propagate the information further.

Table 3.7 Agreement ratio versus influence cross-tabulation

		Influence		Total
		1 (Yes)	-1 (No)	
Agreement Ratio	0	10	13	23
	≤ 0.08	6	19	25
	≤ 0.125	4	18	22
	≤ 0.818	30	0	30
	≤ 0.909	30	0	30
	1	66	0	66
Total		146	50	196

There was high correlation between agreement postings and information propagation in a positive direction ($r = 0.765$, $p < 0.001$, $N = 196$). It could be inferred that cases with high agreement between the blogs would lead to information flow. However, there were some cases where low agreement ratio between the bloggers led to influence between the posts which showed that agreement was not a pre-requisite for information flow. This meant that influence could flow in linked blogs posts even if agreement was not expressed by the linking blogger on the linked blogger in the linking post content.

Table 3.8 shows the agreement value versus sentiment value of the records. Based on the Chi-Square test conducted ($\chi^2 = 112.872$, $df = 4$, $p < 0.001$) with a Lambda value of 0.392 ($p < 0.001$), the results show a weak association between the agreement value and sentiment value. This is because the linking bloggers did not

consistently express agreement towards the linked bloggers though there were sentiments expressed on the topics discussed in the blog posts.

Table 3.8 Agreement value versus sentiment value.

		Sentiment Value			Total
		Same	Neutral / Unknown	Different	
Agreement Value	Agree	31	13	2	46
	Neutral / Unknown	62	30	9	101
	Disagree	1	7	41	49
Total		94	50	52	196

Average copied sentence over total number of sentences

The average copied sentence was computed by dividing the total number of copied sentences over the total number of sentences between the two blogs. Due to the low instances of sentences copying compared to the total number of sentences in postings, sentences copying has minimal effect on influence detection.

Average time difference over all the cases

Average time difference at blog level was computed using total time difference over the number of cases. As shown in Figure 3.3, it was observed that influence cases tend to have shorter posting interval even at blog level. This reinforces the notion that linking bloggers tend to closely monitor linked blogs that are influential, though the highly subjective nature of posting time has to be taken into consideration.

Figure 3.3 Average time difference versus influence frequency occurrence.

Number of cases where authority of linked blog is greater than linking blog

Table 3.9 shows the effect of the blogs' authority value on influence. If authority of the linked blog is higher than the linking blog, both influence and non-influence cases would be higher. That is, authority value has a positive correlation with both influence and non-influence postings. Blogs with high authority value meant there are more blogs linking to it. Naturally, high authority blogs would have both higher influential and non-influential values because of the higher number of linkages. Since authority values cannot differentiate between the influential or non-influence cases, there was a limitation in using it as a factor to predict influence within blogs. This was because, though authority value was important within the same community, the effect of authority value was not relevant for bloggers from different communities. The limitation on the use of authority values to detect influence is reinforced by the observation that linking blogs could express differing sentiments on common topics and had no influence flow.

Table 3.9 Auth(linked blog) > Auth(linking blog) versus influence cross-tabulation

		Influence		Total
		Yes	No	
Auth(linked blog) > Auth(linking blog)	Yes	105	46	151
	No	41	4	45
Total		146	50	196

Community identity versus influence

In the evaluation of community identity versus influence, community identity and influence were significantly related based on the Chi-Square test conducted ($\chi^2 = 129.23$, $df = 1$, $p < 0.001$) with a Lambda value of 0.66 ($p < 0.001$) showing strong association between community identity and influence. As shown in Table 3.10, the number of blogs with same community identity and having no influence is zero. Hence, the results showed that blogs within the same community have a high tendency to be influenced by the linking blogs. The rationale could be that, bloggers from the same community hold a similar perspective on a particular domain. For instance, bloggers who are either Democrats or Republicans have similar political views within their respective communities. When viewpoints are similar, the tendency was for the linked blog to influence the linking blog and for information to propagate, as shown in the evaluation of similar sentiment values between two linked blogs. Thus, the use of common community identity could be used effectively to detect influence.

Table 3.10 Community value versus influence cross-tabulation

		Influence		Total
		Yes	No	
Community value	Same	129	0	129
	Different	17	50	67
Total		146	50	196

3.5 Evaluation of Product Review Domain Blogs

Additional experiments were conducted on a sample product review blog dataset and the results were compared with findings of the political blogs. The objective is to validate the findings for the political blogs with the product review blogs' results. A sample of 236 linked blog post pairs were collected from a larger dataset used in the subsequent influence diffusion detection study done in Chapter 6 with the dataset consisting of blog posts from the following product review blogs: www.9to5mac.com, www.macrumors.com, www.tuaw.com, www.engadget.com, www.techcrunch.com, www.slashgear.com, and www.theverge.com. Similarly, correlations between the measured blog features and influence values were evaluated to determine the effect of the features on influence between the linked and linking blog posts in the product review domain dataset.

3.5.1 Blog Post-Level Data versus Influence of Linked Blog Posts (Product Review Domain)

Sentiment values between two blog posts (product review domain)

As seen in Table 3.11, same sentiment value postings had a high number (94 cases) of influence cases as compared to the low number (3 cases) for non-influence cases. This indicated a strong relationship between influence and postings of same sentiment values, whereby if the linked and linking post pairs expressed same sentiments, tendency was for influence to be present. Further to that, a Chi-Square test conducted confirmed that the relationship between sentiment values and influence was statistically significant ($\chi^2 = 141.58$, $df = 2$, $p < 0.001$) with a Lambda value of 0.646 ($p < 0.001$) showing strong association between sentiment values and influence for the

product review dataset. From the result, there is statistical evidence to relate same sentiment values expressed on the blog topics with influence of the linked postings in the product review blogs, which is similar to the findings for the political blogs.

Table 3.11 Sentiment value versus influence cross-tabulation (product review domain).

		Influence		Total
		Yes	No	
Sentiment Value	Same	94	3	97
	Neutral / Unknown	87	7	94
	Different	7	38	45
Total		188	48	236

There is a higher ratio of Neutral/Unknown Sentiment Value with Influence (87 out of 188 cases) cases in the product review domain dataset as compared to the political domain dataset (44 out of 146 cases) indicating that the linking bloggers, though influenced, are more neutral in their blog postings in the product review domain dataset. This shows that product review bloggers are less forthcoming in expressing their opinions on the topics discussed in their postings as compared to the political bloggers.

Agreement of linking blogger in linked post on linked blogger (product review domain)

The records in Table 3.12 show the relationship between the agreement of the linking post towards linked post and influence exerted by the linked post. The relationship between agreement values and influence was shown to be statistically significant based on the Chi-Square test conducted ($\chi^2 = 66.37$, $df = 3$, $p < 0.001$) with a Lambda value of 0.333 ($p < 0.001$) where weak association is shown between the

two variables. The results are similar to the findings for the political blogs where agreement or neutrality expressed in the linking blog posts on the linked bloggers is correlated to influence by the linked blogs, while disagreement is correlated to non-influence. Nevertheless, the association between agreement values and influence is weaker in the product review blog posts compared to the political blog posts.

Table 3.12 Agreement value versus influence cross-tabulation (product review domain).

		Influence		Total
		Yes	No	
Agreement Value	Agree	26	2	28
	Neutral / Unknown	161	29	190
	Disagree	1	17	18
Total		188	48	236

It was observed that the ratio of agreement and disagreement posts over the total posts (0.19) were lower in the product review domain dataset as compared to the political domain dataset with a ratio value of 0.48. This indicates that political bloggers tended to express agreement or disagreement towards the linked bloggers in their posts, whereas product review blogs were less expressive towards the linked bloggers.

From Table 3.13, the results showed that agreement value was significantly related to the sentiment value between the linked and linking blog posts based on the Chi-Square test ($\chi^2 = 90.53$, $df = 4$, $p < 0.001$) with a Lambda value of 0.23 ($p < 0.001$) where the association between the agreement value and sentiment value is shown to be weak, similar to the results for the political blogs. This means that agreement value itself could not predict the sentiment value as the expression of agreement towards the linked bloggers is not a consistent phenomenon in the product review blogs even though sentiments on the topics were expressed in the blog posts. However, it was

observed that product review blog posts that expressed agreement towards the linked bloggers have same sentiment values between the linked and linking posts, which show that agreement towards the linked blogger usually means that the opinion on the topic by the linking blogger would not contradict the opinions expressed in the linked post.

Table 3.13 Agreement value versus sentiment value (product review domain).

		Sentiment Value			Total
		Same	Neutral / Unknown	Different	
Agreement Value	Agree	19	9	0	28
	Neutral / Unknown	78	85	27	190
	Disagree	0	0	18	18
Total		97	94	45	236

Sentences count

(a) Number of sentences copied from linked post to linking post (product review domain)

The copying of sentences was not prevalent in the product review dataset (14 cases for influenced posts and 3 cases for non-influenced posts out of a total of 236 blog posts), which was similarly observed in the political domain dataset. Possible explanation is that references to the linked content could be done via the link itself and there was no need for sentences to be copied, or copying of content is limited due to copyright concerns.

(b) Copied sentence ratio in linked blog post and linking blog post (product review domain)

A high copied sentence ratio value in the linked and linking blog post indicates a high proportionality of the linked blog content was copied from the linked blog and from the linking blog respectively. However, the low mean values of 0.011 and 0.11 for both copied sentence ratio in the linked and linking post in the product review domain dataset indicates few content were copied by the linking blog posts. This was similar to the political domain's mean values for copied sentence ratio in the linked and linking post (0.05 and 0.14), which shows minimal number of sentences being copied. Hence, the number of sentences copied could not be used to detect blogosphere influence for both the political as well as the product review domain datasets.

(c) Total sentences count in linked blog post and linking blog post (product review domain)

Table 3.14 shows the descriptive statistics for the total sentences count of the product review domain dataset.

Table 3.14 Descriptive statistics for total sentences count (product review domain)

	Total Sentences in linked blog posts		Total Sentences in linking blog posts	
	Influence cases	Non-influence cases	Influence cases	Non-influence cases
Mean	16	31	19	17
Median	13	24	13	11
Mode	9	9	7	8
Max	98	82	92	97
Min	4	3	5	3

It was observed from the product review domain dataset that length of blog post was subjective in nature and was dependent on factors such as writing style, interest level, the number of sub-topics discussed, etc. Hence, there was no specific trend in the number of sentences between influential and non-influential blog posts for the product review domain dataset results.

Time difference values between postings of linked blog and linking blog (product review domain)

The time difference between the posting of the linked and linking postings values for the product review domain dataset are shown in Figure 3.4.

Figure 3.4 Time-difference-value versus frequency of influence (product review domain)

It was observed that influence cases tended to have longer posting interval as compared to the shorter posting interval for the political domain dataset. While it is possible that the political domain bloggers monitor the influential linked blogs closely for updates and re-post in a shorter time interval, it cannot be inferred that short

posting interval is positively correlated to influence as seen in the product review domain results where influence cases could have longer posting interval. Furthermore, the results would have to consider the many variables, such as unpredictable reading and writing habits of bloggers, engagement in other activities, etc. that could affect the time of postings. Therefore, posting time interval is not a reliable indicator of influence.

3.5.2 Blog-Level Data versus Influence of Linked Blogs (Product Review Domain)

Similar sentiment ratio between linked blog and linking blog (product review domain)

From Table 3.15, the similar sentiment ratio was computed through dividing the number of cases with same sentiments over the number of cases that are not neutral at the blog level.

Table 3.15 Similar sentiment ratio versus influence cross-tabulation (product review domain)

		Influence		Total
		1 (Yes)	-1 (No)	
Similar Sentiment Ratio	0	0	9	9
	≤ 0.33	8	13	21
	≤ 0.75	23	8	31
	≤ 0.928	79	15	94
	≤ 1	78	3	81
Total		188	48	236

Pearson correlation test confirmed that there was a high correlation between similar sentiment ratio and influence in a positive direction ($r = 0.587$, $p < 0.001$, $N = 236$), where a Pearson correlation $r \geq 0.5$ value shows correlation, and a correlation value

of $r \geq 0.7$ would indicate a strong relationship (Cohen, 1988). The product review dataset Pearson r value ($r = 0.587$) was smaller than the political dataset value ($r = 0.804$). While the correlation between similar sentiment ratio and influence for the product review dataset may not be as high as the value for the political dataset, influence could be inferred when the product review blog posts have similar sentiments expressed on the common topics as seen in the following example where both linking and linked blog posts expressed negative sentiments on the app:

Linking blog post

Apple has **stricken** 500px's apps from the App Store...

Apple, however, had some **harsher** things to say about the app...

Linking post source : <http://www.slashgear.com/apple-pulls-500px-apps-from-app-store-due-to-nude-photo-search-ability-22266320/>

Linked blog post

The Apple reviewer told the company that the update **couldn't be approved**...

The app was **removed** from the App Store...we also received **customer complaints**...

Linked post source : <http://techcrunch.com/2013/01/22/apple-pulls-500pxs-mobile-apps-from-the-app-store-claiming-its-too-easy-to-search-for-nude-photos/>

Conversely, product review blogs with low similar sentiment ratio tend to inhibit influence flow since their sentiments on the common topics differs. Thus the correlation between different sentiment ratio and influence propagation is inversely proportional.

Agreement ratio of linking blog on linked blog (product review domain)

The agreement ratio was computed from dividing the number of cases with agreement over the number of cases that are not neutral at the blog. A high agreement

ratio value indicates that the linking blog expresses more agreement than disagreement towards the linked blog in the blog postings. On the other hand, a low agreement ratio value shows more disagreements were expressed. Table 3.16 shows the agreement ratio values versus influence cross-tabulation for the product review domain dataset.

Table 3.16 Agreement ratio versus influence cross-tabulation

		Influence		Total
		1 (Yes)	-1 (No)	
Agreement Ratio	0	71	24	95
	≤ 0.188	8	13	21
	≤ 0.5	2	2	4
	≤ 0.667	17	9	26
	1	90	0	90
Total		188	48	236

It was observed that neutral blogs with no agreement expressions could also indicate influence for a product review domain dataset and the Pearson correlation test ($r = 0.309$, $p < 0.001$, $N = 236$) shows no significant relationship between agreement expressed by linking blogs and influence by linked blogs at blog level since $r < 0.4$ (Cohen, 1988).

The product review blogs expressed fewer agreement statements as compared to political blogs, since product review blogs are usually more neutral towards the linked bloggers where the content would usually focus on the product discussion rather than the linked blogger. On the other hand, political blogs tend to be expressive in stating their opinions and support for the linked bloggers. In contrast to political blogs dataset, expressed agreement towards linked bloggers is not indicative of influence at blog level for the product review domain dataset as neutral linking blogs are also significantly influenced. Though agreement/neutral and disagreement expressions towards linked bloggers were clear indicators of influence and non-influence

respectively at blog post level, the blog level relationship between agreement ratio values and influence relationship shows otherwise. At the blog level, agreement ratio measures only the degree of agreement in relation to influence and the results had shown that product review blogs which are influenced could also be neutral towards the linked bloggers. It is noted that though there is no significant relationship between agreement ratio values and influence, agreement towards the linked blogger was still considered together with sentiment values between linking and linked bloggers, and bloggers' influence styles as features in the Influence Diffusion Detection Model, since there is a clear indication of influence when a linking blogger expresses agreement towards the linked blogger.

Average copied sentence over total number of sentences (product review domain)

Similar to the findings for the political domain dataset, results for the product review domain dataset show that copying from the linked blogs' content had minimal effect on influence detection due to the low instances of sentences copying compared to the total number of sentences in the postings.

Average time difference over all the cases (product review domain)

Average time difference at blog level was computed using total time difference over the number of cases. As shown in Figure 3.5, it was observed in the product review domain dataset that influence cases tend to have longer posting interval at blog level in contrast to the results for the political blogs given in Figure 3.3. This shows that while influenced political blogs tend to post in shorter intervals, influenced product review blogs post after a longer time period. In addition to the consideration

that posting time is highly subjective in nature, using posting time interval as a feature to detect influence diffusion would not be reliable.

Figure 3.5 Average time difference versus influence frequency occurrence.

Blog authority and community identity (product review domain)

Blogs with high authority value meant there are more blogs linking to it. However, authority values cannot differentiate between the influential or non-influence cases, and this was observed in the product review domain dataset with linking blogs expressing differing sentiments on common topics and had no influence flow, which means authority value alone could not indicate influence cases.

The community identity versus influence evaluation was not performed for the product review domain dataset as there were no clear indications of community identities for some of the blogs. For example, the blog www.engadget.com writes on both Android and Apple topics, and does not identify itself with either community. Hence, the evaluation of the community identity for the product review dataset was not possible.

3.6 Discussion

It is observed from the results that sentiment expressed by both linked bloggers on common topics, and agreement towards the linked blogger by the linking blogger give high correlation values. This indicates that the results from content analysis are closely related to influence within the blogosphere and would be the most effective approaches to use for detecting influence. Similar sentiments and agreement between two bloggers indicates a higher propensity to further propagate information from the linked blog post via the linking blog post to other blogs with extended links. Hence, the effectiveness of sentiment analysis and agreement detection on the blog content is significant for detecting influence in the blogosphere because the approaches provide a more comprehensive influence analysis through considering the opinions expressed in the blog posts.

In the evaluation of blog post level features, copied sentences and length of blog are found to be non-correlated with influence. This is contrary to previous studies (Agarwal and Liu, 2009; Matsumura et al., 2008; Adar and Adamic, 2005) that had used text similarity between linked blogs to measure influence. Similar text appearing in both the linked and linking blog posts may not indicate influence as linking bloggers often used copied text as references to the linked blog post to further illustrate their own viewpoint. This could happen for both influence and non-influence cases as in linking posts with opposing views could also reference to the linked posts using copied text. Moreover, having positive terms being propagated to the linking posts may not mean there is influence as often the overall opinion of the posting has to be considered to detect influence accurately. Blog length is subjective in nature and is dependent on factors such as writing style, interest level, the number of sub-topics discussed, etc. However, it is observed that linking blogs without influence tend to be longer than

linking blogs with influence as a longer explanation is needed to explain the rationale for non-influence, as further details or arguments are provided to justify alternative viewpoints. From the result, though time difference for linking bloggers to post is shorter if the linked blogger is influential, the subjectivity of time difference in posting has to be taken in consideration. Different bloggers have different routines with regards to posting time. Even individually, bloggers' posting behavior could also vary.

At the blog level, results for sentiment values, agreement, copied sentences, blog length, and time difference are similar and coherent with the corresponding features at blog post level. This shows that these features have the same relationship with influence at both the post and blog level. Therefore, blog level information could be used to reinforce the conclusions made at the post level. In the absence of specific post level information, blog level information is useful in providing further indications of influence in the blogosphere.

The other blog level features, community identity and blog authority exhibited different behavior with respect to their relation with blogosphere influence. Community identity was shown to relate closely with influence in the blogosphere. As observed in the sample blogs collected, blogs within a common political standing, either Democrats or Republican, tend to have common viewpoints and opinions on posted topics. Though authority of the blog within a community indicates some form of influence as popular blogs tend to be closely followed, popularity itself is not a definite indication of influence as influential blogs may not be popular in nature. Moreover, authority values do not apply across other communities as bloggers tend not to influence other bloggers from different communities due to contradictory viewpoints. This shows that the number of links to a blog is not effective in detecting influence.

It is observed that non-influence cases are fewer than influence cases as bloggers may tend to avoid conflicts by not posting contradictory content in their linking postings. This is supported by comments made by bloggers such as the following statement from TalkingPoint.

...As a rule, I try to avoid link war nonsense and getting into spats with other sites. But I'll make a small exception because it illustrates a point...

(Source: http://www.talkingpointsmemo.com/archives/2009/02/key_metric_the_fr_ratio.php)

The dataset from the study in this chapter is extracted from the political domain as the initial assumption is that political blogs contain more opinionated terms, which are easier to analyze for sentiments in this exploratory study. An additional sample dataset was collected from blogs in the product review domain and analyzed as a comparison to the results gathered from the political blogs. It was shown that there is a strong relationship between influence and postings of same sentiment values in the product review blogs similar to the findings for the political blogs. Nevertheless, the political blogs expressed more sentiments in their blog posts as compared to the product review blogs which were more neutral in their opinions.

Agreement and neutrality expressed towards the linked blogger in the blog posts was also shown to relate strongly with influence for the product review blog posts, but with the product review blog posts expressing fewer agreement or disagreement statements towards the linked bloggers than the political blog posts. This could be because political bloggers are usually more explicit in expressing support and agreement for their political friends as compared to product review bloggers. There was no significant correlation between agreement ratio and influence at the blog level in the product review blogs as agreement ratio was defined to measure the degree of agreement excluding the effects of neutral statements in detecting influence. The

results show that blogs which were neutral towards the linked bloggers could also infer influence. Therefore, while the presence of agreement towards the linked bloggers in the blog post is a clear indication of influence, using it as the only feature in detecting influence would not give good results.

Copying of sentences was not prevalent in the product review domain dataset and the results showed that sentence copying and total sentence length were not good indicators of influence diffusion in both the product review and political domain datasets. It was observed that influence cases tended to have longer posting interval in the product review domain dataset, while the political domain dataset had a shorter posting interval. From the results, it cannot be inferred that short posting interval is positively related to influence as seen in the product review domain dataset where influence cases have longer posting intervals. It was also observed in the product review domain dataset that linking blogs could express differing sentiments on common topics and had no influence flow indicating that authority value alone could not determine influence cases, which is congruent with the findings for the political domain dataset. The community identity versus influence evaluation was not performed for the product review domain dataset as there was no clear indication of community identities for some of the blogs. Future work could involve collecting product review domain blogs from specific communities to evaluate the effects of community identities in the product review blogosphere.

3.7 Summary

The work in this chapter studied and analyzed the features of the blogosphere, together with sentiment and agreement analysis, and community identity to determine

the correlation between the features and influence within linked blogs to meet the requirements of research objective (RO1: *Identify the relevant blog features that could be used to detect influence diffusion between linked blog posts*). Posting level features as in length of both linked and linking blogs, number of copied sentences from the linked post to the linking post, and time difference between the linked posts together with results from the blog content and community analysis were evaluated through statistical means based on research method (RM1: *Perform statistical analysis to determine blog features suitable for detecting influence diffusion*). At the blog level, evaluations common to both posting and blog level were found to be coherent. Specific blog level parameters as in community identity exhibited strong correlation with influence, whereas authority of the blogs do not have a direct relation as popularity of the blog is not a clear indication of influence. Posting level features as in number of copied sentences, length of postings, time differences between linked blogs did not exhibit clear indication of influence when analyzed. Features like length of postings and time difference in posting are subjective in determining influence. These features depend on each different blogger's style and the topics discussed in the posting.

Content analysis, using similar sentiment on common topic and agreement of the linking blogger on the linked blogger shows high correlation with influence compared with other blog features. In addition, sentiment and agreement analysis on the whole blog content provide more comprehensive evaluation of influence as the opinion of the posting is taken into consideration. Though community identity was found to be useful for influence detection, it will not be considered in the proposed influence diffusion detection model as the aim of this study is focused on the use of influence styles for influence diffusion detection.

Based on the finding from this study that similar sentiments expressed on common topics between linked blog content are clear indicators of influence flow, the next study in Chapter 4 further explores and establishes a sentiment analysis approach in relation to the second research objective (RO2: *Establish a sentiment analysis approach that takes into consideration the complex relationships between words in blog posts to automatically detect influence diffusion between linking and linked blog posts*).

CHAPTER 4

SENTIMENT ANALYSIS OF BLOG POSTINGS

4.1 Introduction

The study in Chapter 4 applies research method (RM2: *Apply sentiment analysis based on a typed dependencies rules and a subjective phrase lexicon approach*) through the use of class sequential rules (CSR) (Liu, 2006) to accomplish research objective (RO2: *Establish a sentiment analysis approach that takes into consideration the complex relationships between words in blog posts to automatically detect influence diffusion between linking and linked blog posts*). Additional pattern rules were introduced in the process to handle the complex relationships between words, and experimental results show this leads to improvement in sentiment classification performance. Two sets of tests were performed in this study. First, the performance evaluation between the CSR-generated rules and manual rules was carried out, followed by the evaluation of the proposed Sentiment Analysis model based on further complex relationship rules.

The next section provides the background for this study, with Section 4.3 and Section 4.4 describing the CSR and Heuristic Rules methods used in deriving the typed dependencies rules. Section 4.5 provides details of the typed dependencies rules evaluation. The semantic rules used for handling the complex relationships between words are explained in Section 4.6, while Section 4.7 gives an overview of the proposed Sentiment Analysis model as well as the details of the model evaluation process. Finally, Section 4.8 summarizes the studies conducted in Chapter 4.

4.2 Background

Sentiment analysis has been studied at both the document and sentence level with the goal of assigning an overall sentiment polarity for the document or sentence (Riloff and Wiebe, 2003; Turney, 2002). Other studies proposed methods to predict sentiment polarity at the clause or phrase levels to provide a more refined analysis (Wilson et al., 2005; Wilson et al., 2006). Recent sentiment analysis studies (Hassan et al., 2010; Thet et al., 2010; Wilson et al., 2009) focused on the functional relations of words using typed dependency parsing, which provides a refined analysis of the grammar and semantics of textual data. The focus of this study is on sentiment analysis at the phrase level, which is specifically defined as the bigrams in typed dependencies.

AMOD(calamity:NN, great:JJ) for “great calamity” ADVMOD(destroy:JJ, completely:RB) for “completely destroy” DOBJ(killed:VB, innocents:NN) for “killed innocents”

Figure 4.1 Typed dependency bigrams.

Figure 4.1 shows examples of the adjectival modifier (AMOD), adverbial modifier (ADVMOD), and direct object modifier (DOBJ) typed dependencies for the respective phrases with the notations represented as NN:noun; JJ:adjective; RB:adverb; and VB:verb. Previous linguistic approaches in sentiment analysis leveraged semantic dependencies between words to predict sentiments. Wilson et al. (2006) observed that word patterns along with its modifiers could determine the intensity of phrase sentiments. For example, in the phrase “*indiscriminate* killing”, the adjectival modifier “*indiscriminate*” gives the phrase an overall negative sentiment. Hence, the discovery of word patterns within typed dependencies could reveal clues to identifying the phrase polarity.

The fundamental notion of typed dependency is based on the idea that the syntactic structure of a sentence consists of binary asymmetrical relations between the words (Nivre, 2005). Intuitively, by using typed dependencies, the syntactic structure of the sentence will be taken into consideration during the sentiment analysis. A natural language parser is a program that works out the grammatical structure of sentences, for example, the grouping of words as phrases, and the identification of words that are the subject or object of a verb. In this study, the Stanford Parser³ is used to generate typed dependencies from the sentence dataset. Figure 4.2 shows a typed dependency tree generated from the Stanford Parser for the sentence “*The cruel dictator willfully suppressed the citizens.*” The corresponding typed dependency bigrams are listed in Figure 4.3.

Figure 4.2 Typed dependency tree

In this example, the parser generates the typed dependencies output showing the semantic relationship between the bigrams. Typed dependencies facilitate the analysis of semantic relationships between words based on both their grammatical relationships and overall sentence syntactical structure. This approach allows words positioned far apart to be analyzed without neglecting the semantic and syntactic significance that could impact sentiment prediction performance.

³ <http://nlp.stanford.edu/software/lex-parser.shtml>

DET(dictator-3, The-1) AMOD(dictator-3, cruel-2) NSUBJ(suppress-5, dictator-3) ADVMOD(suppress-5, willfully-4) DET(citizens-7, the-6) DOBJ(suppress-5, citizens-7)

Figure 4.3 Generated typed dependencies.

From Figure 4.3, it can be seen that by analyzing the polarity patterns of the bigram words, the polarity of the typed dependencies can be inferred. For example, “AMOD(dictator-3, cruel-2)” could indicate a polarity pattern rule of “AMOD(NN(-),JJ(-))” giving a negative output, where (NN(-)) and (JJ(-)) refer to negative oriented noun term and adjective term respectively. In this study, only the subjectivity of the AMOD, ADVMOD, and DOBJ typed dependencies were evaluated as they are deemed to contain the vast majority of opinionated expressions (Jakob et al., 2009; Wilson et al., 2006). In the Stanford Parser, an adjectival modifier (AMOD) of a noun phrase is any adjectival phrase that serves to modify the meaning of the noun phrase, an adverbial modifier (ADVMOD) of a word is an adverb or adverbial phrase that serves to modify the meaning of the word, and the direct object (DOBJ) of a verb phrase is the object of the verb.

Heuristics are typically used (Shaikh et al., 2008; Thet et al., 2010) to determine typed dependency polarity patterns, which may not comprehensively identify all possible rules. The study in this chapter aims to achieve the research objective of establishing an automatic sentiment detection approach in consideration of the blog posts’ opinions to identify influence diffusion between linked bloggers. This is achieved through applying the Class Sequential Rules (CSR) (Liu, 2006) to derive typed dependencies rules from sequential patterns within the bigrams and taking into account the complex relationships between words. The CSR approach was adopted as

the algorithm could automatically derive rules based on the sequential ordering of the bigram terms. Both linguistic and machine learning methods are combined to automatically detect the sentiment polarity of bigrams in specific typed dependencies. The data was parsed to obtain the typed dependency bigrams of the sentences from the dataset and subsequently the CSR algorithm was applied to derive the polarity rules from sequential patterns within the bigrams. This is followed by the evaluation of the CSR rules compared to the heuristic rules adapted from Thet et al. (2010). An overview of the CSR and HR comparison process is shown in Figure 4.4.

Figure 4.4 Overview of the CSR and HR approach.

4.3 Heuristic Method

A common approach to word pattern discovery is through heuristics, that is, knowledge engineering or manual discovery. Shaikh et al. (2008) considered the semantic relationship between textual components in a sentence and the computation of contextual valence of the words to create word pattern rules. Thet et al. (2010) manually created rules based on grammatical dependencies and prior sentiment scores of the feature terms to compute the sentiment of a clause. This study adapted the

heuristic rules implementation by Thet et al. (2010) as the baseline with the heuristic rules (HR) used listed in Table 4.1.

Table 4.1 Heuristic pattern rules (HR) (Thet et al., 2010)

HR Type	[+] class	[-] class	[0] class
AMOD (governor, dependent)	NN+, JJ+ NN, JJ+ NN+, JJ	NN-, JJ+ NN-, JJ NN+, JJ- NN, JJ- NN-, JJ-	NN, JJ
ADVMOD (governor, dependent)	VB+, RB+ VB, RB+ VB+, RB	VB-, RB+ VB-, RB VB+, RB- VB, RB- VB-, RB-	VB, RB
DOBJ (governor, dependent)	VB+, Obj+ VB+, Obj VB, Obj+	VB+, Obj- VB, Obj- VB-, Obj+ VB-, Obj VB-, Obj-	VB, Obj

An example of a type dependency rule is AMOD(phone:NN[0], great:JJ[+])→[+], where the pattern NN[0], JJ[+] gives a positive polarity output. The neutral class rules, which were assumed to be positive in Thet et al. (2010) were further separated to their own class in this study to make a direct comparison with CSR's positive [+], negative [-], and neutral [0] classes. The typed dependencies rules used in Thet et al. (2010) were derived based on a prior knowledge base. In contrast to a manual heuristic approach, the proposed CSR approach in this study could automatically generate a comprehensive set of rules to provide the sentiment classification. However, performance of automatically generated rules may not be comparable to the manually derived rules. Hence, this study conducts experiments to compare the performance between the manual heuristic approach and the proposed CSR approach in predicting the sentiment polarity output of phrases.

4.4 Class Sequential Rules Method

Liu (2006) introduced Class Sequential Rules (CSR), which are similar to Association Rule Mining (Agrawal and Srikant, 1994) by using the concept of a frequent item set, except that CSR considers the subsequence (called sequential pattern) and class labels of individual records. The CSR method was used to identify the pattern rules for each bigram's polarity class with this study's implementation given as follows:

- Let S be the set of relation data sequences where each sequence is also labelled with a class y . Let I be the set of all items in S , and $Y=\{[+], [-], [0]\}$ be the set of all class labels, and $I \cap Y = \emptyset$.
- The input data D is denoted as $\{(s_1, y_1), (s_2, y_2), \dots, (s_n, y_n)\}$, where s_i is a sequence in S and $y_i \in Y$ is its class. The sequence s_i is represented as a relational triplet for the item records: [item1(POS, polarity), item2(POS, polarity), item3(class)] and [item1(polarity), item2(polarity), item3(class)] for the two respective evaluations. POS denotes the part-of-speech tag.

A **class sequential rule** is of the form $X \rightarrow y$, where X is a sequence, and $y \in Y$. A data instance (s_i, y_i) is said to **cover** a CSR, $X \rightarrow y$, if X is a subsequence of s_i . A data instance (s_i, y_i) is said to **satisfy** a CSR if X is a subsequence of s_i and $y_i = y$.

From the CSR algorithm, both the unigram (e.g., “NN[+], _”) and bigram (e.g., “NN[+], JJ[+]”) pattern rules were extracted based on an experimentally determined minimum support threshold of 0.01 as suggested by Hu and Liu (2004). Additionally, a minimum confidence threshold of 0.5 was used to remove noisy data. The generated

rules are sorted in order of descending F1 measure of their support and confidence values, defined as follows.

$$F1 = \frac{2 * Support * Confidence}{Support + Confidence} \quad (Eq. 4.1)$$

The confidence threshold values were validated by standardizing the support threshold value to 0.01 and varying the minimum confidence threshold value from 0.1 to 0.7. It was observed that beyond a confidence threshold value of 0.7, there were no valid polarity pattern rules generated for certain test records. Ten-fold cross validation for each of the positive [+], negative [-], and neutral [0] class was conducted in the evaluation, where 10% of the dataset was tested upon with the remaining 90% dataset used for training in a recursive manner until all portions of the dataset were tested. The apportioned training data was used to generate the polarity pattern rules and in turn tested on the other portion of test records. Results from the 10 test sets were then averaged to give the overall result. The average F1 scores of the respective confidence threshold values with support threshold value = 0.01, and the average, minimum, and maximum number of pattern rules generated for the ten-fold cross validation tests are shown in Table 4.2.

Table 4.2 Results for respective confidence threshold values

Confidence Value	AvgF1 (%)	AvgNo Pattern	MinNo Pattern	MaxNo Pattern
0.1	84.45	38.6	36	41
0.2	84.71	29.4	27	32
0.3	84.39	27.4	25	29
0.4	84.39	26.3	24	28
0.5	84.39	24.4	22	26
0.6	84.39	23.8	22	26
0.7	84.39	22.3	20	24

From the results of Table 4.2, there is no significant difference in average F1 scores among the tested *Phrase-Level Sentiment Classification* confidence values at the $p < .05$ level for the conditions $[F(6,63)=.007, p=1.000]$ based on one-way ANOVA testing. The results further showed that the number of generated patterns is greatest when the confidence value is 0.1. The support threshold values were further validated by standardizing the confidence threshold value using the arbitrarily chosen value of 0.5 and varying the minimum support threshold value.

Table 4.3 Results for respective support threshold values.

Support Value	AvgF1 (%)	AvgNo Pattern	MinNo Pattern	MaxNo Pattern
0.01	84.97	24.4	22	26
0.02	84.50	16.4	15	17
0.03	83.93	13.5	13	14
0.04	83.35	12.5	11	13
0.05	83.04	10.4	10	11
0.06	83.03	10	10	10
0.07	83.04	10	10	10
0.08	83.04	9.8	9	10
0.09	83.04	9.5	9	10
0.1	83.60	9	9	9

Table 4.3 shows the average F1 scores of the respective support threshold values and the average, minimum, and maximum number of generated pattern rules for the ten-fold cross validation tests. The results of Table 4.3 show no significant difference in average F1 scores among the tested support values at the $p < .05$ level for the conditions $[F(9,90)=.183, p=.995]$ among the respective support threshold values with confidence threshold value = 0.5 based on one-way ANOVA testing. The results also show that the number of generated patterns is greatest when the support value is 0.01. From the validation tests, the support threshold value of 0.01 was used, which provides the highest average F1 score, and an arbitrarily chosen confidence threshold value of 0.5 for subsequent experimental testing.

4.5 Typed Dependencies Rules Evaluation

4.5.1 Dataset Used

The subjectivity dataset from Pang and Lee (2004) contains subjective and objective sentences extracted from a movie review dataset and was used to obtain 1000 records each for the adjectival modifier (AMOD), adverbial modifier (ADVMOD), and direct object modifier (DOBJ) typed dependencies for use in the typed dependencies rules evaluation. However, the subjective and objective labels given in the dataset were not used, as the focus of this study was on phrase level prediction and a separate manually tagged phrase level labels were used instead. The typed dependencies were generated from sentences in the dataset using the Stanford Parser and the polarity pattern rules were derived through the semi-supervised CSR algorithm. Next, the polarity of each word in the typed dependencies bigrams was tagged by matching the word to the subjectivity lexicon terms from Thet et al. (2010), which were derived from SentiWordNet⁴ (Esuli et al., 2006) and the subjectivity lexicon of Wilson et al. (2005). Each word in the typed dependencies bigrams is tagged with one of three polarity values of positive (+), negative (-), or neutral (0).

Table 4.4 Polarity distribution of the subjectivity lexicon terms.

POS	Positive (+)	Negative (-)	Neutral (0)
Adjectives	5375	8392	1027
Adverbs	2010	934	305
Verbs	1149	2303	714
Nouns	5407	8591	1409

Table 4.4 shows the distribution of the subjectivity lexicon terms with respect to their polarity and part-of-speech (POS), where it is observed that there are generally more

---

⁴ <http://sentiwordnet.isti.cnr.it>

negative terms than positive terms, except for the adverb terms. The typed dependencies, along with the polarity tagged bigram words, were then manually annotated to give the overall typed dependency polarities. Manual annotation of the typed dependency bigrams' polarity between two annotators gave a Cohen Kappa (Cohen, 1960) value of 0.78, indicating an acceptable reliability of the coded polarities. The respective typed dependency bigrams' polarity distributions in the coded dataset are shown in Table 4.5. It is observed for the dataset that more opinions are expressed using adjectival words as seen from adjective modifiers (AMOD) having the most sentiment phrases, while direct objects (DOBJ) has the least number. Figure 4.5 shows samples of the coded data records which indicate the relation “typed-dependency(governor-term:POS[polarity], dependent-term:POS[polarity])→ [bigram-polarity]”. Acronyms used here are POS:part-of-speech, NN:noun, JJ:adjective, RB:adverb, and VB:verb.

Table 4.5 Polarity distribution of bigrams in coded dataset.

Type	Positive (+) Class	Negative (-) Class	Neutral (0) Class	Total
AMOD	320	228	452	1000
ADVMOD	245	222	533	1000
DOBJ	196	153	641	1000

AMOD(movie:NN, nice:JJ+)→[+] ADVMOD(nice:JJ+, very:RB+)→[+] DOBJ(upset:VB-, viewers:NN+)→[-]

Figure 4.5 Sample generated output of coded data.

Wilson et al. (2006) found that the sparsity of word occurrences in clauses posed a problem in extracting meaningful patterns. Joshi and Penstein-Rosé (2009) observed that generalizing the typed dependency words into their part-of-speech improves

performance. In this study, the typed dependency bigram features are generalized to their respective part-of-speech and polarity (e.g., JJ+). This reduces the number of distinct features, which increases the statistical significance of the patterns. The effects of generalization were further evaluated by reducing the features to just the polarity of each word in the bigram. The results show that the performance of using only the polarity of terms is comparable to that of using both part-of-speech and bigram polarity.

AMOD(NN[0],JJ[+])→[+] ADVMOD(JJ[+], RB[+])→[+] DOBJ(VB[-], NN[+])→[-]

Figure 4.6 Sample generated output for POS-polarity evaluation.

Figure 4.6 shows samples of the generated output of the POS-polarity records, in the form of “typed-dependency(governor-POS[polarity],dependent-POS[polarity])→[bigram-polarity]”. Figure 4.7 shows samples of the generated output for the polarity records, indicating the relation “typed-dependency(governor[polarity], dependent[polarity])→[bigram-polarity]”.

AMOD([0],[+])→[+] ADVMOD([+],[+])→[+] DOBJ([-],[+])→[-]

Figure 4.7 Sample generated output for polarity evaluation.

At the phrase level, i.e., bigram words, and given a comprehensive subjectivity lexicon, the generated polarity pattern rules can be applied regardless of domain. For example, AMOD(movie:NN[0], bad:JJ[-])→[-] or AMOD(phone:NN[0], lousy:JJ[-])→[-] from a movie review and a product review blog respectively will give similar negative polarity output in the pattern rule evaluation. In addition, by ignoring the bigram terms,

which can be domain dependent, and the part-of-speech and considering only its polarity as a feature in the polarity evaluation, the model is further generalized across domains.

4.5.2 Evaluation Results

The typed dependencies rules evaluation compares the performance of the CSR generated rules and the manual rules adapted from Thet et al. (2010). The following sections provide the results of the three major typed dependencies for CSR and HR.

AMOD Evaluation

The list of generated AMOD CSR rules is shown in Table 4.6, with the top 10 generated AMOD CSR rules sorted with respect to their F1 scores listed in Table 4.7.

Table 4.6 CSR rules for AMOD typed dependency.

[+] class patterns	[-] class patterns	[0] class patterns
● NN+,_ (beauty) ● JJ+,_ (interesting)	● NN-,_ (problem)	● VB,_ (known) ● CC,_ (or)
● _, JJ+ (great) ● _,VB+ (winning)	● _,JJ- (forgettable) ● _,VB- (deteriorating)	● _,JJ (recent) ● _,VB describe) ● _,NN (story) ● _,DT (the)
● NN, JJ+ (movies, greatest) ● NN+,JJ+ (insight, great) ● NN+, JJ (accomplished, most) ● NN,VB+ (vision, expanding)	● NN, JJ- (concept, stale) ● NN-, JJ- (struggle, desperate) ● NN-, JJ (poignancy, certain) ● NN-, JJ+ (problem, biggest) ● NN+, JJ- (tribute, hollow) ● NN, VB- (flaws, infuriating)	● NN, JJ (years, recent) ● NN, VB (member, surviving) ● NN, DT (ending, the)
*Note: The rules in the table only consider the part-of-speech and polarity of the bigram words. The words in parenthesis are shown as examples only.		

The rules are in the order of “governor-POS[polarity], dependent-POS[polarity]”, with the unigram pattern rules represented as “governor-POS, _” and “_, dependent-POS”. In comparison with the HR patterns in Table 4.1, CSR discovered additional patterns containing verbs like “*expand*” (VB+), “*baffling*” (VB-), and “*known*” (VB), which are actually adjectival verbs, explaining their presence in AMOD. The discovery of additional pattern rules shows CSR’s capability to provide greater in-depth analysis of typed-dependency bigram patterns. However, as with any rule-based approach, the ordering of rules must be deliberated carefully. Simply ordering them based on F1 values poses a problem in CSR because a generic unigram pattern rule of higher priority may inadvertently override a more specific bigram rule of lower priority from a different class. To overcome this, the specific bigram pattern rules take precedence over generic unigram rules. For example, the more specific bigram pattern rule AMOD(NN+,JJ-) $\rightarrow$ (-) with F1=0.53 overrides the unigram pattern rule AMOD(NN+,_) $\rightarrow$ [+] with a higher F1=0.61.

Table 4.7 Top 10 AMOD CSR rules.

#	F1	Governor	Dependent	Class
1	1.00	NN (member)	VB(surviving)	0
2	1.00	VB (known)	_	0
3	0.99	NN-(poignancy)	JJ- (certain)	-
4	0.99	NN(vision)	VB+(expanding)	+
5	0.99	NN(subplot)	VB-(baffling)	-
6	0.98	_	VB+(rising)	+
7	0.95	NN+ (insight)	JJ+ (great)	+
8	0.85	NN (concept)	JJ- (stale)	-
9	0.81	NN- (problem)	_	-
10	0.80	NN (movie)	JJ+ (great)	+

As seen in Table 4.8, the F1 values for the four methods, POS-polarity and polarity features for both CSR and HR approach do not deviate much. However, CSR found additional rules such as AMOD(NN,VB+) $\rightarrow$ [+], AMOD(NN,VB-) $\rightarrow$ [-] and others for

the neutral [0] class, which were absent in HR. As a result, CSR enjoys up to 4% better recall (96.84 versus 92.78) for the polarized classes, and 6% better precision for the neutral [0] class (95.83 versus 89.18) compared to HR. From Table 4.8, although the increase in average F1 for the polarity features method is only 0.4% (85.37 versus 84.97) for CSR and close to 2% (85.87 versus 83.95) for HR, it shows that part-of-speech is not essential in the bigram pattern rules. In fact, part-of-speech tagging introduces errors, lowering the recall (up to 6% lower for [+] and [-] classes) in HR.

Table 4.8 Results for AMOD POS-polarity and polarity rules

Method	Class	Precision (%)	Recall (%)	F1 (%)	AvgF1 (%)
CSR (POS-polarity)	[+]	78.20	96.84	86.37	84.97
	[-]	80.44	94.25	86.52	
	[0]	95.83	71.97	82.04	
HR (POS- polarity)	[+]	79.38	92.78	85.45	83.95
	[-]	81.77	91.80	86.09	
	[0]	89.18	73.34	80.31	
CSR (polarity)	[+]	78.56	98.92	87.42	85.37
	[-]	79.31	96.52	86.75	
	[0]	98.66	70.46	81.95	
HR (polarity)	[+]	79.41	98.92	87.95	85.87
	[-]	79.58	98.31	87.72	
	[0]	98.66	70.45	81.95	

Table 4.9 AMOD polarity rules.

Sequence type	CSR patterns			HR patterns		
	[+] class	[-] class	[0] class	[+] class	[-] class	[0] class
Governor only	+, _	-, _	0, _	N.A.	N.A.	N.A.
Dependent only	_, +	_, -	_, 0	N.A.	N.A.	N.A.
Governor, Dependent	+, + 0, + +,0	-, - 0, - -, 0 +, -	0, 0	0, + +, + +, 0	0, - -, 0 -, - +, - -, +	0, 0

In Table 4.9, the HR AMOD(-,+) $\rightarrow$ [-] pattern is covered in CSR by the AMOD(-,_) $\rightarrow$ [-] unigram rule. However, the CSR AMOD(,+) $\rightarrow$ [+] unigram rule with higher (F1=0.85) priority overrides the lower (F1=0.36) priority CSR AMOD(-,_) $\rightarrow$ [-] rule, which explains the lower positive [+] class precision and lower negative [-] class recall for the CSR polarity method, when compared to the HR rules as shown in Table 4.1. Therefore, the priority of the unigram pattern rules for different classes affects the performance when overlapping rules of lower priority are overridden. For example, the bigram AMOD(disaster[-], biggest[+]) should be classified as negative [-] class. However, the higher priority CSR AMOD(,+) $\rightarrow$ [+] rule overrides the lower priority CSR AMOD[-,_] $\rightarrow$ [-] rule and classified the bigram as positive [+] based on the polarity of the dependent term “biggest”.

ADVMOD Evaluation

The generated ADVMOD CSR pattern rules are listed in Table 4.10, with the top 10 generated ADVMOD CSR pattern rules sorted with respect to their F1 scores given in Table 4.11. CSR generates more pattern rules than HR due to the greater variety of part-of-speech types found in ADVMOD. In the Stanford Parser, the adverbial modifier can contain an adverb or an adverbial phrase, which can contain words that operate adverbially and are not restricted to adverbs. This leads to more possible pattern combinations in describing sentiment phrase for adverbial modifiers. The rules generated showed that CSR is able to detect significant pattern rules like ADVMOD(JJ-, RB-), which contain the different types of part-of-speech not discovered in HR. From Table 4.12, the ADVMOD HRs’ performance is significantly lower than the CSR performance. This is due to the variability of the part-of-speech (e.g., JJ, NN, etc.) type records, which are not discovered in HR. HR polarity method

performance improved significantly over the HR POS-polarity method, demonstrating the effectiveness of using the more general polarity features. This study's implementation classifies all unmatched bigrams as in the neutral [0] class, which explains the high recall and low precision (54.49%) values for the neutral [0] class in the HR POS-polarity method. Thet et al. (2010), in their implementation, classified all unmatched bigrams using a generalized default rule that leveraged the sentiment scores, which might have provided better performance.

Table 4.10 CSR generated pattern rules for ADVMOD typed dependency.

Positive class patterns	Negative class patterns	Neutral class patterns
● JJ+,_ (thoughtful) ● VB+,_ (magnified) ● NN+,_ (genius)	● JJ-,_ (unoriginal) ● VB-,_ (torturing) ● NN-,_ (lie)	● VB,_ (showed) ● NN,_ (product) ● JJ,_ (observed) ● RB+,_ (much) ● RB,_ (soon) ● DT,_ (a) ● IN,_ (on)
● _,JJ+ (important)	● _,RB- (unfortunately) ● _,JJ- (crap)	● _,RB (only) ● _,IN (at) ● _,NN (movie) ● _,VB (turn) ● _,JJ (next) ● _,DT (the)
● JJ+, RB+ (beautiful, incredibly) ● JJ+, RB (serene, seemingly) ● VB+, RB+ (delighted, undoubtedly) ● VB+, RB (engrossing, so) ● NN+, RB+ (fun, simply)	● JJ-, RB+ (intrusive, simply) ● JJ-, RB (flashy, visually) ● JJ-, RB- (dull, unspeakably) ● VB-, RB+ (bugs, constantly) ● JJ, RB- (dull, lethally) ● VB, RB- (dubbed, poorly) ● JJ+, RB- (interesting, barely) ● VB-, RB (forgotten, long) ● NN-, RB (violence, only) ● NN-, RB+ (labor, just) ● JJ-, DT (bad, that) ● VB-, RB- (mugs,mercilessly)	● VB, RB (find, only) ● VB, RB+ (cut, just) ● NN, RB+ (comparison, much) ● NN, RB (boys, only) ● JJ, RB (predictable, finally) ● JJ, RB+ (similar, so) ● VB, IN (finishing, at) ● RB+, RB (much, as) ● RB, RB (about, only) ● JJ+, RB (aware, partly)

Table 4.11 Top 10 ADVMOD CSR rules.

#	F1	Governor	Dependent	Class
1	1.00	NN(boys)	RB(only)	0
2	1.00	VB-(falling)	RB-(short)	-
3	0.99	VB(find)	RB(only)	0
4	0.93	JJ+(good)	RB+(really)	+
5	0.93	JJ-(stupid)	RB-(insanely)	-
6	0.93	JJ(untold)	RB(largely)	0
7	0.89	VB+(give)	RB+(actually)	+
8	0.84	JJ(paced)	RB-(poorly)	-
9	0.83	JJ+(aware)	RB(partly)	0
10	0.80	NN-(lie)	—	-

Table 4.12 Results for ADVMOD POS-polarity and polarity rules

Method	Class	Precision (%)	Recall (%)	F1 (%)	AvgF1 (%)
CSR (POS-polarity)	[+]	78.49	75.39	76.21	82.81
	[-]	82.36	94.72	87.81	
	[0]	86.62	82.58	84.42	
HR (POS-polarity)	[+]	40.00	23.42	29.20	41.57
	[-]	81.51	19.10	29.99	
	[0]	54.49	82.99	65.54	
CSR (polarity)	[+]	75.06	80.94	77.47	83.10
	[-]	80.75	98.15	88.38	
	[0]	89.58	78.34	83.45	
HR (polarity)	[+]	52.53	97.51	67.93	74.34

The generated ADVMOD polarity rules are shown in Table 4.13. Polarity conflicts were observed for the case of the ADVMOD(0,+) $\rightarrow$ ([+][0]) rule. In the ADVMOD(0,+) pattern, which is classified neutral [0] for CSR and positive [+] for HR, though HR correctly identified positive [+] class bigrams such as ADVMOD (orchestrates, beautifully), the vast majority of bigrams matching this particular pattern are in fact in the neutral [0] class, which only CSR was able to find. Examples of such neutral [0] class bigrams include ADVMOD(acted, mostly) and ADVMOD (built, entirely). Further analysis of the bigram word semantics is required to resolve such polarity conflicts.

Table 4.13 ADVMOD polarity rules.

Sequence type	CSR patterns			HR patterns		
	(+) class	(-) class	(0) class	(+) class	(-) class	(0) class
Governor only	+, _	-, _	0, _	N.A.	N.A.	N.A.
Dependent only	_, +	_, -	_, 0	N.A.	N.A.	N.A.
Governor, Dependent	+, + +, 0	-, - -, + -, 0 +, - 0, -	0, 0 0, +	+, + +, 0 0, +	-, - -, + -, 0 +, - 0, -	0, 0

DOBJ Evaluation

The CSR generated pattern rules are listed in Table 4.14, while the top 10 DOBJ CSR pattern rules are given in Table 4.15.

Table 4.14 CSR generated pattern rules DOBJ typed dependency

Positive class patterns	Negative class patterns	Neutral class patterns
● VB+,_ (deliver)	● VB-,_ (hate)	● VB,_ (describe) ● DT,_ (the) ● NN,_ (movie)
● _,NN+ (insight) ● _,JJ+ (alive) ● RB+ (enough) ● _,VB+ (love)	● _,NN- (headache) ● _,JJ- (weird)	● _,NN (thing) ● _,DT (a) ● _,VB (makes) ● _,CD (two) ● _,JJ (other) ● _,IN (at)
● VB, NN+ (give, blessing) ● VB+, NN (good, effort) ● VB+, NN+ (deserves, dignity) ● VB, RB+ (make, much) ● VB+, RB+ (enjoying, much) ● VB, JJ+ (made, richer) ● VB+, JJ+ (growing, more) ● VB, VB+ (reading, love)	● VB, NN- (gives, unease) ● VB-, NN (upset, viewers) ● VB-, NN+ (lost, ability) ● VB-, NN- (suffering, failure) ● VB-, PRP (depress, you) ● VB, JJ- (say, least) ● VB-, DT+ (unrewarding, all) ● JJ-, NN (clueless, combination) ● VB-, DT (fooled, some) ● JJ-, VB (surreal, dabbling)	● VB, NN (describe, vision) ● VB, PRP (call, me) ● VB, DT (catch, some) ● VB, WP (does, what) ● VB, CD (do, what)

Table 4.15 Top 10 DOBJ CSR rules

#	F1	Governor	Dependent	Class
1	1.00	VB(reading)	VB+(love)	+
2	1.00	VB+(growing)	JJ+(more)	+
3	1.00	VB(catch)	DT(some)	0
4	1.00	VB(describe)	NN(vision)	0
5	1.00	VB+(enjoying)	RB+(much)	+
6	1.00	—	JJ-(weird)	-
7	0.99	VB(act)	JJ(other)	0
8	0.99	VB(say)	JJ-(least)	-
9	0.93	VB(gives)	NN-(unease)	-
10	0.89	VB-(suffering)	NN-(failure)	-

Table 4.16 Results for DOBJ POS-polarity and polarity rules.

Method	Class	Precision (%)	Recall (%)	F1 (%)	AvgF1 (%)
CSR (POS-polarity)	(+)	76.61	54.73	62.84	80.95
	(-)	86.79	97.11	91.53	
	(0)	86.11	91.12	88.49	
HR (POS-polarity)	(+)	68.81	55.41	61.26	79.54
	(-)	85.80	95.45	90.05	
	(0)	85.96	88.80	87.32	
CSR (polarity)	(+)	75.18	56.20	63.72	81.45
	(-)	85.56	100	92.09	
	(0)	86.87	90.34	88.54	
HR (polarity)	(+)	70.24	99.56	82.22	88.09
	(-)	85.42	98.85	91.50	
	(0)	99.62	83.08	90.57	

From Table 4.16, there is significant improvement in performance by using polarity features (F1 of positive HR polarity rule = 82.22%) versus part-of-speech polarity features (F1 of positive HR POS-polarity rule = 61.26%). Improvements in F1 for HR were across the board at around 8%, whilst improvements for CSR were marginal at around 0.5%. This could be due to the presence of unigram rules in CSR that overrides the lower and correct priority rules, which affect performance. On the other hand, HR polarity contains a more comprehensive list of bigram pattern rules that give a high recall performance. The DOBJ polarity rules are similar except for the CSR unigram rules and the HR(+,0)→(+) rule, which was not discovered in the CSR

method as shown in Table 4.17. The confidence of $(\_,0) \rightarrow (+)$ is at 0.13, which is less than the minimum threshold of 0.5, and was not large enough for the $CSR(+,0) \rightarrow (+)$ rule to be generated, and the pattern for $(+,0) \rightarrow (+)$ was covered using the unigram $CSR(+,\_) \rightarrow (+)$ rule. As a result, the more generic unigram rule caused a lower recall value for the CSR polarity positive (+) class compared with HR polarity positive (+) class because the neutral (0) class unigram rule $CSR(\_,0) \rightarrow (0)$ has a higher $F1=0.79$ value, which overrides the positive unigram rule $CSR(+,\_) \rightarrow (+)$ of $F1=0.68$ in classifying $(+,0)$ patterns. This shows the dependency of CSR on the dataset polarity frequency.

Table 4.17 DOBJ polarity rules.

Sequence type	CSR patterns			HR patterns		
	(+) class	(-) class	(0) class	(+) class	(-) class	(0) class
Governor only	+, _	-, _	0, _	N.A.	N.A.	N.A.
Dependent only	_ , +	_ , -	_ , 0	N.A.	N.A.	N.A.
Governor, Dependent	+, + 0, +	0, - -, - +, -	0, 0	+, + +, 0 0, +	-, - -, + -, 0 +, - 0, -	0, 0

4.5.3 Discussion

This study explored the use of the CSR method to derive the polarity-predicting pattern rules at the blog posts' phrase level, in consideration of the neighboring terms of the target term. Results showed that CSR is capable of automatically generating a comprehensive list of bigram part-of-speech patterns compared to a heuristic approach, which improves subsequent sentiment classification performance. It was also shown that rules using polarity features perform better than those using part-of-speech polarity features, especially for the heuristic approach. In general, the heuristic rules

perform similarly to the CSR rules, which could be due to the careful construction of the knowledge base. The exception is for ADVMOD typed dependency where the heuristic rules did not include rules for the varied types of part-of-speech due to a restricted scope to verb phrases. This shows that the heuristic approach is not able to detect new pattern rules beyond the knowledge base, while CSR is able to discover new and existing pattern rules based on evidential support of the pattern item sets using computational analysis. Further to that, the rule discovery requires process is automatic, requiring only a manually tagged answer key to generate the CSR rules. The discovery of new additional pattern rules generally improves the overall performance of the CSR model. Furthermore, they could be used in linguistics studies to further enhance the knowledge base on the semantic dependencies of bigrams. CSR is also able to generate unigram pattern rules that are more generic in scope as compared to the bigram pattern rules. The unigram rules are important as they could cover the less frequent patterns and improve the robustness of the CSR model by reducing the effect of noisy data. It is observed that unigram rules improve recall performance while bigram rules tend to improve precision performance.

As shown in the results, performance for the rules is high when polarity sign is used as a feature. This could be because there is a common polarity pattern for the various part-of-speech terms within the specific typed dependencies. The generalization of the feature terms leading to the corresponding increase in their frequency count has enhanced the significance of the important pattern rules, which improves the overall prediction performance. Moreover, evaluating just the polarity of words could increase the prediction model performance as additional part-of-speech tagging may introduce errors.

The generated POS-polarity pattern rules can be generalized across domains as context has lesser effect at the bigram phrase level with only the part-of-speech and polarity of the bigram words being considered during processing. Similarly, the polarity-only pattern rules are domain independent where rules are processed without the need to consider the context. This is seen in the examples, where the AMOD(NN-, JJ-) POS-polarity pattern rules or AMOD((-),(-)) polarity-only pattern rule will give a negative output regardless of the domain analyzed.

Possible causes of error in the tests are the wrongly-parsed typed dependencies such as AMOD(regard:VB, be:VB), and sentiment lexicon with polarity errors in the typed dependency parsing are less than 5% and are controlled using the support and confidence threshold values, while the detected words with wrong polarity tags are corrected by the coders during the annotation process. It is noted that for CSR, the execution priority has to be given to the specific bigram rules over the generic unigram rules to maintain good performance. Furthermore, in CSR the execution priority of the overlapping rules from different classes poses a problem. A generic but higher priority rule could override a specific and lower priority rule from another class, causing the test item to be wrongly classified by the more generic rule. Nonetheless, with all learning algorithms, the discovery of rules in CSR is dependent on the dataset used where the distribution and frequency of the polarity cases could determine the rules to be derived, and as a result impact performance.

In addition, polarity conflict cases were systematically processed according to the priority of the generated rules in this study. A separate semantic analysis on the bigrams' words and part-of-speech to resolve polarity conflicts through identifying semantic relationships of the bigram words would have provided better performance. Though the results showed that rules using generalized features performed well,

analysis of the bigrams' semantics including the words and part-of-speech is still required to identify the more complex relationships between words, which are observed to influence phrase polarity. This study additionally analyzed these complex relationships between words by considering the intensifiers, maximisers, mitigators, minimisers, and negators found in the phrases, as well as taking into account the subjectivity of phrases that could affect the sentiment polarity output with further details given in the next Section 4.6.

4.6 Handling Complex Relationships

Though polarity pattern rules using generalized features could perform well, a semantic analysis of the words is still required to identify the more complex relationships between words that could influence phrase polarity. It was observed in Wilson et al. (2005) and Polanyi and Zaenen (2006) that there exist more complex relationships between words that could influence the sentiment polarity of phrases. Examples of these complex relationships include negation (e.g., “*not*” in “*not good*”), subjective terms found in domain terms (e.g., “*star*” in “*star trek*”), subjective terms appearing as neutral (e.g., “*fiction*” in “*science fiction*”), neutral terms appearing as subjective (e.g., “*red*” in “*red carpet*”), and intensifier and mitigation terms.

Intensifiers such as “*very*” intensify the sentiments of its adjacent term (e.g., “*very good*”), while mitigators such as “*few*” reduces the sentiment (e.g., “*few support*”). Wilson et al. (2005) considered a word to be an intensifier if it appears in a list of intensifiers and if it precedes a word of the appropriate part of speech (e.g., an intensifier adjective must come before a noun). This study compiled intensifier words from those listed in Quirk et al. (1985) intensifiers identified from existing entries in

the subjectivity lexicon, and intensifiers identified during exploration of the developmental data. Quirk et al. (1985) described the effects of intensifiers as scaling upwards from an assumed norm, and identified two intensifier types: maximizers (e.g., absolutely, completely, and perfectly) and boosters (e.g., very much, a lot, and deeply). Mitigators are described as generally having a lowering effect on the force of the term and could be categorized into four groups: approximators (e.g., almost, nearly, and as good as), compromisers (e.g., kind of, sort of, quite, and rather), diminishers (e.g., mildly, partly, and somewhat), and minimizers (e.g., barely, hardly, and little). Polanyi and Zaenen (2006) observed that there are contextual valence shifters that could influence the sentiment polarity of sentences. These include negatives (e.g., “*is not*”), intensifiers (e.g., “*deeply*”), presuppositional items (e.g., “*barely*”), modals (e.g., “*might*”), and irony. Negation can be local (e.g., “*not good*”), or involve longer distance dependencies such as the negation of the proposition (e.g., “*does not look very good*”). Simple negatives include “*never*”, “*none*”, “*nobody*”, “*nowhere*”, “*nothing*”, “*neither*”, etc. Polanyi and Zaenen (2006) further observed that modal operators set up a context of possibility or necessity and in texts they initiate a context in which valence terms express an attitude towards entities which do not necessarily reflect the author’s attitude towards those entities in an actual situation under discussion. Although previous studies (Wilson et al., 2005; Polanyi and Zaenen, 2006) have observed more types of complexities between words, this study focused only on the complex relationships found between words in a phrase. These complex relationships in a phrase are not easily detected using polarity pattern rules, and hence further analysis on the semantic meaning of words in a phrase is required to improve the phrase-level sentiment polarity classification performance. For example, subjective words can be found in a neutral sentiment polarity phrase such as the negative word “*close*” in the

neutral phrase “*close up*” as in “*take a close up look*”. The sentiment polarity of words in a phrase could also be reversed, for example the negative word “*break*” in the positive phrase “*break through*”.

Multi-word subjective phrases that influence sentiment polarity output can be found in sentences. However, as typed dependencies contain only the grammatical relationship within bigrams, they could not be used to detect the subjectivity within multiple words in a phrase. For example, in the sentence “*The ra-ra skirts were all the rage in the 1980s*”, typed dependency analysis alone could not detect the positive expression “*all the rage*.” as shown in Figure 4.8. A multi-word analysis is required to identify the subjective phrases.

DET(skirts-3, The-1)
AMOD(skirts-3, ra-ra-2)
NSUBJ(rage-7, skirts-3)
COP(rage-7, were-4)
PREDET(rage-7, all -5)
DET(rage-7 , the -6)
PREP(rage-7, in-8)
DET(1980s-10, the-9)
POBJ(in-8, 1980s-10)

Figure 4.8 Subjective phrases across typed dependencies.

In previous studies, Wilson et al. (2005) used a machine-learning approach to predict the phrase sentiment polarity by identifying the appropriate subjective phrase features, while Moilanen and Pulman (2007) analyzed the word constituent within phrases to detect sentiment polarity. The use of a linguistic approach in considering the more complex relationships between words, including subjective phrases to improve the performance of phrase-level sentiment polarity classification as well as to provide a clear explanation of the grammatical relationships that gives the sentiment polarity are further discussed. In this study, subjective phrases are defined as phrases that contain

sentiment polarity orientation. For example, the sentence "*The team that brought to life the worlds of Avatar is breaking new ground*" contains the positive phrases "*brought to life*" and "*breaking new ground*", which cause the overall sentence sentiment output to be positive. However, as typed dependencies contain only the grammatical relationship within bigrams, they could not detect the subjectivity within multiple words in a phrase. As can be seen from Figure 4.9, using the typed dependency polarity pattern rules that relied on the grammatical relationships, and the polarity patterns within the bigrams are not sufficient to determine the subjectivity found within the multi-word expressions, which are otherwise neutral terms on their own. For example, using the $DOBJ(breaking[0],ground[0]) \rightarrow [0]$ and $AMOD(ground[0],new[0]) \rightarrow [0]$ typed dependencies polarity pattern rules would have resulted in an incorrect neutral polarity output for the positive phrase "*breaking new ground*".

DET(team, The)
NSUBJ(breaking, team)
NSUBJ(brought, that)
RCMOD(team, brought)
PREP(brought, to)
POBJ(to, life)
DET(worlds, the)
DOBJ(brought, worlds)
PREP(worlds, of)
POBJ(of, Avatar)
AUX(breaking, is)
AMOD(ground, new)
DOBJ(breaking, ground)

Figure 4.9 Generated typed dependencies.

On the other hand, as seen from Figure 4.10, by extracting and analyzing the verb phrases "*brought to life*" and "*breaking new ground*" for subjectivity, the positive sentiment polarity expressed by the phrase can be detected correctly. Hence, there is a

need to evaluate the expressions within phrase entities for predicting the phrase subjectivity accurately. Examples of phrase entities include adverbial phrase (ADVP), noun phrase (NP), verb phrase (VP), etc.

```
((ROOT
(S
(NP
(NP (DT The) (NN team))
(SBAR
(WHNP (WDT that))
(S
(VP (VBN brought)
(PP (TO to)
(NP (NN life)))
(NP
(NP (DT the) (NNS worlds))
(PP (IN of)
(NP (NNP Avatar))))))
(VP (VBZ is)
(VP (VBG breaking)
(NP (JJ new) (NN ground))))))
```

Figure 4.10 Generated phrase structure tree.

In order to handle the subjective phrases that could influence the sentiment polarity output, a prior knowledge base is used to detect subjectivity in the multi-word expressions to identify phrase subjectivity and adjust the output polarity accordingly. For example, a prior positive and negative polarity tag could be assigned to the phrase “*feel at home*” and “*red handed*” respectively. This subjectivity phrase lexicon containing phrase-lexicon terms could be built from extracting subjective phrases from the training dataset and other knowledge base containing phrase information (e.g., phrasal verbs dictionary and idioms dictionary), and humans can manually annotate the

phrases with the mutually agreed sentiment polarity. Examples of subjective phrase terms are shown in Table 4.18 and in Appendix A.

Table 4.18 List of sample subjective phrases.

Polarity	Example Phrases
Positive	red carpet, above par, cutting edge, stand the test
Negative	set you back, arm and a leg, wet blanket, wear out
Neutral	an hour long, up close, big apple, dead even

In the steps to evaluate phrase subjectivity, the proposed CSR typed dependencies polarity pattern rules are processed before the subjective phrase polarity rules. In order to relate the typed dependencies to their respective phrase entity, there is a need to analyze the phrase structure tree where the grouped typed dependencies are evaluated under the corresponding phrase according to the phrase head word (=H).

```
(VP (VB=H get)
  nsubj(get, he)
  aux(get, will)
  prep_into(get, water)
  prep_with(get, director)
  (NP (JJ hot) (NN=H water)))
  amod(water, hot)
  (NP (DT the) (NN=H director))))))
  det(director, the)
```

Figure 4.11 Phrase entities with the corresponding typed dependencies.

For example, from Figure 4.11 in the sentence “*he will get into hot water with the director*”, the typed dependencies with “*get*” as their governor term are identified with the verb phrase “VP (VB=H *get*)”. This is similarly done for the governor terms “*water*” and “*director*” with their corresponding phrase entities.

```

(ROOT
  (S=H
    (NP (PRP=H he))
    (VP=H (MD=H will)
      (VP (VB=H get)
        (PP (IN=H into)
          (NP (JJ hot) (NN=H water))))
        (PP (IN=H with)
          (NP (DT the) (NN=H director)))))))))

```

Figure 4.12 Generated phrase structure tree.

From Figure 4.12, focusing on the negative expression "*get into hot water*", the lower nested level noun phrase "(NP (JJ **hot**) (NN=**H water**))" are evaluated using the $AMOD(water[0], hot[+]) \rightarrow [+]$ typed dependency pattern rule to give an interim positive output. Next, using a bottom-up approach in evaluating the rules, the nested level of each item in the phrase structure tree is taken into consideration to ensure that only items from the same level are evaluated and the resultant polarity is propagated to subsequent levels for further evaluation. The positive AMOD output is recursively evaluated with the corresponding neutral prepositional (PP) typed dependencies and the AUX typed dependency "AUX(*get*, *will*)" to obtain the verb phrase "*will get into hot water*", and the verb phrase is matched with the terms from the subjective phrase lexicon to get the negative polarity output for the phrase "*get into hot water*". The phrase structure tree is used to identify the nested level of iteration items to ensure that only same-level items are evaluated before moving on to higher level items. In this study, the typed dependency tree is not used as it is harder to identify the phrase structure from a typed dependency tree, as shown in Figure 4.13 compared with the phrase structure tree in Figure 4.12.

Figure 4.13 Typed dependency tree.

4.7 Sentiment Analysis Model Evaluation

This study proposes a Sentiment Analysis model that uses typed dependencies rules as well as further consideration of the complex relationships between words to predict the sentiment polarity output. An overview of the proposed Sentiment Analysis model is shown in Figure 4.14.

Figure 4.14 Overview of the proposed Sentiment Analysis model.

First, the sentence dataset is parsed using the Stanford Parser to generate the typed dependencies for the words. Second, the words are matched with a list of lexicons and the polarities of words are adjusted accordingly. Next, the rules are recursively

evaluated to give the output polarity of the corresponding level of the phrase structure tree. Fourth, at each level, the words in the phrase are matched against a subjective phrase lexicon to extract multi-word expressions that are subjective. In the final recursion, the sentence's sentiment polarity is predicted.

The typed dependency polarity pattern rules used in the evaluation, as shown in Table 4.19 to Table, 4.22, were adopted from Thet et al. (2010) and Tan et al. (2012), which also considers the complex relationships between words in a phrase through the use of a subjective phrase lexicon. Sample coding for the rules is given in Appendix C.

Table 4.19 AMOD typed dependency rules

(G)overnor	(D)ependent	Sentiment Score	Examples
+ / 0*	+ / 0	$+(G +((1-G)* D))$	<i>Fantastic(D)[1.0]** Product(G)[0.0] → [1.0]</i>
+ / 0	-	$- D $	<i>Worst(D)[-1.0] Scenario(G)[0.0] → [-1.0]</i>
-	+ / 0	$-(G +((1-G)* D))$	<i>Strong(D)[0.5] Complaint(G)[-1.0] → [-1.0]</i>
-	-	$-(G +((1-G)* D))$	<i>Little(D)[-1.0] Awkward(G)[-1.0] → [-1.0]</i>
*: 0 indicates neutral, **: [ ] denotes sentiment polarity score			

Table 4.20 ADVMOD typed dependency rules

(G)overnor	(D)ependent	Sentiment Score	Examples
+ / 0	+ / 0	$+(G +((1-G)* D))$	<i>Enthusiastically(D)[1.0] Responded(G)[0.0] → [1.0]</i>
+ / 0	-	$- D $	<i>Glaringly(D)[-1.0] Stared(G)[0.0] → [-1.0]</i>
-	+ / 0	$-(G +((1-G)* D))$	<i>Probably(D)[0.0] Disappointed(G)[-1.0] → [-1.0]</i>
-	-	$-(G +((1-G)* D))$	<i>Strictly(D)[-1.0] Prohibited(G)[-0.05] → [-1.0]</i>

Table 4.21 DOBJ typed dependency rules

(G)overnor	(D)ependent	Sentiment Score	Examples
+ / 0	+ / 0	$+(G +((1-G)* D))$	<i>loved</i> (G)[1.0] the <i>phone</i> (D)[0.0] → [1.0]
+ / 0	-	$- D $	<i>face</i> (G)[0.0] an <i>onslaught</i> (D)[-0.5] of windows messages → [-0.5]
-	+ / 0	$-(G +((1-G)* D))$	<i>killed</i> (G)[-0.5] <i>servers</i> (D)[0.0] → [-0.5]
-	-	$-(D + (1- D)* G)$	<i>destroyed</i> (G)[-0.5] the <i>virus</i> (D)[-1.0] → [-1.0]

Table 4.22 NSUBJ typed dependency rules

(G)overnor	(D)ependent	Sentiment Score	Examples
+ / 0	+ / 0	$+(G +((1-G)* D))$	The <i>discussions</i> (D)[0.0] <i>resolve</i> (G) [0.5] the problems → [0.5]
+ / 0	-	If $ D > G $ output $- D $ Else output $ G $	The <i>failures</i> (D)[-1.0] <i>changes</i> (G)[0.0] the power source → [-1.0]
-	+ / 0	$- G $	The <i>thing</i> (D)[0.0] really <i>bugged</i> (G)[-0.5] me → [-0.5]
-	-	$-(G +((1-G)* D))$	The <i>virus</i> (D)[-1.0] <i>destroyed</i> (G)[-0.5] the codes → [-1.0]

The adjectival modifier (AMOD) typed dependency and adverb modifier (ADVMOD) typed dependencies were evaluated as these are deemed to contain the vast majority of opinionated expressions (Jakob et al., 2009; Wilson et al., 2006). The AMOD of a noun phrase is any adjectival phrase that modifies its meaning. Similarly, an ADVMOD of a word is an adverb or adverbial phrase that modifies its meaning. The governor is a term or phrase in the typed dependency bigram that is modified by the dependent. The direct object (DOBJ) typed dependency and nominal subject (NSUBJ) typed dependency are evaluated to handle the predicate (VP-to-NP) relation and the relation between noun phrase (NP) and verb phrase (VP), respectively in a sentence. In addition, the negation (NEG) typed dependency is evaluated to detect negation in the sentence, where the output is the reverse of the recursion polarity.

The intensity of sentiment polarity affects the eventual output polarity. Hence, polarity score values were assigned to the lexicon terms to account for the intensity of sentiment polarity. The scoring algorithm used in the typed dependency rules were adopted from Thet et al. (2010). Due to the subjective nature of sentiment polarity, instead of using a continuous score value, a discretized score range within the values [-1, -0.5, 0, 0.5, 1] was used to generalize the sentiment polarity classification of terms into the respective categories of sentiment polarity intensity. The sentiment scores of the words were derived from the subjectivity lexicon by Wilson et al. (2005) where the words are tagged with the sentiment intensity based on a strong or weak subjectivity category. The subjectivity categories correspond to scores that indicate intensities ranging from high negative [-1], negative [-0.5], neutral [0], positive [0.5], to high positive [1]. Examples of the prior sentiment scores assigned to the terms are shown in Table 4.23.

Table 4.23 List of prior sentiment scores assigned to word terms

Category	Score	Examples
High negative	-1	Conspire, abuse
Negative	-0.5	Messy, defect
Neutral	0	Obvious, thought
Positive	0.5	Unique, impress
High positive	1	Awesome, righteous

Table 4.24 Lexicon based rules.

(M)odifier	(T)erm	Sentiment Score	Examples
Negation	T	$-1 * T$	Hardly good
Intensify	T	$(2 * T)$, max value = ± 1	Highly recommended
Mitigate	T	$(0.5 * T)$	Slightly better
Maximize	T	$1 * \text{Polarity}(T)$	Totally bad
Minimize	T	$0.25 * T$	Least effort

To handle the more complex relationships between words, additional lexicons and polarity rules, as shown in Table 4.24, are incorporated into the Sentiment Analysis model. The idea is to identify key expressions that could alter the intensity of phrase intensity. These lexicon based rules include negators (20 terms), intensifiers (42 terms), mitigators (45 terms), maximizers (43 terms), and minimizers (11 terms). The lexicon terms are collected from the training dataset and the given lists from Quirk et al. (1985), with examples shown in Appendix A. This study further analyzes the multi-word expressions at phrase-level in order to detect phrase subjectivity and adjust the output polarity to the correct polarity. A subjectivity phrase lexicon containing 5330 terms was built from the training dataset and a list of sentiment polarity annotated verb and idiom phrases.

From Table 4.24, in the negation rule, the polarity score of the word that is modified by the negator would be reversed. For the intensify rule, the polarity score of the modified word is doubled, but limited to a value of ± 1 in consideration for the increasing effect of the intensifier. Conversely, the polarity of the modified word is halved in the mitigate rule. The polarity score is maximized to ± 1 in the maximize rule to denote the upper extreme of the intensity scale. In the minimize rule, the polarity score of the modified word is severely reduced to account for the lower extreme of the intensity scale. Additionally, the study uses a domain lexicon (40 terms) to correct domain-specific subjective words (e.g., “*hard disk*” and “*open source*”). Next, the typed dependencies to be evaluated are grouped under the corresponding phrase according to the phrase head word. The phrases are then matched with the subjective phrase lexicon to give the annotated phrase polarity. The study uses a bottom-up approach in evaluating the rules, taking into consideration the nested level of each item in the phrase structure tree to ensure that only items from the same level are evaluated

and the resultant polarity propagated to subsequent levels for further evaluation. For example, in the positive expression “*no doubt, it is good*”, if the nested levels are not considered, the evaluation would have incorrectly negate the neutral output of (“*good[+]*” + “*doubt[-]*”). While other studies (Prasad et al., 2006; Ruppenhofer et al., 2008; Somasundaran and Wiebe, 2009) proposed determining the polarity of discourse relations in a sentence through learned probabilities, this study performs simple discourse analysis through analyzing the effects of connectives such as “however” and “but” at typed dependencies level. A fine-grained linguistic approach was used in this study as compared to the approach of using learned probabilities because the latter method does not consider the relationship between words in each specific post. There is a need for closer attention to the specific discourse sentiments at clause level to identify the opinions from the textual content. Nevertheless, future work could involve evaluating the automatic detection of clause-level typed dependencies polarity pattern rules to improve sentiment prediction performance. Samples of clause-level typed dependencies are given in Appendix G.

The use of typed dependency pattern rules provides a foundation for analyzing the semantic orientation of words with direct subjectivity expressions. Further analysis on the complex relations between words could be applied to decipher the polarity conflict situations to enhance sentiment polarity prediction performance.

4.7.1 Dataset Used

In this sentiment analysis model evaluation, the blog dataset available from Zafarani and Liu (2009) was used, where long sentences (10 words or longer) were extracted in an attempt to handle the more complex relationship between words cases. The dataset contains blog content extracted from TUAW (www.tuaw.com), a blog

dedicated to disseminating information on Apple products and services. A total of 402 sentences were used for lexicon building, and another 1000 sentences were used as the test data, which contain a balanced distribution of 300 positive and 300 negative sentences, and 400 neutral sentences. Two coders manually classified the sentiment orientation of each sentence into positive (+), negative (-), or neutral (0). The intercoder agreement Cohen's kappa index (Cohen, 1960) is within the acceptable range at 0.78. The conflicting tags by the two coders were reviewed and then manually re-classified and used as the answer keys.

4.7.2 Evaluation Results and Discussion

The baseline, typed dependency (TD)-only, and proposed Sentiment Analysis model approaches were used in the evaluation testing. The baseline used the average count of the positive terms and negative terms in the sentence to predict the polarity of the sentence. In the TD-only approach, only the AMOD, ADVMOD, NSUBJ, DOBJ, and NEG typed dependency polarity pattern rules were used in the evaluation. Lastly, the proposed Sentiment Analysis model was tested. All three tests were conducted on the testing dataset, and their results were compared using the F1 score measure, with the formula given as $F1 = (2 \cdot \text{Precision} \cdot \text{Recall}) / (\text{Precision} + \text{Recall})$. The test results for the baseline, (TD)-only, and Sentiment Analysis model are shown in Table 4.25, Table 4.26, and Table 4.27 respectively.

Table 4.25 Baseline test results.

Polarity	Precision	Recall	F1-Score	Average F1-Score
Neutral	56.64%	65.00%	60.54%	54.85%
Positive	51.69%	61.00%	55.96%	
Negative	62.57%	39.00%	48.05%	

Table 4.26 Typed dependency (TD)-only test results.

Polarity	Precision	Recall	F1-Score	Average F1-Score
Neutral	53.04%	89.50%	66.60%	56.43%
Positive	64.44%	38.67%	48.33%	
Negative	83.45%	40.33%	54.38%	

Table 4.27 Proposed Sentiment Analysis model test results.

Polarity	Precision	Recall	F1-Score	Average F1-Score
Neutral	55.94%	76.31%	64.56%	59.39%
Positive	61.27%	58.00%	59.59%	
Negative	74.71%	42.33%	54.04%	

From the results tables, the average F1-score for the baseline approach (54.85%) is relatively lower than the TD-only (56.43%) and the Sentiment Analysis model (59.39%) approach. The baseline approach only considers the number of positive and negative terms and does not take into account the possible syntactic or semantic relationships between words that could influence the sentence sentiment polarity. This is seen in the case for the expression “*remove the bug*”. The baseline approach would have counted two negative terms (“*remove*” and “*bug*”) and output a negative sentiment polarity for the sentence. In the TD-only and Sentiment Analysis model approach, the DOBJ(remove[-],bug[-])→[+] polarity pattern rule would have given the correct positive sentiment output. The low negative F1-score (48.05%) of the baseline could be due to the baseline approach not being able to detect negation of positive sentiments resulting in a low recall of negative sentiment polarity cases.

Although the Sentiment Analysis model performed only around 3% better than the TD-only approach in terms of overall F1-score, its positive F1-score (59.59%) is more than 10% better than that (48.33%) for the TD-only approach. This is because the test dataset contains a large proportion of long sentences, whose structure could

significantly influence the sentiment polarity output; the TD-only approach is unable to account for the syntactic phrase structure of the sentence while analyzing the words' sentiment polarity. Additionally, the Sentiment Analysis model approach considered the complex relationships between words, such as subjective phrases, intensifiers, mitigators, maximizers, minimizers, and domain terms, which could provide a more accurate representation of the sentence sentiment polarity. However, the small improvement (2.96%) in overall F1 performance could be due to the fact that the test dataset may not contain enough subjective phrases that are found in the subjective phrase lexicon. Applying the subjective phrase analysis on a dataset which contains a significant amount of subjective phrases that alter the polarity outcome could improve overall sentiment prediction performance. It was observed that subjective phrases not only determine the sentiment polarity of phrases, but could also alter the intensity of sentiment polarity through intensification (e.g., “long since” in “he has long since won the award”), maximization (e.g., “the one and only”), mitigation (“more or less”), minimization (e.g., “not going anywhere” in “the product is not going anywhere”).

The baseline approach could sometimes predict a correct sentiment polarity output without considering the subjective phrases as shown in the single count for “good” in the positive sentence, “Now I don't(negation) use the term back seat(- expression) lightly since Macs still being in the car for the ride is a good(+) thing.”. Nevertheless, the Sentiment Analysis model provides a deeper analysis of the relationships between words in predicting the output.

Errors in the Sentiment Analysis model include subjective phrases that are not identified in the subjective phrase lexicon. Examples include the negative term “cut” which is neutral in polarity for the expression “you can download, print, cut out and fold together”. While the subjectivity lexicon contains quality annotation of subjective

terms, there are still subjective words (e.g., recognize, trained, etc.) that are not included. It is also possible to provide a wrong sentiment polarity to words due to different context. For example, the word “*little*” is considered negative in the subjectivity lexicon, but is neutral in context in the sentence “*put up a little gallery with a bunch of the template...*”. Errors were also found in the typed dependencies parsing. For example, negation was not identified in the phrase “no bargaining chips”, where only the *det(chips, no)* and *nn(chips, bargaining)* typed dependencies were generated.

4.8 Summary

The study in this chapter established a sentiment detection approach in consideration of the complex relationships between words to identify influence diffusion between the linked bloggers to fulfill the research objective (RO2: *Establish a sentiment analysis approach that takes into consideration the complex relationships between words in blog posts to automatically detect influence diffusion between linking and linked blog posts*). This was achieved through generating the typed dependency rules automatically to predict phrase-level polarity using the CSR approach and further consideration for the complex relationships between words. This study evaluated the effectiveness of CSR derived rules for three major typed dependencies against the heuristic based approach. Results showed the CSR approach generated a more comprehensive list of rules which improved sentiment prediction performance. In addition, it was observed that using polarity rules provides results that are on-par, if not better compared to POS-polarity rules. This will improve the CSR and heuristic based approaches and remove the need to identify the part-of-speech

(POS). A second evaluation test on the significance of the complex relationships between words was conducted and results showed that performance in sentiment polarity prediction improved when semantic rules were incorporated in consideration of these complex relationships.

The results and findings of this sentiment analysis study were used to detect the persona of bloggers through analyzing the sentiments expressed towards the bloggers in the subsequent study on bloggers' influence style in Chapter 5, as well as to detect the sentiment values between links in the proposed influence diffusion detection model in Chapter 6.

CHAPTER 5

BLOGGERS' INFLUENCE STYLE ANALYSIS

5.1 Introduction

The aim of the study in this chapter is to profile the influence styles of bloggers based on automatic detection and analysis of the linked blog posts through applying research method (RM3-i: *Determine engagement style, persuasion style, and persona of the bloggers*) in response to the research gap (RG3: *Influence diffusion could not be described using a single detached feature based on the existence of links or similar sentiment polarity*). The proposed influence style model describes the influence of bloggers based on three possible styles: engagement style, persuasion style, and persona of the bloggers. Engagement style indicates the frequency, scope, originality, and consistency of the blog postings. Persuasion style refers to appeals to reasons or emotions displayed in the blog posts, and persona is the degree of compliance exerted by the bloggers. The identified influence styles would give a better description and understandings of the influence exerted by the bloggers, and are used as features in the influence detection model in Chapter 6.

Section 5.2 describes the background of the study followed by Section 5.3 which gives an overview of the proposed influence style model. Section 5.4 provides the influence style analysis framework where the methods for identifying the influence styles are given. Details of the dataset used in this study and descriptions of the blogger influence style profiling evaluation process are provided in Section 5.5 and 5.6

respectively. Section 5.7 discusses the evaluation results with the summary of study given in Section 5.8.

5.2 Background

Previous studies focused on a single notion that influence exists between linked bloggers (Adar and Adamic, 2005; Agarwal and Liu, 2008) or bloggers with similar sentiment polarity (Leskovec et al, 2010; Cai et al., 2011; Li et al., 2011), but ignored details of other aspects of influence. Bloggers are not restricted to a monolithic description of influence and may differ in their influence styles. Knowing the manner in which influence is exerted can provide a better understanding of the influence diffusion process between linked bloggers.

A blog site could contain a single blogger, and is referred to as a personal blog, while a blog site containing a group of bloggers is known as a multi-author blog as shown in Figure 5.1 for the blog site MacRumors ([http://www. macrumors.com](http://www.macrumors.com), retrieved on 29 Oct 2012).

Figure 5.1 Bloggers of multi-author blog site MacRumors.

Multi-author blogs have certain characteristics amongst their bloggers such as common shared goals and guidelines, and the editor guides and organizes the activities of the

blog. One reason for having multiple authors on a blog is the wider variety of opinions and ideas that could cater to more readers. Nevertheless, the greater variety of opinion expressions could lead to differences in the bloggers' influence styles. Influence style refers to the manner in which the blogger exerts influence through the blog postings. For example, a blogger could be subjective in persuasion influence style as seen in the expression "*The much anticipated iPad mini by Apple is finally here.*" as compared to the non-subjective expression "*Apple has launched the iPad mini.*" indicating an objective style with both expressions communicating the same message.

Figure 5.2 Description of blog site MacRumors (<http://www.macrumors.com>). Retrieved on 29 Oct 2012.

By identifying each blogger's influence style, this study provides more in-depth profiles of the bloggers. For example, bloggers' influence could be better described through identifying the bloggers as active participants or just passive sharers, whether they focus on specific topics or products, as seen in Figure 5.2, or they are broad in their scope, whether they express ideas in a rational or subjective manner, or whether their posts are received positively or negatively by the readers. The influence styles provide a better understanding of the bloggers' influence, which could be used as features to improve influence detection performance.

5.3 Influence Style Model

An overview of the Influence Style Model with descriptions of the influence styles and types, together with the corresponding models is shown in Table 5.1. The influence style model adapted the four general influence types from the Klout Influence Matrix⁵ shown in Figure 5.3 to describe the engagement style of the bloggers.

Figure 5.3 Klout's Influence Matrix

⁵ <http://hm-2onmymind.com/2011/03/26/klout-a-social-influencemetric/klout-influence-matrix> retrieved on 5 May 2012.

Table 5.1 Influence style model.

Influence Style	Type	Description	Corresponding model
Engagement	Participating-Listening	Describes the blogs' or bloggers' participation level in posting specific target topics.	<i>Authority</i> (Cialdini, 2001)
	○ Participating	Refers to blogs or bloggers that actively post articles on target topics and share information readily.	E.g., <i>Feeder</i> from Klout influence matrix
	○ Listening	Refers to blogs or bloggers that do not share much but could follow the blog posts with a low level of posting activities on target topics.	E.g., <i>Observer</i> from Klout influence matrix
	Broad-Focused	Describes the scope of the topics discussed in the blog posts.	<i>Authority</i> and <i>Scarcity</i> (Cialdini, 2001)
	○ Broad	Refers to bloggers that post a wide range of topics.	E.g., <i>Broadcaster</i> from Klout influence matrix
	○ Focused	Refers to bloggers that concentrate on specific topics which they are normally good at.	E.g., <i>Specialist</i> from Klout influence matrix
	Creating-Sharing	Describes the level of blog posts' content originality.	<i>Informational influence</i> (Cialdini and Goldstein, 2004), <i>Scarcity</i> (Cialdini, 2001), <i>Objective consensus</i> approach (Mackie, 1987) E.g., <i>Thought Leader</i> from Klout influence matrix
	○ Creating	Refers to blogs or bloggers that are original in their blog post content.	E.g., <i>Curator</i> from Klout influence matrix
	○ Sharing	Refers to blogs or bloggers that tend to share by restating the content of the linked blog posts.	

Table 5.1 Influence style model (cont.).

Influence Style	Type	Description	Corresponding model
Engagement	Consistent-Casual	Describes the consistency and duration of blogs' or bloggers' ability to garner in-links.	<i>Consistency and Commitment</i> (Cialdini and Trost, 1998), <i>Authority and Reciprocity</i> (Cialdini, 2001)
	○ Consistent	Refers to blogs or bloggers that exhibits continual ability to garner in-links over a duration of time.	E.g., <i>Pundit</i> from Klout influence matrix
	○ Casual	Refers to blogs or bloggers that is irregular and intermittent in their ability to garner in-links.	E.g., <i>Explorer</i> from Klout influence matrix
Persuasion	Subjective-Objective	Describes the types of Persuasion style	<i>Emotional influence</i> (Brehm, 1966); <i>Affect and Arousal</i> (Cialdini and Goldstein, 2004)
	○ Subjective	Refers to blogs or bloggers who are opinionated and expressive in stating their viewpoints	
	○ Objective	Refers to rationale blogs or bloggers who express their opinions through appeals to reasons.	
Persona	Positive-Negative-Controversial	Describe the types of Persona	
	○ Positive	Refers to blogs or bloggers with a strong degree of compliance from their followers	<i>Compliance and Conformity</i> (Kelman, 1958)
	○ Negative	Refers to blogs or bloggers with strong degree of disagreement and dissent from their followers.	<i>Dissent</i> (Milgram, 1963)
	○ Controversial	Refers to blogs or bloggers that creates controversies from their postings and are both likely to be challenged or supported by many.	<i>Controversy</i> (Cai et al., 2011)

The Klout Influence Matrix developed the concept of influencer type with their individual style in relation to the online social media domain to describe the influence styles of bloggers. According to Klout there are 16 different types of online social media influencers with their individual engagement style, ranging from those who actively share and participate, to those who like to explore, observe and report. The 16 types of influencers identified are: Curator; Broadcaster; Syndicator; Feeder; Tastemaker; Celebrity; Thought Leader; Pundit; Dabbler; Conversationalist; Observer; Explorer; Socializer; Networker; Activist; Specialist. This study limits its scope to the analysis of the four broad based *Participating-Listening*, *Broad-Focused*, *Creating-Sharing*, and *Consistent-Casual* types as analyzing the specific style of each blogger type would require further contextual information that is not found in the blog post content. While Klout developed the concept of influencer type with their individual style, this study propose methods to measure engagement influence style based on the *Participating-Listening*, *Broad-Focused*, *Creating-Sharing*, and *Consistent-Casual* influence style types through analyzing the blog content and features.

Engagement style refers to the frequency, scope, originality, and consistency in which the influencer presents the blog post content. A blogger who is a subject expert would tend to garner more in-links as well as post frequently within the specific target topics scope indicating a high level of authority (Cialdini, 2001). The bloggers' authority could be derived from the depth and extent of dominance in the target topics, measured through the *Participating-Listening*, *Broad-Focused*, and *Consistent-Casual* influence types respectively in this study. The *Participating-Listening* type describes the bloggers' participation level in posting specific target topics. The listening type usually do not post much, but rather follow the blog posts with a low level of posting activities, while a participating type would actively post articles on target topics and

share information readily. The *Broad-Focused* type describes the scope of the topics discussed in the blog posts. A broad type blogger would post a wide range of topics, and a focused type would concentrate on specific topics on which they are domain experts. *Creating-Sharing* type indicates the originality of the blog posts' content. A sharing blogger tends to restate the content of the linked blog posts. On the other hand, a creating blogger posts original content. For the *Consistent-Casual* style, the *consistent* type refers to bloggers that could continually attract in-links over a duration of time, whereas a *casual* blogger is irregular or intermittent in their number of in-links. The extent to which one's commitments are made actively is a powerful determinant of the likelihood of compliance (Cialdini and Trost, 1998). The influence style model measures commitment through the consistency of the target blogger's ability to garner in-links.

Persuasion is a process of influence through appeals to reason or appeals to emotion (Brehm, 1966). Subjectivity, that is, opinionated or emotional phrases expressed in the blog content could provide clear evidence of the blogger's persuasion style (Cialdini and Goldstein, 2004). This study evaluates persuasion through analyzing the subjectivity expressed in the blog posts content.

Kelman (1958) defined *Compliance*, a form of conformity as the expression of agreement towards the people or groups who are influential to the individual, while possibly keeping one's own original beliefs. The degree of compliance towards the blog site is measured based on the persona of the bloggers in the model. Persona style shows the degree of compliance exerted by the influencer, and is assessed through analyzing the sentiments on common topics between the in-link blog post and target blog post. *Positive persona* describes bloggers with high positive influence, where their links from others often indicate approval and agreement. Milgram (1963)

observed that dissent within a social network can greatly wane the strength of an influence. The influence style model measures dissent through evaluating the negative persona of bloggers. *Negative persona* represents bloggers with high negative influence, and their links from others usually express disagreement or distrust. The *Controversial persona* represents bloggers that are both likely to be challenged or supported by many (Cai et al., 2011), which is shown in the high number of agreement and disagreement in-link blog posts. Blog influence is specific to the topics or aspects discussed in the blog posts (Agarwal et al., 2008; Somasundaran and Wiebe, 2010). The influence style model considers the bloggers' influence in relation to specific target topics by identifying the topics, aspects, and features discussed in the blog posts.

5.4 Influence Style Analysis Framework

Figure 5.4 shows the influence style analysis framework. In-links are incoming links to the target blog post, while out-links are links to other blog posts from the target blog post. The in-link blogs URL were extracted from the Technorati.com blog directory, where the in-link blog posts were downloaded from the respective in-link blog sites. The “*search for more reactions*” feature found in Technorati.com was used to extract the in-link to the target blog site. This is followed by downloading of the target blog posts using the target blog URLs extracted from the in-link blog posts with the bloggers identified from the author tags of the postings. Subsequently, the out-link blogs URLs are extracted from the target blog posts, and used to download the blog posts from the out-link blog sites. The *Participating-Listening*, *Broad-Focus*, *Consistent-Casual* type of the Engagement style were analyzed from the target blog posts, while the *Creating-Sharing* type was analyzed between the target blog posts and

the out-link blog posts. The *Persuasion* style is determined based on the subjectivity study on the target blog posts, and the *Persona* style analysis is performed between the in-link blog posts and the target blog posts. The *Participating-Listening* type is measured by the number of blog post titles that contain the target topic and related feature terms. In this study, the scope of the target topics is limited to three main topics: iPad, iPhone, and Mac. The count value is then normalized by the total number of posts for each respective blog site to give the Participating-Listening score. A high percentage of target topic or related feature postings would indicate that the blogger is of Participating type, and conversely, a low percentage shows a Listening type.

Figure 5.4 Influence style analysis framework.

The *Broad-Focused* type is measured by the number of unique target topics and their related features found in the blog post titles. A wide range of topics and their related features discussed would indicate a Broad type, while a limited number of topics and their related features discussed refer to a Focused type. The *Creating-Sharing* type is evaluated through analyzing the similarity between the target blog posts content and its

out-link blog posts content to give the *Creating-Sharing* score based on the Jaccard coefficient defined as the size of the intersection divided by the size of the union of the two linked posts content (A and B) shown as: $J(A_{linking\ blog\ post}, B_{linked\ blog\ post}) = |A \cap B| / |A \cup B|$. A high similarity value would mean that the blogger shares most of the out-link blog post content. On the other hand, the target blog post contains more original content if the similarity value is low. A separate analysis for self-links found in the blog posts is performed to differentiate the *Creating-Sharing* style between self-link and non-self-link posts. Self-links refer to blog posts having links within the same blog site. The overall *Creating-Sharing* type is evaluated by taking the average of the self-linked and non-self-linked *Creating-Sharing* scores. The *Creating-Sharing* style analysis measures the similarity between the target blog posts and *out-link* blog posts to identify the style of the target blogger and differs from the content similarity approaches used in previous studies (Agarwal and Liu, 2009; Matsumura et al., 2008; Song et al., 2007) where the similarity between the target blog posts and the *in-link* blog posts is used to directly measure influence. The *Consistent-Casual* type is determined by the number of target blog site postings in a near-term, mid-term, and long-term duration linked by the in-link posts within the analyzed time period.

Persuasion style is detected through analyzing the subjectivity expressed by the bloggers in their blog posts by counting the number of subjective terms in the target blog posts. This is done by matching the subjectivity terms from Wilson et al. (2005) with the target blog posts terms. The matched number of subjectivity terms is then normalized with the length of the blog post to give the Persuasion Style score. Wiebe et al. (2004) observed that the percentage of opinionated words over total number of word instances for subjective content ranged from 7% to 12%. The 7% benchmark is used as the threshold for subjectivity in the influence style model. Examples of

subjective and objective phrases found in the blog posts are given in Figure 5.5 and Figure 5.6. It can be seen from Figure 5.5, subjective terms such as ‘*major forthcoming*’, ‘*come under fire*’, ‘*lack of*’ and ‘*impressive*’ expressed subjectivity in the post, resulting in the blog post being subjective in persuasion style. On the other hand, the absence of subjective terms in Figure 5.6’s post made the blog post’s style objective and reporting in nature.

*iOS redesigned App Store has leaked a **major forthcoming** update to Twitter’s official iPhone application in its update history section. Twitter in recent months has **come under fire** due to potential new changes to its mobile application API and the **lack of** power-user features in its official mobile device apps. It appears that this controversy may change in the coming days as leaked release notes show some **impressive** new changes.....*

(<http://9to5mac.com/2012/07/08/major-twitter-for-iphone-update-incoming-more-interactive-enhanced-notifications/>). Retrieved on 2 Dec 2012.

Figure 5.5 Example of subjective phrases.

The Electronic Product Environmental Assessment Tool (EPEAT) is a program that ranks computers and displays based on environmental attributes in accordance with IEEE 1680.1-2009. A clear and consistent set of environmental performance criteria for the design of personal computer products including notebook computers desktop computers and computer displays is provided thereby providing an opportunity to secure market recognition for efforts to reduce the environmental impact of electronic products. This standard is intended to provide a tool for government institutional and corporate purchasers. Product manufacturers may also use this tool to earn recognition in the consumer market recognizing that certain criteria may not be applicable to all types of purchasers.....

(<http://9to5mac.com/2012/07/07/apples-new-designs-force-it-to-abandon-epeat-green-standard-potentially-cutting-off-big-customers/>). Retrieved on 2 Dec 2012.

Figure 5.6 Example of objective phrases.

The bloggers’ *Persona* is identified by detecting the sentiments expressed between the in-link and target blog posts. In the influence style model, sentiment analysis rules based on typed dependency rules were adapted from the study in Chapter

4 and (Tan et al., 2012). The typed dependency polarity pattern rules used include the adjectival modifier (AMOD), adverbial modifier (ADVMOD), direct object modifier (DOBJ), and nominal subject (NSUBJ) rules. The typed dependency polarity pattern rule is denoted as “typed-dependency(governor-term:[**polarity**],dependent-term:[**polarity**])→[**polarity**]”. For example, the adjectival modifier pattern rule “AMOD([**neutral**],[+])→[+]” was used for the phrase “*great phone*”: AMOD(phone:[**neutral**],great:[+]) to give a positive polarity value. The typed dependency polarity pattern rules are evaluated using a bottom up approach within a clause. For example, starting from the bottom phrase in the phrase structure tree, as shown in Figure 5.7 for the clause “*I absolutely hate the current state of Facebook for the iPad.*”, the initial noun phrases (NP) (indicated by P1) would give a neutral output as there are no evaluated typed dependency polarity pattern rules. The typed dependencies generated for the clause are shown in Figure 5.8.

Figure 5.7 Phrase structure tree of clause.

<u>Typed dependencies in P1</u>	<u>Typed dependencies in P2</u>	<u>Typed dependencies in P3</u>
prep_of(state, Facebook)	det(state, the)	nsubj(hate, I)
det(iPad, the)	amod(state, current)	advmod(hate, absolutely)
prep_for(Facebook, iPad)		dobj(hate, state)

Figure 5.8 Typed dependencies of clause.

For the subsequent phrase (P2), the adjectival modifier pattern rule “AMOD([**neutral**], [**neutral**])→[**neutral**]” for the phrase “*current state*”: AMOD(state:[**neutral**], current:[**neutral**]) would give a neutral output. In the recursive evaluation, the preceding lower level sentiment polarity output is used as the governor term polarity of the subsequent typed dependency polarity pattern rule. Hence, the governor term “state” in the subsequent direct object pattern rule would inherit the neutral polarity of the preceding rule “AMOD(state:[**neutral**], current: [**neutral**])→[**neutral**]” evaluation output. With this, the direct object pattern rule “DOBJ([-], [**neutral**])→[-]” for the phrase “*hate state*”: DOBJ(hate:[-], state:[**neutral**]) gives a negative polarity sentiment output within the verb phrase indicated by P3. The subsequent adverb modifier pattern rule “ADVMOD([-], [**intensify**])→[**intensified** -]” for the phrase “*absolutely hate*”: ADVMOD(hate:[-], absolutely:[**intensify**]) will also yield a negative polarity. The intensifier term “absolutely” further increase the negative value of the polarity, which is recursively input into the governor term in the nominal subject pattern rule “NSUBJ([-], [**neutral**])→[-]” for the phrase “*I hate*”: NSUBJ(hate:[-], I:[**neutral**]) to give the eventual negative sentiment polarity output for the clause in the final recursive step. The overall blog post's sentiment on the target topic is derived by aggregating the individual clause sentiments for the target topics.

This study analyzes the relationships between the sentiments and target topics by providing a fine-level of syntactic and semantic analysis of the words in the typed dependencies, where each target term is matched against the target topic lexicon. The

syntactic rules used are shown in Table 5.2. The target topic lexicon was created by listing the aspects and features of the target topics from Apple product related websites (e.g., www.apple.com), with examples of the lexicon records shown in Table 5.3.

Table 5.2 Syntactic rules for finding targets of opinions.

ADJECTIVAL MODIFIER Rule: <i>amod(target, opinion)</i> In words: The opinion is an adjectival modifier of the target Example: <i>amod(iPhone, great)</i>
DIRECT OBJECT Rule: <i>dobj(opinion, target)</i> In words: The target is the direct object of the opinion Example: <i>dobj(like, camera)</i>
NOMINAL SUBJECT Rule: <i>nsubj(opinion, target)</i> In words: The target is the subject of the opinion Example: <i>nsubj(iPhone, empowers)</i>

Table 5.3 Examples of target topics and their overall aspects and features.

Target Topics	Overall Aspects	Features
iPad, iPhone, Mac	Connectivity Display Application Operating system	Thunderbolt, Network, Wifi, etc. Touch-Screen, Pixel, Retina, etc. Safari, iTunes, Siri, iCloud, MobileMe, etc. iOS, MacOS, etc.

The target topic scope of the three main topics: iPad, iPhone, and Mac were identified by matching the target terms and feature terms in the lexicon list with terms found in the noun phrases of the blog post sentences parsed using the Stanford Parser. Feature terms are words describing the product’s functionalities and properties, while aspect terms describe the general characteristics of the product. For example, in the clause “*Despite the many advantages of the iPhone’s multi touch-screen, a lack of tactile feedback remains its biggest disadvantage*”, the identified feature “*touch-screen*” was mapped to the display aspect of the iPhone target topic. The predicted triplet output (*target topic, aspect, sentiment polarity*) for the clause would be (“iPhone”,

“Display”, “Negative”). The overall blog post's sentiment on the target topic is derived by aggregating the individual clause's sentiments for the target topics. Compliance could be inferred from the similar sentiments expressed on common target topics and related features between the linked blog posts, even if the sentiments are negative. This can be seen from the linked blog posts that exhibited negative sentiments on the common feature term “*MobileMe*” as shown in Figure 5.9.

Figure 5.9 Example of compliance between linked blog posts.

Figure 5.10 Example of non-compliance between linked blog posts.

On the other hand, high dissimilarities in sentiments show disagreement towards the target blog posts, which in turn indicate the linked bloggers' negative persona. This is

seen in Figure 5.10, where the disagreement is shown in the negative sentiments and positive sentiments expressed by the in-link blog post and the target blog post respectively on the feature term “iOS 6”.

Table 5.4 Threshold values of corresponding influence style scores.

Influence style	Low	Normal	High
Engagement Style			
○ Participating-Listening			
- Listening	15%< <i>s</i> (score)<=20%	20% < <i>s</i> <= 50%	<i>s</i> > 50%
- Participating	10% < <i>s</i> <= 15%	5% < <i>s</i> <= 10%	<i>s</i> <= 5%
○ Creating-Sharing			
- Sharing	15% < <i>s</i> <= 20%	20% < <i>s</i> <= 50%	<i>s</i> > 50%
- Creating	10% < <i>s</i> <= 15%	5% < <i>s</i> <= 10%	<i>s</i> <= 5%
○ Broad-Focused			
- Broad	Comparative Evaluation		
- Focused			
○ Consistent-Casual			
- Consistent	Comparative Evaluation		
- Casual			
Persuasion Style			
○ Subjective	7% < <i>s</i> <= 10%	10% < <i>s</i> <= 12%	<i>s</i> > 12%
○ Objective	5% < <i>s</i> <= 7%	3% < <i>s</i> <= 5%	<i>s</i> <= 3%
Persona Style			
○ Positive	60% < <i>s</i> <= 70%	70% < <i>s</i> <= 80%	<i>s</i> > 80%
○ Controversial		40% < <i>s</i> <= 60%	
○ Negative	30% < <i>s</i> <= 40%	20% < <i>s</i> <= 30%	<i>s</i> <= 20%

In order to differentiate the influence styles of each individual bloggers quantitatively, the threshold values of each influence style type were chosen carefully based on the corresponding scales shown in Table 5.4. The scales were derived from the distribution of scores (s) for each influence type, having the majority score distribution classified into the normal scale. The Participating-Listening score, *Creating-Sharing* score, and Persuasion Style score are used to determine the scoring scale for the respective influence style types. The Persona Style score is based on the

percentage of similar sentiments postings, while the Broad-Focused type and Consistent-Casual type are evaluated through relative comparison amongst the bloggers as the features are specific to the blog sites' context.

5.5 Dataset Used

Two influential blog sites (MacRumors, Technorati Authority = 742 and 9to5Mac, Technorati Authority = 779, retrieved on 1 July 2012) from Technorati.com were identified based on their high Technorati Authority values. A total of 753 and 892 in-link posts dated from 1 July 2012 to 23 July 2012 were extracted for MacRumors and 9to5Mac respectively. From the in-link posts, the target blog site posts with MacRumors having 425 posts and 9to5Mac having 391 posts were identified. These numbers refer only to the linked target blog posts within the analyzed period, and not all the blog sites' posts. Within the target blog posts, 1948 out-link posts for MacRumors and 2297 out-link posts were extracted for 9to5Mac.

Table 5.5 MacRumors bloggers and their respective blog post count.

Blogger Name	Number of target blog post
MacRumors_A	255
MacRumors_B	87
MacRumors_C	78
MacRumors_D	3
MacRumors_E	1
MacRumors_F	1

Further to that, the authors of each blog post with the number of blog posts for the MacRumors bloggers and 9to5Mac bloggers were identified as shown in Table 5.5 and Table 5.6 respectively. 9to5Mac blog site had 10 bloggers as compared to

MacRumors with 6 bloggers. Though there were more bloggers in 9to5Mac, the number of target blog posts linked by the in-link posts was lower compared with MacRumors.

Table 5.6 9to5Mac bloggers and their respective blog post count.

Blogger	Number of target blog post
9to5Mac_A	105
9to5Mac_B	88
9to5Mac_C	60
9to5Mac_D	43
9to5Mac_E	39
9to5Mac_F	25
9to5Mac_G	17
9to5Mac_H	11
9to5Mac_I	2
9to5Mac_J	1

A validation of the methods was performed to verify the performance of the influence style model and the threshold values used. Approximately 12% (100 records out of 816) of the total target blog post records were extracted for the inter-coder reliability testing in the *Persona* evaluation, with a balanced distribution for the positive and negative tagged samples. A further 100 records were used for the subjectivity inter-coder reliability testing, where the samples were distributed evenly between the subjective and objective thresholds. A sample of 100 records (Approximately 2.5% of the total number of out-links) was used in the reliability testing for the *Creating-Sharing* evaluation with a balanced distribution between creating records and sharing records. The threshold values used in the validation are *Creating-Sharing* score (15%), Persuasion style score (7%), and Persona score (Positive>60%, Negative<=40%). The computed Kappa values for the two person coders' inter-reliability testing are as follow: *Creating-Sharing* evaluation (Kappa value=0.97), *Persuasion* evaluation (Kappa value=0.69), and *Persona* evaluation

(Kappa value=0.61), which are in the acceptable range (Cohen, 1960). The conflicting tags by the two coders were reviewed and manually re-classified and used as the answer keys.

5.6 Evaluation of Influence Style Model

An evaluation of the influence style model was performed in this study, where the F1 score defined as $(F1\ score = (2 * Precision * Recall) / (Precision + Recall))$ was used to measure the performance of the influence style model in relation to the precision and recall. F1 scores for both Sharing type (90.9%) and Creating type (90.1%) classification are high. This shows the *Creating-Sharing* type predictions made by the model to be reliable as validated by the coders. On the other hand, F1 scores for Persuasion evaluation are relatively lower (Objective type F1 score=7.1%, Subjective type F1 score=85.4%). The threshold value (7%) for Objective type classification was too stringent resulting in the poor performance. Instead, a threshold value of 10% improved the influence style model's performance (Objective type F1 score=65.8%, Subjective type F1 score=78.7%), while a threshold value of 12% gave F1 scores of Objective type=36.9%, and Subjective type=49.1%. The results also show that persuasion style could vary even within individual blogger posts as seen from the variation between the maximum, minimum, and average subjective scores of the bloggers. The F1 scores for positive Persona (74.6%) are higher than the scores for negative Persona (54.1%), because of the assumption that neutrality is considered as agreement, resulting in the higher scores for predicting similar sentiments between blog posts. Performance of the Persona evaluation is dependable on the sentiment

analysis. Therefore, any improvement in the sentiment analysis process would result in better performance for the Persona evaluation.

MacRumors Bloggers Analysis

Engagement Style: Participating-Listening Type

The Participating-Listening value was computed by counting the number of post titles that contained the target topic terms or related aspect terms, and normalized by the total number of target blog posts in each blog site. The aspect terms found in the blog posts were collated to the target topics to provide the aggregated result at target topic level. As seen from Table 5.7, the MacRumors bloggers had high participation rate (greater than 90%) on the target topic terms with respect to the other topics postings.

Table 5.7 Participating-Listening values.

Blogger	Participating-Listening values of target topics				Other topics
	iPad	iPhone	Mac	Total	
MacRumors_C	51(32.7%)	64(41.0%)	27 (17.3%)	142(91.0%)	14(9.0%)
MacRumors_D	7(41.2%)	7(41.2%)	2(11.8%)	16(94.1%)	1(5.88%)
MacRumors_A	231(34.4%)	238(35.5%)	176(26.2%)	645(96.1%)	26(3.9%)
MacRumors_B	82(40.4%)	80(39.4%)	34(16.7%)	196(96.5%)	7(3.5%)
MacRumors_E	0(0.0%)	0(0.0%)	1(100.0%)	1(100.0%)	0(0.0%)
MacRumors_F	1(50.0%)	1(50.0%)	0(0.0%)	2(100.0%)	0(0.0%)

This shows that the MacRumors bloggers were active participants towards Apple products related discussions. However, there was a slight deviation to other topics discussion by MacRumors_C (other topics' Participating-Listening score = 9.0%) and MacRumors_D (other topics' Participating-Listening score = 5.88%), indicating more participation in other non-Apple related products. This study profile MacRumors_C and MacRumors_D as normal participating, with the rest of the bloggers

(MacRumors_A, MacRumors_B, MacRumors_E, and MacRumors_F) as high participating type based on their other topics' Participating-Listening scores being less than 5%. Though MacRumors_E and MacRumors_F had only few postings, their postings were specific to Apple related products leading to a low other topics' Participating-Listening score and subsequent high participating profile.

Engagement Style: Broad-Focused Type

The Broad-Focused type was evaluated by analyzing the unique target topics and the related features contained in the blog post titles for the MacRumors bloggers. MacRumors_A, MacRumors_B, and MacRumors_C were wide in their scope by covering comprehensively the target topics and related aspect features pertaining to Apple brand products. On the other hand, MacRumors_D was restricted in the scope of topics discussed. Only iPhone, iPad, Siri, iOS, iTunes, iCloud, and Mac topics were mentioned in MacRumors_D's postings. Further to that, MacRumors_E had a limited posting on the Mac topic, while MacRumors_F discussed only on applications pertaining to iPhone and iPad, indicating a focused or limited scope of topics. Nonetheless, the coverage of target topics and aspect features may be limited as the target topic and related aspect-feature list were manually created based on prior knowledge of the Apple brand product domain. Performing automatic topic discovery in future work would improve the coverage of target topics, aspects, and features.

Engagement Style: Creating-Sharing Type

The *Creating-Sharing* type was analyzed through the similarity between the target blog posts content and its out-link posts content based on the Jaccard coefficient. Table 5.8 shows the maximum, minimum, and average Jaccard coefficient values with

computed standard deviations. MacRumors_D had a low average self-links engagement style score (4.1%) indicating low content similarity between the self-linked posts. This meant that the blog post content was highly original even within self-linked posts, which shows a high creating style. The other MacRumors bloggers had *Creating-Sharing* scores between 5% and 10%, which classified them as normal creating type. It is noted that MacRumors bloggers were creative and original in its blog posts content even between self-linked content as seen from the similar engagement style values between the self-linked posts and non-self-linked posts.

Table 5.8 Engagement style (Jaccard coefficient) scores.

Blogger	Self-links (%)			Non-self-links (%)		
	Max	Min	Avg (Std dev)	Max	Min	Avg (Std dev)
MacRumors_E	15.4	0.7	9.1 (6.6)	NIL	NIL	NIL
MacRumors_C	79.5	0.2	8.2 (8.2)	19.3	1.2	7.8 (4.0)
MacRumors_B	96.1	0.3	7.9 (6.1)	15.3	0.3	7.2 (3.9)
MacRumors_A	98.1	0.3	7.8 (5.8)	16.4	0.5	7.6 (4.1)
MacRumors_F	6.5	6.5	6.5 (0.0)	7.5	7.5	7.5 (0.0)
MacRumors_D	14.8	0.5	4.1 (4.6)	NIL	NIL	NIL

Engagement Style: Consistent-Casual Type

The *Consistent-Casual* type measures the consistency of the target blog site to garner in-links over a period of time with respect to the number of target blog posts that are linked by the in-link posts within the analyzed period. Through registering the dates for the linked target blog posts, the blog posts were grouped according to the near-term (Jan 2012 to Jul 2012), mid-term (Jan 2011 to Dec 2011), and long-term (until year 2010) time period. Figure 5.11 and Figure 5.12 show a consistent number of links between the in-link blog posts and target blog posts for MacRumors_A,

MacRumors_B, and MacRumors_C in the near-term and mid-term duration respectively.

Figure 5.11 Near-term linked posts count of top 3 MacRumors bloggers.

Figure 5.12 Mid-term linked posts count of top 3 MacRumors bloggers.

Similarly, there were links found in the long-term duration for the three top bloggers, indicating a consistent style. On the other hand, MacRumors_D had two links in the near-term duration, and one link in the long-term duration. Both MacRumors_E and MacRumors_F had only one link in the long-term duration. The results show MacRumors_D, MacRumors_E, and MacRumors_F to be casual bloggers with inconsistent link history.

Persuasion Style

The persuasion style was measured based on the amount of subjectivity expressed in the target blog post content. From Table 5.9, MacRumors_A, MacRumors_B, MacRumors_C, MacRumors_D, and MacRumors_F were highly subjective in persuasion with their persuasion style scores greater than 12%. On the other hand, MacRumors_E had a mildly objective persuasion style as seen from the low persuasion style score of 6.9%. From the high subjectivity values, it could be inferred that MacRumors bloggers tend to be subjective rather than objective in their persuasion style, which indicates MacRumors to be a subjective product review blog site instead of being objective and reporting in style.

Table 5.9 Subjectivity values

Logger	Subjectivity values (%)		
	Max	Min	Avg (Std dev)
MacRumors_D	15.6	13.8	15.2 (0.5)
MacRumors_F	13.7	13.7	13.7 (0.0)
MacRumors_B	13.9	10.6	12.7 (0.8)
MacRumors_C	14.2	9.2	12.5 (1.2)
MacRumors_A	15.7	10.5	12.2 (1.5)
MacRumors_E	6.9	6.9	6.9 (0.0)

Persona Style

In the persona analysis, the similar sentiments and dissimilar sentiments ratios between the in-link blog posts and the target blog posts towards the respective target topics were evaluated. From Table 5.10, it can be seen that MacRumors_A, MacRumors_C, MacRumors_F had Persona Style scores greater than 80% indicating a high positive persona style. The other MacRumors blogger MacRumors_B was also of a normal positive persona style, inferring that MacRumors was a blog site that was generally positively received by its readers. However, MacRumors_D had a Persona Style score of 56.7% showing a controversial style within the positively received MacRumors blog site. This means that MacRumors_D blog posts readers expressed both agreement and disagreement on her postings. The evaluation result shows that bloggers from within a blog site could differ in their persona style. The persona style evaluation is not applicable for MacRumors_E because none of the blog postings were related to the target topics.

Table 5.10 Similar sentiment and dissimilar sentiment values of MacRumors bloggers.

	Target Topic							
	iPad		iPhone		Mac		Average	
	Similar Senti-ment	Dissi-milar Senti-ment	Similar Senti-ment	Dissi-milar Senti-ment	Similar Senti-ment	Dissimi-lar Senti-ment	Similar Senti-ment	Dissimi-lar Senti-ment
MacRumors_F	NIL	NIL	100.0%	0.0%	NIL	NIL	100.0%	0.0%
MacRumors_A	82.4%	17.6%	83.6%	16.4%	91.0%	89.0%	89.3%	10.7%
MacRumors_C	96.2%	3.8%	85.2%	14.8%	81.0%	19.0%	87.5%	12.5%
MacRumors_B	81.8%	18.2%	76.5%	23.5%	77.1%	22.9%	78.5%	21.5%
MacRumors_D	100.0%	0.0%	20.0%	80.0%	50.0%	50.0%	56.7%	43.3%
MacRumors_E	NIL	NIL	NIL	NIL	NIL	NIL	NIL	NIL

Macrumors Bloggers' Influence Profiles

Table 5.11 shows the evaluated profiles of the Macrumors bloggers. The results show the MacRumors bloggers to be participating and focused on the target topics as evaluated by the influence style model.

Table 5.11 Profiles of MacRumors bloggers influence.

Blogger	Engagement style				Persuasion style	Persona style
	Participating-Listening type	Broad-Focused type	Creating-Sharing type	Consistent-Casual type		
MacRumors_A	High Participating	Focused	Creating	High Consistent	High Subjective	High Positive
MacRumors_B	High Participating	Focused	Creating	Consistent	High Subjective	Positive
MacRumors_C	Participating	Focused	Creating	High Consistent	High Subjective	High Positive
MacRumors_D	Participating	Low Focused	High Creating	Casual	High Subjective	Controversial
MacRumors_E	High Participating	Low Focused	Creating	Low Casual	Low Objective	N.A.
MacRumors_F	High Participating	Low Focused	Creating	Low Casual	High Subjective	High Positive

However, there were varying levels of participation and focus amongst the bloggers. In addition, the bloggers were evaluated to be creating in style, with MacRumors_D being more creative and original in blog posts content. MacRumors_A, MacRumors_B, and MacRumors_C had consistent influence over the duration of time compared to the other bloggers who were casual in their influence style. The Macrumors bloggers expressed high degree of subjectivity in their blog posts, except for MacRumors_E who is objective and reporting in nature. Likewise, the bloggers are positive in their persona, while MacRumors_D exhibits a controversial style. These differences in influence style show that bloggers within a common blog site could still exhibit

varying style of influence, reflecting the need for further in-depth analysis to differentiate the bloggers' influence style. By providing an in-depth analysis on the bloggers' influence style, the bloggers' influence could be further described in detail.

9to5Mac Bloggers Analysis

The 9to5Mac blog site had 10 bloggers as compared to MacRumors with 6 bloggers, where the greater number of individual bloggers in 9to5Mac would create more diversity in influence style amongst the bloggers. This is seen from the Participating-Listening evaluation results in Table 5.12, where the high Participating-Listening scores of 9to5Mac_J and 9to5Mac_D, and varying low to high scores of the other bloggers indicate differing engagement styles.

Table 5.12 Participating-Listening values.

Blogger	Participating-Listening values of target topics				Other topics
	iPad	iPhone	Mac	Total	
9to5Mac_J	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1(100.0%)
9to5Mac_D	27 (36.5%)	29 (39.1%)	6 (8.1%)	62 (83.7%)	12(16.3%)
9to5Mac_B	88 (40.9%)	77 (35.8%)	21 (9.7%)	186 (86.4%)	29(13.6%)
9to5Mac_E	39 (42.4%)	39 (42.4%)	7 (7.6%)	85 (92.4%)	7 (7.6%)
9to5Mac_C	31 (32.6%)	45 (47.3%)	13 (13.6%)	89 (93.5%)	6 (6.5%)
9to5Mac_H	13 (37.1%)	17 (48.5%)	3 (8.5%)	33 (94.1%)	2 (5.9%)
9to5Mac_A	114 (40.1%)	126 (44.3%)	30 (10.5%)	270 (94.9%)	14 (5.1%)
9to5Mac_F	18 (40.0%)	21 (46.7%)	4 (8.8%)	43 (95.5%)	2 (4.5%)
9to5Mac_G	13 (36.1%)	14 (38.9%)	9 (25.0%)	36 (100.0%)	0 (0.0%)
9to5Mac_I	4 (44.4%)	4 (44.4%)	1 (11.1%)	9 (100.0%)	0 (0.0%)

In addition for the Broad-Focused type evaluation, 9to5Mac_A, 9to5Mac_B, 9to5Mac_C, and 9to5Mac_E were evaluated to be loosely focused on the Apple related target topics with more other non-target topics discussed, while the other 9to5Mac bloggers were restricted to the target topics range in their postings. From Table 5.13,

the duplication of content from non-self-link posts was minimal for the 9to5Mac bloggers, possibly due to copyright constraints on usage of other sites' content.

Table 5.13 Creating-Sharing (Jaccard coefficient) scores.

Blogger	Self-links			Non-self-links		
	Max	Min	Avg (Std dev)	Max	Min	Avg (Std dev)
9to5Mac_D	98.6%	0.4%	18.9% (28.5%)	93.6%	0.4%	10.6% (15.9%)
9to5Mac_B	96.3%	0.4%	32.7% (36.1%)	85.4%	0.7%	9.5% (6.7%)
9to5Mac_C	97.9%	0.2%	28.8% (29.2%)	93.8%	0.1%	8.5% (7.3%)
9to5Mac_E	95.4%	0.7%	32.2% (31.6%)	21.0%	0.6%	7.9% (3.4%)
9to5Mac_A	98.1%	0.5%	32.6% (33.5%)	85.9%	0.8%	7.8% (5.9%)
9to5Mac_I	95.3%	13.5%	67.3% (46.5%)	84.0%	3.2%	7.2% (9.6%)
9to5Mac_F	95.3%	1.9%	31.8% (31.0%)	39.7%	1.4%	7.2% (4.3%)
9to5Mac_H	98.0%	0.7%	31.5% (32.8%)	13.3%	2.3%	6.8% (2.2%)
9to5Mac_J	NIL	NIL	NIL	8.3%	4.0%	6.8% (1.8%)
9to5Mac_G	96.9%	0.5%	43.5% (43.0%)	12.1%	1.2%	6.0% (2.0%)

However, content similarity between self-links blog posts was high with 9to5Mac_I and 9to5Mac_G having *Creating-Sharing* scores greater than 40%. This shows that though 9to5Mac bloggers do observe the copyrights of other blogs' content, the tendency is to share their own blog content through the linked blog posts, indicating a sharing type in the influence style. The variations in the *Creating-Sharing* style amongst the bloggers are seen in the wide range of average similarity values (9to5Mac_I=67.3% to 9to5Mac_G=6.0%) between the 9to5Mac bloggers. It is noted that the similarity measure is affected by writing style as seen in short blog posts which tended to give higher Jaccard Coefficient Scores due to the higher similarity between short texts. In the Consistent-Causal type evaluation, it is observed that 9to5Mac_A, 9to5Mac_B, 9to5Mac_C, 9to5Mac_D, and 9to5Mac_E had high number of target posts linked by in-link posts in the near-term time period as seen in Figure 5.13. However, only 9to5Mac_A, 9to5Mac_B, and 9to5Mac_E had limited number of

links (less than 6 for each blogger) in the mid-term duration. In the long-term duration, only 9to5Mac_A and 9to5Mac_F had one linked posting each. This shows that the influence consistency of the 9to5Mac bloggers is limited and differing across the bloggers, for example with 9to5Mac_A being more consistent than 9to5Mac_J, who had limited number of linked posts in the near-term duration.

Figure 5.13 Near-term linked posts count of 9to5Mac bloggers.

The persuasion style scores of the 9to5Mac bloggers are within the range of 10% to 12%, reflecting a subjective persuasion style type for the bloggers with similar level of subjectivity expressed in their blog posts. Only 9to5Mac_D and 9to5Mac_E had a highly subjective persuasion style with a score greater than 12%. Likewise, the 9to5Mac bloggers were positive in their Persona style, with the Persona style scores well above 70%, except for 9to5Mac_E who is controversial in style, and 9to5Mac_J, who did not post on the related target topics.

From Table 5.14, the influence style profiles of the 9to5Mac bloggers show a participating and focused style with the exception of 9to5Mac_D who is less active in the target related postings.

Table 5.14 Profiles of 9to5Mac bloggers influence.

Blogger	Engagement style				Persuasion style	Persona style
	Participating-Listening type	Broad-Focused type	Creating-Sharing type	Consistent-Casual type		
9to5Mac_A	High Participating	Low Focused	Sharing	Consistent	Subjective	Positive
9to5Mac_B	Low Participating	Low Focused	Sharing	Consistent	Subjective	High Positive
9to5Mac_C	Participating	Low Focused	Low Sharing	Low Consistent	Subjective	Positive
9to5Mac_D	Low Listening	Focused	Low Creating	Casual	High Subjective	Positive
9to5Mac_E	Participating	Low Focused	Sharing	Consistent	High Subjective	Controversial
9to5Mac_F	High Participating	Focused	Low Sharing	Casual	Subjective	High Positive
9to5Mac_G	High Participating	High Focused	Sharing	Casual	Subjective	High Positive
9to5Mac_H	Participating	Focused	Low Sharing	High Casual	Subjective	High Positive
9to5Mac_I	High Participating	Focused	Sharing	High Casual	Subjective	High Positive
9to5Mac_J	High Listening	High Focused	Creating	High Casual	Subjective	N.A.

The 9to5Mac bloggers, other than 9to5Mac_J, were sharing rather than creating in style. There were differing degrees of consistency amongst the bloggers with slightly more than half of them are casual in their influence style. On the whole, the 9to5Mac bloggers were positive in persona style, indicating general consensus from their readers, and they were also subjective in their persuasion style. In general, bloggers from popular sites have positive persona where in-link bloggers express agreement towards them. Nevertheless, there were bloggers who displayed different

persona from the other bloggers as seen in MacRumors_D from Macrumors, and 9to5Mac_E from 9to5Mac who had controversial persona styles, where they received both agreeing and disagreeing statements from other bloggers.

5.7 Discussion

This study proposed a novel approach in analyzing blog posts at the blogger-level to give an in-depth analysis of the individual blogger's influence style. Manual observations and description of the blog sites in Technorati.com revealed both MacRumors and 9to5Mac to be specialized blog sites focusing on Apple brand-related products, which are indicated by the Participating and Focused types of their bloggers in the evaluation results. It is observed that individual bloggers could differ from the overall blog site and other bloggers in influence style. For example, the overall MacRumors blog site was highly participating towards the main target topics, while individually MacRumors_C and MacRumors_D were less participating on the main target topics, with a more diverse range of topics discussed. Similarly, the 9to5Mac bloggers had differing Participating and Focus type levels. The level of participation and focus on topics could also differ between the individual bloggers. Being able to differentiate the participation level and topic focus of individual bloggers allows blog readers to determine the bloggers' depth and scope of knowledge in respective topics. Though bloggers could differ in influence style, there are certain rules, such as the non-infringement of copyrights commonly observed by bloggers resulting in a similar *Creating-Sharing* influence style for the bloggers, especially within non-self-link posts. Influence consistency is also observed to differ between bloggers. For example, MacRumors_D, MacRumors_F, and MacRumors_E were casual in their influence

style compared with the other bloggers from MacRumors, while 9to5Mac_E, 9to5Mac_B, and 9to5Mac_A were more consistent in the more casual 9to5Mac blog site. Both the MacRumors and 9to5Mac bloggers, except for MacRumors_E were evaluated to be subjective in their persuasion style based on their subjectivity score. The homogeneity in persuasion style is indicative of a product review blog site, which is generally subjective in nature as compared to an objective news reporting blog site. It is noted that the bloggers studied came from multi-blogger sites, and it is possible that they were governed or responding to a “house style” which might increase similarities within a site. Future work could examine the influence style differences within a blog site and across different blog sites.

This study attempted to derive deterministic thresholds for the scores to provide an objective measurement to the respective influence style evaluation. Through this, a detailed breakdown of the bloggers’ influence styles could be provided. However, a deterministic threshold approach is subjective to the data used and could not be generalized across other domains and content. Subsequent study in Chapter 6 describes the influence styles relative to each blogger through comparing the influence style values amongst the bloggers. Influence style of a blogger is time dependent, where the bloggers’ influence on specific topics is based on the analyzed period as seen from the *Consistent-Casual* influence style result. A different period of analysis could have identified the influence style differently since the measurement values would have changed. In this study, the target topic and related aspects list were manually created based on prior knowledge of the Apple-product domain. Future work could explore automatic topic discovery to detect a more comprehensive range of topics discussed in the blog content.

The results from this study show that influence style is specific to individual bloggers, and the blog site influence style is not representative of all bloggers. By analyzing the influence style of individual bloggers, this study was able to provide a fine-grained description of the bloggers' influence. The knowledge on individual blogger's influence style would be used in the subsequent study in Chapter 6 to detect influence diffusion between linked bloggers. For example, the Participating, Focused, Creating, Consistent, Low Subjective, and Highly Positive influence style of MacRumors_C could be used as features to predict influence propagation.

5.8 Summary

In contrast to previous studies which used a monolithic definition of influence for blogs, the study in this chapter further describes the influence of bloggers in terms of engagement, persuasion, and persona style in applying research method (RM3-i: *Determine engagement style, persuasion style, and persona of bloggers*) to meet the partial requirements of the research objective (RO3: *Develop an influence diffusion detection model using sentiment values between linking and linked bloggers, agreement towards the linked bloggers, and bloggers' influence styles as possible features to improve influence diffusion detection between linked bloggers*). The detailed analysis of influence could differentiate the influence style of bloggers and provide a more comprehensive description of the influence exerted. Evaluation results show there are differences in influence style between individual bloggers and the overall blog site as well as amongst the bloggers. This indicates a need to analyze individual blogger's influence style to provide further in-sights into the bloggers' influence profile. The provision of the bloggers' influence style would enhance the

influential blog search experience through the detailed description of the manner in which the bloggers exert influence. Further to that, a more comprehensive and accurate portrayal of the blog influence through the evaluated influence profiles would in turn improve influence diffusion detection between linked bloggers.

CHAPTER 6

INFLUENCE DIFFUSION DETECTION BETWEEN LINKED BLOGGERS

6.1 Introduction

The study in this chapter aims to meet the research objective (RO3: *Develop an influence diffusion detection model using sentiment values between linking and linked bloggers, agreement towards the linked bloggers, and bloggers' influence styles as possible features to improve influence diffusion detection between linked bloggers*) through applying research method (RM3-ii: *Develop the influence diffusion detection model based on the Independent Cascade model and Linear Threshold model using similar sentiments on common topics between the linked and linking bloggers, agreement towards the linked blogger, and influence styles as features*) by developing the Influence Diffusion Detection (IDD) model to detect influence flow between linked bloggers. Influence is described through the engagement style, persuasion style, and persona of the bloggers based on the study in Chapter 5.

Different variants of the IDD model: Influence Diffusion Detection-Influence Style-Sentiments Independent Cascade (IDD-IS-Sent) (IC), Influence Diffusion Detection-Influence Style-Sentiments-Agreement Independent Cascade (IDD-IS-Sent-Agree) (IC), Influence Diffusion Detection-Influence Style-Sentiments Linear Threshold (IDD-IS-Sent) (LT), and Influence Diffusion Detection-Influence Style-Sentiments-Agreement Linear Threshold (IDD-IS-Sent-Agree) (LT) using similar sentiment values, agreement values, and influence styles as features were developed and tested, with both the Independent Cascade (IC) model and the Linear Threshold

(LT) model evaluated as possible approaches in this study. The IDD model was built based on a real-life blog posts dataset using the influence styles of the bloggers. The results showed the IDD model performed well in detecting influence flow within a network of bloggers with average F1-scores at 76%.

The next section provides a background of the study, with details of the IDD model given in Section 6.3, where the influence styles score measurements and the influence diffusion detection methodology are illustrated. Section 6.4 describes the dataset used in the study, while the model evaluation process and results are given in Section 6.5. This is followed by the discussion on the results and summary of study in Section 6.6 and Section 6.7 respectively.

6.2 Background

Previous studies had proposed using blog features (Adar and Adamic, 2005; Agarwal and Liu, 2008), similarity comparison (Matsumura et al., 2008; Song et al., 2007), and community detection (Agarwal et al., 2008; Ghosh and Lerman, 2008) to detect blog influence. However, the studies did not consider the subject matter of the blog post and thus, would not measure influence accurately as influence is subjective in nature and is often dependent on the contextual details. This is seen in Figure 6.1 where in-links to blog post A are used to indicate the presence of influence flow from A to the other blog posts. But a blog post with many in-links may not necessarily indicate the blogger is influential as the linking posts could contain disagreement towards the blogger as shown in Figure 6.2 where the negative sentiment expressed between blog post A and B shows non-influence flow despite the existence of the link between them.

Figure 6.1 Link frequency approach.

Figure 6.2 Sentiment links approach.

Figure 6.3 Sentiment links and influence style approach.

Chapter three of this study found that the sentiment values between the linking and linked blog posts are significantly correlated to the influence of the linked bloggers. From here, the study in Chapter 4 attempted to improve sentiment prediction performance through applying the typed dependencies polarity rules and subjective phrase lexicon to handle the complex relationship between words in a phrase. Recent studies attempted to detect influence between blog posts by analyzing the sentiments expressed in the posts (Leskovec et al., 2010; Cai et al., 2011; Li et al., 2011). These studies focused on a single notion that influence exists between blogs with positive sentiments, but ignored the details in other aspects of influence. In addition to the sentiment feature used in previous studies, this study further explore the use of influence styles as additional features in detecting influence diffusion. Influence is a complex concept that could be further described through the manner in which bloggers actively engage readers, or how bloggers express ideas in an objective or subjective manner, or whether bloggers' posts are received positively or negatively by the readers. This can be seen in Figure 6.3 where the author of blog post A is described to be participating, focused, creating, and consistent in engaging the readers, objective in persuasion style, and having a persona that is positively received by readers. With the

identified influence styles, the manner in which influence flow occurs between the blogs could be further described.

Directed graphs were used in previous studies (Kempe et al., 2003; Yang and Leskovec, 2010; Lim et al., 2011) to model the spread of an idea or innovation through a network. Each individual node in the graph is either active (indicating adaptation of the innovation) or inactive. The Linear Threshold model (Granovetter, 1978) and the Independent Cascade model (Goldenberg et al., 2001) are both fundamental models for the diffusion of information through a network. The Linear Threshold approach is used to study the use of influence styles in detecting influence diffusion that is initiated from all the active neighbors of a blogger versus the Independent Cascade approach, which only considers influence diffusion from a single neighbor at each step of the evaluation process. The Independent Cascade model provides specific influence analysis between two linked bloggers to provide a more precise evaluation, while the Linear Threshold model conceptualizes a situation where a blogger is linked to multiple bloggers and could be influenced simultaneously by all neighboring bloggers instead of only one blogger. This study aims to evaluate the performance between the Independent Cascade approach and the Linear Threshold Model through the use of influence styles and sentiment links as features in detecting influence diffusion between linked bloggers.

6.3 Influence Diffusion Detection Model

In the proposed IDD model, influence is further described in terms of the bloggers' *Engagement style*, *Persuasion style*, and *Persona style* based on the study done in Chapter 5. An overview of the influence style analysis methods is summarized in

Table 6.1, and the influence style analysis framework adapted from the study in Chapter 5 is shown in Figure 6.4.

Table 6.1 Influence style analysis methods.

Influence Style	Type	Description
Engagement	Listening-Participating	Describes the blogs' or bloggers' participation level in posting specific target topics.
	Broad-Focused	Describes the scope of the topics discussed in the blog posts.
	Creating - Sharing	Describes the level of blog posts' content originality.
	Consistent-Casual	Describes the consistency and duration of blogs' or bloggers' ability to garner in-links.
Persuasion	Subjective-Objective	Describes the types of Persuasion style
Persona	Positive-Negative-Controversial	Describe the types of Persona.

Figure 6.4 Influence style analysis framework.

The in-link blog posts URLs were extracted from Technorati.com, with the in-link blog post content downloaded from the respective blog sites. This is followed by downloading of the target blog posts using the target blog URLs extracted from the in-link blog posts. Subsequently, the out-link blogs URLs are extracted from the target

blog posts, and used to download the blog posts from the out-link blog sites. The *Participating-Listening*, *Broad-Focus*, *Consistent-Casual* type of the Engagement style are analyzed from the target blog posts, while the *Creating-Sharing* type is analyzed between the target blog posts and the out-link blog posts. The *Persuasion* style is determined based on the subjectivity study on the target blog posts, and the *Persona* style analysis is performed between the in-link blog posts and the target blog posts.

6.3.1 Influence Styles Score Measurement

The influence styles score measurement methods were adopted from the study in Chapter 5 and (Tan et al., 2013), with changes to the *Broad-Focused (BF)* and *Consistent-Casual (CC)* influence style type measurements made, where the normalized scores of each influence style type are further measured for computation in the IDD model.

Engagement Style Analysis. The *Participating-Listening (PL)* type is evaluated as the active participation of the bloggers in specific target topics, and is measured by the number of blog post titles that contain the target topic and related feature terms. In this study, the scope of the target topics is limited to Apple-related products. The count value is then normalized by the total number of posts for each respective blogger to give the *PL-score*. A high percentage of target topic or related feature postings would indicate that the target blog site is of the *Participating* type, and conversely, a low percentage shows a *Listening* type. The *Broad-Focused (BF)* type represents the scope of the topics discussed within the blog site. A wide range of topics and their related features discussed would indicate a *Broad* type, while a limited number of topics and their related features discussed refer to a *Focused* type. The *BF-score* is measured by

the number of unique target topics and their related features found in the blog post titles and normalized by the total number of topics in the network. The *Creating-Sharing (CS)* type measures the originality of the content posted by the blog site, and is evaluated through analyzing the similarity between the target blog posts content and its out-link blog posts content based on the Jaccard coefficient. The Jaccard coefficient is defined as the size of the intersection divided by the size of the union of the two linked posts content (A and B) as shown in Equation 6.1:

$$J(A_{(linking\ blog\ post)}, B_{(linked\ blog\ post)}) = (|A \cap B|) / (|A \cup B|) \quad (\text{Eq. 6.1})$$

A high similarity value would mean that the target blog post shares most of the out-link blog post content. On the other hand, the target blog post contains more original content if the similarity value is low. The *CS-score* is given as (1 minus Jaccard coefficient) to reflect the propensity to influence for bloggers who are creative and original in their content. The *Consistent-Casual* type indicates the consistency and duration of the target blog post's ability to garner in-links, and is determined by the number of target blog site postings in an arbitrarily chosen near-term (within 1 month), mid-term (between 1 to 3 months), and long-term (beyond 3 months) duration linked by the in-link posts within the analyzed time period. The *CC-score* is computed by assigning a nil score for near-term posts, a score of 0.5 for mid-term posts, and a score of 1.0 for long-term posts. This gives consideration for the consistency of the blogger to harness links for its mid and long term postings. The assumption is that bloggers with links to their long-term posts tend to be more consistent in their influence. The assigned scores are summed and normalized over the number of posts for each blogger.

Persuasion (PES) Style Analysis. The subjectivity expressed by the bloggers in their blog posts is analyzed through counting the number of subjective terms in the

target blog posts. This is done by matching the subjectivity terms from Wilson et al. (2005) with the target blog posts terms. The matched number of subjectivity terms is then normalized with the length of the blog post. The *PES-score* is derived from taking the average score of the blogger’s total postings.

Persona (PER) Style Analysis. The blogger’s persona is identified by detecting the sentiments expressed between the in-link and target blog posts. It is possible to analyze comments and feedbacks as an observation of readers’ response to the blog posts. However, the scope of this study is limited to detecting only the influence styles between linked blog posts. This study limits the detection of persona style through analyzing the sentiments on common topics between the linking and linked blog posts as there are not many explicitly mentioned agreement or disagreement cases between linked blog posts. Nevertheless, detecting agreement towards the linked bloggers could have improved the analysis on the persona style of bloggers. In the IDD model, the sentiment analysis rules were adapted from the study in Chapter 4 with the *PER-score* of each blogger given as the ratio of number of similar sentiments posts over the total number of posts for each blogger. Sentiment analysis is performed on clauses at the aspect level of the target topics using the typed dependency polarity pattern rules. This is because aspect level analysis is shown to yield better performance (Thet et al., 2010). The typed dependency polarity pattern rule is denoted as “*typed-dependency(governor-term:[polarity], dependent-term:[polarity])→[polarity]*”, for example in the adjectival modifier pattern rule “AMOD(phone:[**neutral**], great:[+])→[+]” yields a positive sentiment polarity for the phrase “great phone”. The other typed dependency polarity pattern rules used include the adverbial modifier (ADVMOD), direct object modifier (DOBJ), and nominal subject (NSUBJ) rules. The subjectivity lexicon from Wilson et al. (2005) is used to tag the prior polarity of the governor term and dependent term.

Table 6.2 Examples of target topics, aspects and features.

Target	Overall Aspects	Features
iPad, iPhone, Mac	Connectivity	Thunderbolt, Network, Wifi, etc.
	Display	Touch-Screen, Pixel, Retina, etc.
	Application	Safari, iTunes, Siri, iCloud, MobileMe, etc.
	Operating system	iOS, MacOS, etc

Table 6.3 Summary of influence styles score measurement methods.

Influence Style	Type	Measurement Method
Engagement	Participating-Listening (PL) score	Measure the number of blog post titles that contain the target topic and related feature terms. The count value is then normalized by the total number of posts for each respective blogger to give the <i>PL-score</i> .
	Broad-Focused (BF) score	The number of unique target topics and their related features found in the blog post titles is normalized by the total number of topics in the network to give the <i>BF-score</i> .
	Creating-Sharing (CS) score	Similarity between linked blog posts is measured based on the Jaccard coefficient. The <i>CS-score</i> is given as (1 minus Jaccard coefficient).
	Consistent-Casual (CC) score	<i>CC-score</i> is computed by assigning a nil score for near-term posts, a score of 0.5 for mid-term posts, and a score of 1.0 for long-term posts.
Persuasion	Subjective-Objective (PES-score)	The matched number of subjectivity terms is normalized with the length of the blog post to give each posts subjectivity value. The <i>PES-score</i> is derived from taking the average score of the blogger's total posts subjectivity values.
Persona	Positive-Negative-Controversial (PER-score)	<i>PER-score</i> of each blogger is given as the ratio of number of similar sentiments posts over the total number of posts for each blogger.

A lexicon of specific target topics with the related aspects and features was manually created for the identification of target topics with examples shown in Table 6.2 as well as Appendix A. The target topics were identified by matching the target terms and

feature terms in the lexicon list with terms found in the noun phrases of the sentence. For example, the identified “*Facebook*” feature for the clause “*I absolutely hate the current state of Facebook for the iPad.*” is mapped to the application aspect of the “*iPad*” target topic, which gives the predicted triplet output (*target topic, aspect, sentiment polarity*) for the clause as (“*iPad*”, “*Application*”, “*Negative*”). The overall blog post's sentiment on the target topic is derived by aggregating the individual clause sentiments for the target topics, where the eventual result shows the in-link blog post is negative on the “*iPad*” target topic. A summary of the influence styles score measurement methods is given in Table 6.3.

6.3.2 Influence Diffusion Detection Methodology

The Independent Cascade (IC) model (Goldenberg et al, 2001) and Linear Threshold (LT) model (Granovetter, 1978) are the two diffusion models used in this study. For the Independent Cascade model, a blogger could be independently influenced by each of the neighboring blogger based on an influence probability. While in the Linear Threshold model, a blogger has a threshold value such that the blogger is influenced if the weighted sum of the active neighbors exceeds the threshold. The two models characterize two different aspects of social interaction. The Independent Cascade approach models individual and independent interaction and influence among linked bloggers. On the other hand, the Linear Threshold model considers the overall threshold behavior of the linked neighbors in influencing the blogger.

Phase 1 of this study found similar sentiments between the linked and linking bloggers, and agreement towards the linked bloggers to be significantly correlated to influence diffusion. The Influence Diffusion Detection (IDD) model incorporates both

similar sentiments and agreement values in addition to the bloggers' influence styles as features to improve influence diffusion detection performance. Four variants of the IDD model were developed and tested as follow:

- Influence Diffusion Detection-Influence Style-Sentiments Independent Cascade (IDD-IS-Sent) (IC) model used both bloggers' influence styles and similar sentiments between the linked and linking bloggers as features in the Independent Cascade approach.
- Influence Diffusion Detection-Influence Style-Sentiments Linear Threshold (IDD-IS-Sent) (LT) model used both bloggers' influence styles and similar sentiments between the linked and linking bloggers as features in the Linear Threshold approach.
- Influence Diffusion Detection-Influence Style-Sentiments-Agreement Independent Cascade (IDD-IS-Sent-Agree) (IC) model used bloggers' influence styles, similar sentiments between the linked and linking bloggers, and agreement towards the linked blogger as features in the Independent Cascade approach.
- Influence Diffusion Detection-Influence Style-Sentiments-Agreement Linear Threshold (IDD-IS-Sent-Agree) (LT) model used bloggers' influence styles, similar sentiments between the linked and linking bloggers, and agreement towards the linked bloggers as features in the Linear Threshold approach.

Influence Diffusion Detection - Independent Cascade approach

This study applies the *Independent Cascade* (IC) approach to introduce the concept of influence styles for detecting bloggers' influence diffusion in a network. In

the IC approach, the aim is to study the use of influence styles in detecting influence diffusion that is initiated from a single linked blogger versus the *Linear Threshold* (LT) approach (Granovetter, 1978) that considers influence from all active neighbors. The bloggers' network is defined as a directed network (or equivalently graph) $G = (V, E)$, where V is the set of nodes (referred to as bloggers) and E a set of links, where each link is denoted by $e = (v, w) \in E$ and $v \neq w$, meaning there exists a directed link from a node v to a node w . For each directed link, a real value $\kappa_{v,w}$ with $0 < \kappa_{v,w} < 1$ is specified as the diffusion probability through link (v, w) . In the IC model, the diffusion process proceeds from a given initial active set $D(0)$ in the following way. When a node $v(\in D(t))$ first becomes active at time-step t , it is given a single chance to activate each currently inactive child node w , and the attempt succeeds with probability $\kappa_{v,w}$. If v succeeds, then w becomes active at time-step $t+1$, i.e., $w(\in D(t+1))$. If multiple parent nodes of w first become active at time-step t , then their activation attempts are sequenced in an arbitrary order, but all the attempts are performed at time-step t . Whether or not v succeeds, v will not make any further attempts to activate w in subsequent rounds. The diffusion process starts with an initial activated set of nodes, and then continues until no further activation is possible. The influence diffusion probability of each link is derived through introducing the *Influence Style Function* (F_{IS}) of a blogger given in Equation 6.2:

$$\begin{aligned}
F_{IS}(i) = & \left\{ \left[W_{PL} \times \frac{PL_{i-score} - PL_{min\ score}}{PL_{max\ score} - PL_{min\ score}} \right] + \left[W_{BF} \times \frac{BF_{i-score} - BF_{min\ score}}{BF_{max\ score} - BF_{min\ score}} \right] + \right. \\
& \left[W_{SC} \times \frac{SC_{i-score} - SC_{min\ score}}{SC_{max\ score} - SC_{min\ score}} \right] + \left[W_{CC} \times \frac{CC_{i-score} - CC_{min\ score}}{CC_{max\ score} - CC_{min\ score}} \right] + \left[W_{PES} \times \right. \\
& \left. \left. \frac{PES_{i-score} - PES_{min\ score}}{PES_{max\ score} - PES_{min\ score}} \right] + \left[W_{PER} \times \frac{PER_{i-score} - PER_{min\ score}}{PER_{max\ score} - PER_{min\ score}} \right] \right\} / 6
\end{aligned} \tag{Eq. 6.2}$$

The Influence Style Function (F_{IS}) is computed by taking the normalized average of the six influence scores. In the study, a simple approach is used by taking all weight (W) values to be 1 in the model evaluation. Future studies would attempt to identify optimal weight values to improve performance. The influence diffusion probability from blogger A to the respective linked blogger B is computed as: $\kappa_{A,B} = Prob(\text{Blogger A to influence}) \times Prob(\text{Blogger B is influenced})$. In the IDD-IS-Sent (IC) model, the probability of Blogger A to influence is inferred through the (F_{IS}) score of the influencing Blogger (A), while the probability of Blogger B is influenced is derived from the similar sentiment values expressed when Blogger B is linked to Blogger A. Hence,

$$\kappa_{A,B} = F_{IS}(A) \times \frac{\text{number of similar sentiment posts}_{A,B}}{\text{number of linked posts}_{A,B}} \quad (\text{Eq. 6.3})$$

where $F_{IS}(A)$ is the normalized influence style function score of Blogger A, while $\{\text{number of similar sentiment posts}_{A,B} / \text{number of linked posts}_{A,B}\}$ is the similar sentiment ratio between Blogger A and Blogger B. Based on the method used in the study in Chapter 3, the similar sentiment value at blog post level between Blogger A and Blogger B is determined by adding the positive or negative sentiments expressed on the common topics discussed by each linking and linked post. A same resultant polarity would indicate similar sentiment and conversely, different polarity indicates non-similar sentiment. A balanced polarity means neutral sentiment value. However, only similar sentiment blog posts were considered in the calculation of the similar sentiment ratio value in Equation 6.3.

To illustrate how influence style is used to detect influence in the IDD-IS-Sent (IC) model, the influence style scores for Blogger A ($F_{IS}(A)=0.73$) and the similar sentiment value between Blogger A and Blogger B ($\Omega_{A,B}=0.75$), with the computed

influence diffusion probability of Blogger A towards Blogger B given as ($\kappa_{A,B} = F_{IS}(A)(=0.73) * \Omega_{A,B}(=0.75) = 0.55$) are shown in Figure 6.5.

Figure 6.5 Independent Cascade Scoring for linked bloggers.

Figure 6.6 Linear Threshold Scoring for linked bloggers.

Blogger B will be activated if ($\kappa_{A,B}=0.55$) is greater than an arbitrary number referred as the Independent Cascade (IC) weight value. In this study, the IC weight value is determined through extensive experiments to be 0.025 for optimal performance of the IDD-IS-Sent (IC) model, with the experimental results given in Appendix H. From Equation 6.3, the influence diffusion probability ($\kappa_{A,B}$) is derived from the ability of Blogger A to influence and the degree to which Blogger B is influenced. From here, the IDD-IS-Sent (IC) model attempts to relate influence diffusion probability to influence diffusion by comparing influence diffusion probability against the IC weight as an influence threshold value.

Influence Diffusion Detection – Linear Threshold approach

In addition, the *Linear Threshold* (LT) approach is applied in the IDD-IS-Sent (LT) model to study the use of influence styles in detecting influence diffusion that is initiated from all the active neighbors of a blogger versus the Independent Cascade

approach which only considers influence diffusion from a single neighbor at each step of the evaluation process. The Linear Threshold model conceptualizes a more realistic situation in a blogosphere network as bloggers are usually linked to multiple bloggers, hence they could be influenced by their neighboring bloggers instead of only one blogger.

In the Linear Threshold model approach described by Kempe et al. (2003), a node v is influenced by each neighbor w according to a weight $b_{v,w}$ such that the total weight of its neighbors is less than 1. Each node v chooses a threshold θ_v uniformly at random from the interval $[0, 1]$, which represents the weighted fraction of v 's neighbors that must become active in order for v to become active. Given a threshold, and an initial set of active nodes A_0 (with all other nodes inactive), the diffusion process flows in discrete steps where in step t , all nodes that were active in step $t-1$ remain active and a node v is activated when the total weight of its active neighbors is at least θ_v :

$$\sum_{w \text{ active neighbor of } v} b_{v,w} \geq \theta_v \quad (\text{Eq. 6.4})$$

Thus, the Linear Threshold model intuitively represents the different latent tendencies of nodes to be influenced when their neighbors are. This study limits the influence diffusion analysis to immediate neighboring bloggers and deviates from previous approaches (Kempe et al., 2003; Java, 2006) that proposed a network-wide influence flow analysis. In this study, the weight $b_{v,w}$ of a node v taken as Blogger A that is linked to Blogger B is computed as the influence diffusion probability from Blogger A to the linked Blogger B given as $\kappa_{A,B}$ in Equation 6.3. Concurrently, Blogger B is linked to Blogger C, whom has a weight similarly computed as the influence diffusion probability from Blogger C to Blogger B given as $\kappa_{C,B}$. Blogger B is activated if the

combined normalized weights $(\kappa_{A,B} + \kappa_{C,B})/2$ of the linked Bloggers A and C to Blogger B is greater than an arbitrary threshold value, which was determined to be 0.015 through extensive experiments as described in Appendix H. From Figure 6.6, where both Blogger A and Blogger C have influence links to Blogger B, the influence style score for Blogger C ($F_{IS}(C) = 0.45$) and the similar sentiment value between Blogger C and Blogger B ($\Omega_{C,B} = 0.03$) are computed to give the influence diffusion probability of Blogger C towards Blogger B ($\kappa_{C,B} = F_{IS}(C)(=0.45) * \Omega_{C,B}(=0.03) = 0.013$). Similarly, the influence diffusion probability of Blogger A towards B is given as $\kappa_{A,B} = F_{IS}(A)(=0.73) * \Omega_{A,B}(=0.75) = 0.55$. The Linear Threshold approach further considers the overall influence of a blogger's neighbors in computing the influence diffusion probability, hence the normalized influence diffusion probability towards Blogger B is $\{\kappa_{A,B}(=0.55) + \kappa_{C,B}(=0.013)\}/2 = 0.28$. Thus, Blogger B is activated as the total influence diffusion probability is greater than the threshold value of 0.015. In this study, all neighboring blogger link pairs (Blogger A-to-Blogger B and Blogger C-to-Blogger B) are considered to be active in the test evaluations when the analyzed blogger (Blogger B) is activated as the neighboring bloggers have contributed to influencing the blogger. It can be seen that Blogger C could activate Blogger B when the Linear Threshold model is used as compared to the Independent Cascade model where Blogger C's influence diffusion probability alone would be lower than the required threshold value. This illustrates the ability of the Linear Threshold model to consider the overall influence effect of the linked bloggers in a networked blogosphere.

From the results of the exploratory study in Chapter 3, there is no significant relationship between agreement ratio values and influence, as influence could also occur in cases where there is no agreement expression. However, agreement towards the linked blogger was still considered in the IDD model as there is a clear indication

of influence when agreement exists. Evaluations were conducted to include agreement expressions towards the linked bloggers as a feature in both the IDD-IS-Sent-Agree (IC) and IDD-IS-Sent-Agree (LT) models. The Sentiment Analysis model developed in Chapter 4 was used to predict agreement towards the linked blogger through detecting agreement or positive polarity words expressed towards the linked blogger as the subject. For example, in the expression “*The always reliable 9to5mac is reporting...*” the processed AMOD([0], [+])→[+] typed dependency polarity rule would detect the neutral [0] name of the linked blogger “9to5mac” in the governor term and the positive polarity [+] of the dependent term “reliable” to give a positive polarity output towards the linked blogger. Agreement towards the linked bloggers is considered in the influence diffusion probability equation as shown in Equation 6.4, where the influence diffusion probability from blogger A to the respective linked blogger B is similarly computed as: $\kappa_{A,B} = Prob(\text{Blogger A to influence}) \times Prob(\text{Blogger B is influenced})$, but the probability of Blogger B is influenced is derived from the similar sentiment values expressed when Blogger B is linked to Blogger A as well as any agreement expressed towards linked Blogger A. Hence, the influence diffusion probability equation is given as:

$$\kappa_{A,B} = F_{IS}(A) \times (\text{similar sentiment ratio} + \text{agreement ratio}) \quad (\text{Eq. 6.4})$$

where $F_{IS}(A)$ is the normalized influence style function score of Blogger A, and the similar sentiment ratio between Blogger A and Blogger B is calculated as $\{number\ of\ similar\ sentiment\ posts_{A,B} / number\ of\ linked\ posts_{A,B}\}$ and the agreement ratio is calculated through dividing the number of posts with agreement expressed towards the Blogger A by the number of posts between Blogger A and Blogger B and is given as

$\{number\ of\ agreement\ posts_{B,A} / number\ of\ linked\ posts_{A,B}\}$ with the Prob(Blogger B is influenced) value ranging from $0 \leq \text{similar sentiment ratio} + \text{agreement ratio} \leq 1$. For both Independent Cascade and Linear Threshold models that used sentiments and agreement as features, Blogger B was activated if the IC weight and threshold values, determined through experiments similar to those described in Appendix H, exceeded 0.05 and 0.025 respectively.

6.4 Dataset Used

Seven blog sites (www.9to5mac.com, www.macrumors.com, www.tuaw.com, www.engadget.com, www.techcrunch.com, www.slashgear.com, and www.theverge.com) were identified from Technorati.com for the evaluation. The “search for more reactions” feature found in Technorati.com was used to extract the in-link posts to the target blog site. A total of 107 bloggers with 9107 in-link posts, 8879 target posts, and 19317 out-link posts were extracted from the blog sites from 11 Feb 2013 to 5 April 2013 in the dataset. A network of the bloggers’ influence links is given in Appendix D. Separately, five linked bloggers from the dataset network are described to illustrate the influence styles of individual bloggers through the Influence Style model.

6.5 Model Evaluation and Results

Influence Style Evaluation

Table 6.4 shows the influence style scores of the five linked bloggers with their blog sites given in brackets respectively. The Participating-Listening type results show

Blogger1 (PL-score=0.99) was actively posting on the target topics with minimal discussions on other topics, while Blogger4 was less participating with a PL-score of 0.27 with fewer postings on the target topics. Blogger2, Blogger3, and Blogger5 were the least participating on the target topics with PL-scores less than 0.2. In evaluating the scope of topics discussed of the five bloggers, Blogger1 and Blogger4 had a broad scope (BF-score=0.99) by covering most of the target topics. On the other hand, Blogger2 (BF-score=0.20) was more focused on specific target topics in relation to the other bloggers. Both Blogger1 and Blogger2 had higher CS-scores (>0.60), which shows them to be original in their blog post content. Blogger3 and Blogger5 were less original in their content with CS-scores of >0.50 each, while Blogger4 was seen to share the most linked post content with a CS-score of 0.32. The bloggers except for Blogger1 had similar Persuasion Style scores within the range of 0.5 to 0.55 showing a similar persuasion style.

Table 6.4 Influence style scores.

Influence style	Blogger1 (Macrumors. com)	Blogger2 (Engadget .com)	Blogger3 (Macrumors. com)	Blogger4 (Tuaw .com)	Blogger5 (Techcrunch. com)
Engagement style					
○ Participating-	0.99	0.10	0.14	0.27	0.17
○ Broad-Focused	0.99	0.20	0.64	0.99	0.41
○ Creating-Sharing	0.66	0.62	0.51	0.32	0.52
○ Consistent-Casual	0.40	0.10	0.33	0.18	0.47
Persuasion style	0.47	0.53	0.55	0.54	0.52
Persona style (PER)	0.75	0.75	0.62	0.79	0.63

On the other hand, Blogger1 was more objective in style as shown in his lower PES-score (0.47) indicating fewer subjective terms were used in the blog post content. From the Persona score results, it could be seen that Blogger1, Blogger2, and Blogger4 had high levels of agreement from their linking bloggers as shown in the high PER-score

range (0.75 to 0.79). On the other hand, Blogger3 and Blogger5 were deemed to be controversial with their mid-ranged PER-score range (0.62 to 0.63). In the persona analysis, the similar sentiments and dissimilar sentiments ratios between the in-link blog posts and the target blog posts towards the respective target topics for the bloggers were evaluated. As observed in Tan et al. (2012), neutrality expresses agreement between the linked blog posts. Neutral polarity is considered to be similar to positive or negative polarity where opinionated and neutral pairs would give similar sentiment outputs. Compliance could be inferred from the similar sentiments expressed on common target topics and related features between the linked blog posts, even if the sentiments are negative. This can be seen from the linked blog post that exhibited negative sentiments on the common feature term (iOS) as shown in Figure 6.7.

Figure 6.7 Example of compliance between linked blog posts.

On the other hand, high dissimilarities in sentiments show disagreement towards the target blog posts, which in turn indicate the linked bloggers' negative persona. This is seen in Figure 6.8, where the disagreement is shown in the positive sentiments and negative sentiments expressed by the in-link blog post and the target blog post respectively on the feature term "LTE".

Figure 6.8 Example of non-compliance between linked blog posts.

Influence Style Profiles

Figure 6.9 shows the graph of the aggregated engagement style scores, as well as the persuasion and persona scores of the five linked bloggers which indicates the relative degree of each influence style amongst the bloggers. As there is no gold standard or ground truth to evaluate influence styles, the approach used in this study was to compare the relative degree of each influence style amongst the bloggers as the best possible way to measure influence styles.

Figure 6.9 Graph of Influence Style Scores.

The influence style profiles of the bloggers are given in Table 6.5. For example, it can be seen that Blogger1 is highly engaging, objective, and positive in persona. Each blogger had an influence style which differs from bloggers from different blog sites as well as bloggers from within the same blog site. This is seen in Blogger1 and Blogger3 who are from macrumors.com, but Blogger3 displayed a less engaging, more subjective, and lower positive persona as compared to Blogger1. By further analyzing the influence styles at blogger level, the proposed IDD model could provide in-depth details of the influence exerted by the individual bloggers further differentiating the bloggers in terms of influence style.

Table 6.5 Influence Style Profiles.

BloggerID	Engagement Style	Persuasion Style	Persona Style
Blogger1	Highly Engaging (0.76)	Less Subjective (0.47)	Positive (0.75)
Blogger2	Less Engaging (0.25)	Subjective (0.53)	Positive (0.75)
Blogger3	Engaging (0.41)	Subjective (0.55)	Controversial (0.62)
Blogger4	Engaging (0.44)	Subjective (0.54)	Positive (0.79)
Blogger5	Engaging (0.39)	Subjective (0.52)	Controversial (0.63)

Influence Diffusion Detection Evaluation

Influence diffusion detection was evaluated at blogger level to assess the performance of the IDD model based on the Independent-Cascade (IC) and Linear Threshold (LT) approaches. This study's focus is on the influence diffusion between the immediate linking bloggers and not on the spread of influence across the network. Therefore, all the linking bloggers were assumed to have the possibility to influence the target blogger, and would be in an active state while the target blogger was being analyzed.

The in-degree weight defined as the number of in-links over the total number of out-links for a blogger, as well as the similar sentiment value between two linked bloggers were used in the baseline comparison to the IDD model. The similar sentiment value is computed based on the ratio of the number of similar sentiment posts between the pair of linked bloggers over the total number of posts between the linked bloggers. For the in-degree baseline approach, a node is activated when the in-degree weight of the influencing blogger is higher than that of the influenced blogger. In the second baseline approach, similar sentiment values greater than 0.5 between the linked bloggers were used as indication of influence diffusion. In addition, a third baseline comparison using the similar sentiment values between linked bloggers as the diffusion probability values in a Sentiments-Independent Cascade (IC) approach was used in the evaluation. Extensive testing showed optimal performance is achieved when the IC weight value for the Sentiments-IC model equals 0.025.

In the evaluation testing, 514 linked blog posts of 362 unique blogger pairs were manually classified into 265 influencing and 97 non-influencing blogger pairs. The two coders' Cohen's kappa index (Cohen, 1960) was computed to be 0.68, with the conflicting tags reviewed and then manually re-classified and used as the answer keys. Samples of the manually coded data for influence and non-influence cases are given in Appendix E. The performance of the models in detecting influence diffusion was measured using the F1-score measure given as $F1 = (2 \cdot \text{Precision} \cdot \text{Recall}) / (\text{Precision} + \text{Recall})$. In the Independent Cascade and Linear Threshold model, a blogger-pair consisting of a neighboring blogger and the evaluated blogger is deemed to be influencing when the evaluated blogger is activated. In the Linear Threshold model approach, the influence diffusion analysis is constrained to immediate neighboring bloggers due to the limited span in the chain of linked bloggers found in

the manually coded data as well as the scope of this study that is confined to the analysis between direct-linked bloggers. Further tests were conducted using the Independent Cascade (IDD-IS-Sent-Agree) (IC) and Linear Threshold (IDD-IS-Sent-Agree) (LT) models, which include agreement towards the linked blogger as a feature.

The results of the evaluation testing are shown in Table 6.6. It can be seen that the IDD-IS-Sent (IC) and IDD-IS-Sent (LT) models, both with F1-score=0.76, performed better than the In-Degree (F1-score=0.45), Sentiments-only (F1-score=0.63), and Sentiments-Independent Cascade (F1-score=0.68) baseline approaches. The poor performance in the In-Degree approach could be attributed to the fact that a blogger's post may not indicate influence but can also express disagreement towards the linked blogger's posts.

Table 6.6 Influence diffusion detection evaluation test results.

	In-Degree	Sentiments-Only	Sentiments (IC)	IDD-IS-Sent (IC)	IDD-IS-Sent (LT)	IDD-IS-Sent-Agree (IC)	IDD-IS-Sent-Agree (LT)
Precision	0.63	0.72	0.73	0.73	0.72	0.79	0.73
Recall	0.35	0.56	0.64	0.79	0.81	0.77	0.79
F1-score	0.45	0.63	0.68	0.76	0.76	0.78	0.76

Both the sentiment value approaches performed better than the In-Degree approach as similar sentiment expressions between linked bloggers provide a clearer indication of influence through the common stance on the topics discussed. The higher performance of the Sentiments Independent Cascade (IC) approach versus the Sentiments-Only approach shows the capability of the Independent Cascade method in modelling influence diffusion as compared to using only the sentiment values between linked bloggers. Both IDD-IS-Sent (IC) and IDD-IS-Sent (LT) models performed well against the baseline methods in detecting influence diffusion within a blogger network.

This could be because the IDD-IS-Sent (IC) and IDD-IS-Sent models relate closely to influence diffusion by further considering the bloggers' influence styles, where each of the influence style characteristic provides clearer indication of the blogger's propensity to influence. The similar F1-scores between IDD-IS-Sent (IC) and IDD-IS-Sent (LT) indicate on par performance in detecting influence diffusion flow for both approaches. However, recall for IDD-IS-Sent (LT) is better and this could be because the IDD-IS-Sent (LT) model considers the influence probability from all active neighbours of a blogger resulting in higher probability of activating the neighbour-blogger pair as compared to the IDD-IS-Sent (IC) model where only a single neighbour is evaluated at each step. From the results of the IDD-IS-Sent-Agree (IC) and IDD-IS-Sent-Agree (LT) models, the detection of agreement towards the linked blogger further reinforced the presence of influence between the linked and linking bloggers, improving precision performance. Agreement as a feature was less effective for the Linear Threshold approach as seen in the lower precision scores. This is because the Linear Threshold approach considered the influence of neighbouring bloggers collectively, which could have mitigated the impact of a linked blogger with agreement expressions. It was also observed that the recall scores for both IDD-IS-Sent-Agree (IC) and IDD-IS-Sent-Agree (LT) models were lower as precision increased.

This study further analyzes the relation between influence styles scores and the bloggers' ability to influence to determine the significant influence styles that could detect influence diffusion. From the independent sample t-test results between influence styles and influence diffusion indicator shown in Table 6.7, it could be seen that the influence styles, except for the Consistent-Casual (CC) style, are significantly different ($p \leq 0.05$) in their $\mu(\text{yes})$ mean score and $\mu(\text{no})$ mean score, indicating their ability to detect influence. The difference in mean scores for CC-style was

insignificant ($p>0.05$), which shows both influential and non-influential bloggers to have similar consistency in their blog posts' links. This also means CC-style was not able to differentiate a blogger's ability to influence and could be omitted from the Influence Style Function computation.

Table 6.7 T-test results.

Influence Styles	p	Influence diffusion	
		$\mu(\text{yes})$	$\mu(\text{no})$
Participating-Listening	0.001	0.31	0.17
Broad-Focused	0.013	0.59	0.49
Creating-Sharing	0.000	0.49	0.42
Consistent-Casual	0.182	0.41	0.44
Subjectivity	0.013	0.49	0.51
Persona	0.002	0.79	0.76

Further experiment on the effect of topic on the use of influence style to detect influence diffusion showed Persona Style to be topic dependent. When the PER-scores of the bloggers were computed based on sentiments expressed on both Apple and Android related posts, performance of the IDD-IS-Sent (IC) model was lower at (Precision=0.72, Recall=0.64, F1-score=0.68) compared to when only bloggers' Apple related posts were used to compute the PER-scores with results (Precision=0.73, Recall=0.79, F1-score=0.76) shown in Table 6.6. In addition, t-test results between influence styles and influence diffusion indicator showed Persona to be insignificant ($p>0.05$) when both Apple and Android topics were considered in the PER-scores. A possible explanation could be that a blogger who is positively recognized in the "Apple" topic domain is usually not equally agreed upon for Android related postings, thus rendering Persona style not suitable for general topic influence diffusion detection.

6.6 Discussion

The knowledge on individual blogger's influence style was used in the IDD model to detect influence diffusion in the blogger network. The IDD model performed better than the baseline methods as the model further maps the influence characteristics of the bloggers to the links between bloggers. This provides a closer match to the influence patterns in the blogger network which gives better performance for detecting influence diffusion as compared to the use of the presence of links and their sentiment polarity.

In comparison, IDD-IS-Sent (IC) has higher precision while IDD-IS-Sent (LT) has higher recall. The higher number of influencing blogger-pairs in the dataset could have resulted in a higher recall value for the IDD-IS-Sent (LT) model due to the Linear Threshold model's propensity to classify bloggers as influencing through combining the influence probabilities of neighboring bloggers. On the other hand, precision performance of the IDD-IS-Sent (LT) model was affected because there was a possibility that non-influencing pairs were unwantedly activated as a result of the influence probability scores of active neighbors. For example in Figure 6.6, there could be no influence diffusion from Blogger C to Blogger B due to the low similar sentiments value between them, but nonetheless the Blogger C-Blogger B pair was activated due to the overall influence probability scores being higher than the threshold value.

Precision performance for both IDD-IS-Sent-Agree (IC) and IDD-IS-Sent-Agree (LT) models were higher and this could be because agreement expressions towards the linked bloggers further indicate the ability of the linked bloggers to exert influence and could be used to detect influence diffusion between the bloggers more precisely.

The use of the Consistent-Casual influence style type was found to be not significant in detecting influence diffusion, which means that the number of in-links to past blog postings has no effect on influence diffusion detection. This shows that the timeline of blog postings is not a clear indicator of influence diffusion in a blogger network. The study also revealed influence diffusion detection to be topic specific and performance would suffer if the Persona influence style is computed based on generic topics.

In addition to detecting influence diffusion, the IDD model further describes the manner and style in which influence is diffused through the influence path in the network by providing the bloggers' influence styles. From the results, it could be seen that influential bloggers within the blogger network tend to be participating, broad, creative, objective, and have a positive persona. The influence styles of each blogger along the influence path are plotted to visualize the pattern of influence styles in the influence flow.

Figure 6.10 Influence diffusion path.

For example, Figure 6.10 shows an extract of the influence path of the five evaluated bloggers, where it can be seen there is influence flow from Blogger1 who is highly participating, more objective, and positive in persona towards bloggers who have influence styles that are listening, more subjective and less positive in persona.

It is noted that influence style can be specific to blog domains as seen in a reputed political blogger that posts infrequently, but holds much weight compared to an influential product review blogger who is actively engaging the readers. Nevertheless, the IDD model could be generalized across blog domains to detect influence diffusion through identifying the influence styles of individual bloggers of each domain. This study has assumed equal weight for the influence styles in the influence diffusion probability (κ) computation by taking the normalized average of the influence scores. Future work could further analyze the importance of each influence style in determining the optimal performance of the IDD model. For example, persuasion style could be more significant in determining influence diffusion between opinionated political bloggers and be given a higher weight in the analysis.

Though this study has shown the IDD model to detect and describe influence diffusion effectively, it is noted that bloggers' influence styles are not static and are dependent on the period of analysis. Hence, the effectiveness of the IDD model is specific to the period of analysis. A different period of analysis might have identified the influence style differently since the measurement values would have changed. The IDD model is also dependent on the performance of the sentiment analysis approach. Any performance changes to the sentiment analysis would affect the results of the IDD model.

6.7 Summary

The study in this chapter developed the Influence Diffusion Detection (IDD) model using the influence styles of bloggers as additional features to detect influence flow between linked bloggers to fulfill the research objective (RO3: *Develop an influence diffusion detection model using sentiment values between linking and linked bloggers, agreement towards the linked bloggers, and bloggers' influence styles as possible features to improve influence diffusion detection between linked bloggers*). In contrast to previous studies which used a monolithic definition of influence for bloggers, the IDD model further describes the influence of bloggers in terms of engagement, persuasion, and persona style. The detailed analysis of influence could differentiate the influence style of bloggers, and provide a more comprehensive description of the influence exerted in response to the research gap (RG3: *Influence diffusion could not be described using a single detached feature based on the existence of links or similar sentiment polarity*).

The results show there are differences in influence style between individual bloggers, which means influence is exerted uniquely among the bloggers. The provision of the influence style would enhance the influential blogger search experience through the detailed description of the manner in which the bloggers exert influence. Further to that, a more comprehensive and accurate portrayal of the blogger's influence through the evaluated influence profiles would in turn improve influence detection within the blogosphere, which answered the research question (RQ3: *Will automatic influence style analysis help detect influence diffusion between linked bloggers?*).

Both the Independent Cascade and Linear Threshold models were evaluated as possible approaches to be used in the IDD models. In comparison, results showed that

the IDD-IS-Sent (IC) had higher precision while IDD-IS-Sent (LT) had higher recall when similar sentiments and influence styles were used as features. In addition, the use of agreement expressions towards the linked blogger as a feature provides better precision performance for the IDD-IS-Sent-Agree (IC) and IDD-IS-Sent-Agree (LT) models as agreement provides a clearer indication of influence diffusion between the linked bloggers.

The IDD model relates closely to the influence characteristics of the bloggers and performed well compared to the in-degree and sentiment values baseline approaches. In addition to detecting influence diffusion in a blogger network, the IDD model describes the influence style of each blogger in the influence path, providing a fine-grained description of the manner in which influence is diffused.

CHAPTER 7

CONCLUSION

7.1 Introduction

This chapter provides the conclusion of this research study by reviewing the objectives of the research and comparing this work with related work, followed by a discussion on the contributions, limitations, and future work. Towards this end, the chapter is organized into six sections. Section 7.2 reviews the objectives of each phase of the study, and discusses the applied methods to address the three research objectives. A comparison of this research with other related research is given in Section 7.3. The following Section 7.4 provides the contribution as well as practical applications of this research. Next, the limitations of the research are presented in Section 7.5, with the recommendations for future work given in Section 7.6. Finally, the chapter brings the research to a close with the concluding remarks.

7.2 Review of Research Objectives

The overall aim of this dissertation study is to develop an influence detection model to automatically detect the influence flow from the linked blogger to the linking blogger, where the ability to detect influence diffusion in the blogosphere could be used to identify influential bloggers and the chain of information flow. The research study is done in three phases; the first phase performed an exploratory study to identify blog features suitable for detecting influence diffusion, while the second phase

established a sentiment analysis method in consideration of the blog posts' sentiments on the common topics. The third phase focused on developing an influence diffusion detection model that uses bloggers' influence styles in addition to similar sentiments and agreement values as features to improve performance as well as provide a better understanding on how influence is diffused in the blogosphere.

7.2.1 Study on blog features analysis

An overview of the first phase of the research's methods and objective is shown in Figure 7.1.

Figure 7.1 Overview of the research's first phase approach.

Previous studies (Adar and Adamic, 2005; Agarwal and Liu, 2008) used blog features such as number of in-links, number of out-links, length of posts, etc. to relate to influence. As stated in the first research gap (RG1: *The use of simple blog features to detect influence diffusion may not yield highly accurate results as influence itself is a complex concept that requires more in-depth analysis, where other intrinsic blog features that provide further details on the influence diffusion process could be used to improve influence diffusion detection performance*), using these simple features may

not necessary indicate influence as influence is a complex concept that could be better detected through more in-depth analysis. This led to the initial study of this research in Chapter 3, which applied research method (RM1) through exploring the possible approaches and methods to detect influence diffusion between linked blog posts. Blog features were analyzed through statistical methods to determine the possible features that could detect influence within the linked blog posts. The study investigated three common approaches (quantitative blog features analysis, content analysis, and community identity) and used the combination of blog features, content analysis, community identity, and manual interpretation in the evaluation of the effectiveness of each method in detecting influence in the blogosphere. Blog features evaluated in the study are the blogs' authority rank, text similarity as in number of copied sentences from linked post to linking post, length of posting as in the number of sentences in the blog post, time difference between postings, and community identity. For content analysis, sentiment and agreement analysis were applied since sentiments expressed on common topics by two linked posts and agreement expressed by the linking blogger on the linked blog within the blog post could be indicators of influence between the blog posts.

The results of the study showed that the opinion stances between two linked bloggers on specific topics are reliable indicators of influence, which means an analysis on the blog post content and the comparison of the sentiments expressed on topics with the linked blog post will provide better interpretation of influence. Though community identity was found to be useful for influence detection, it was not considered in the proposed influence diffusion detection model as the aim of this dissertation study is focused on the use of influence styles for influence diffusion detection. Subsequently, the study in Chapter 4 attempted to establish a sentiment

analysis method that takes into account the opinions expressed in the blog post content through a typed dependencies and subjective phrase lexicon approach.

7.2.2 Study on sentiment analysis

In the second phase of the research, the study in Chapter 4 applied research method (RM2) using typed dependencies rules and a subjective phrase lexicon to analyze the sentiments expressed in the blog posts. These include the possible ways to construct the lexicons for the semantic analysis of words to handle the complex phrases' relationships. An overview of the second phase approach is given in Figure 7.2.

Figure 7.2 Overview of the research's second phase approach.

The study adopted the typed dependencies rules and subjective phrase lexicon analysis approach to predict the sentiments within blog posts as it could provide an in-depth grammatical and lingual syntactical explanation for the predicted results. The focus of the sentiment analysis was at the phrase level, which the study specifically defines as the bigrams in typed dependencies. The use of class sequential rules (CSR) was

explored to automatically learn the typed dependency patterns, and the performance between CSR and a heuristic method were compared. The study also introduced polarity-only pattern rules and the results show this leads to improvement in classification performance.

The findings from the study show that CSR is capable of automatically generating a comprehensive list of bigram part-of-speech patterns compared to a heuristic approach and consequently, it was shown CSR generated rules could improve sentiment classification performance. It was also shown that rules using polarity features perform better than those using part-of-speech polarity features. The typed dependencies rules considered the phrases' grammatical and syntactic structures when predicting the sentiments expressed on specific target topics. Through this, the typed dependencies rules provided an in-depth understanding of the sentiments expressed on the topics in the blog posts. Hence, the research objective (RO2) to establish a sentiment analysis approach that takes into consideration the complex relationships between words in blog posts to automatically detect influence diffusion between linking and linked blog posts was fulfilled through the use of automatically generated typed dependencies and additional semantic rules. The derived typed dependencies and semantic rules are subsequently used to detect bloggers' persona style and to determine the sentiments between linked blog posts in the proposed Influence Diffusion Detection (IDD) model introduced in Chapter 6.

7.2.3 Study on influence style profiling

The third research gap (RG3) states that influence diffusion could not be described using a single detached feature based on the existence of links or similar sentiment polarity. This is because knowing the polarity of opinions expressed

between linked bloggers does not reveal how influence is diffused. Consequently, the third phase of the research study focused on understanding how, as in the manner in which influence is diffused in the blogosphere through analyzing the bloggers' influence styles and on developing an influence diffusion detection model that uses the bloggers' influence styles, in addition to similar sentiments and agreement values as features to improve performance. An overview of the third phase of the research is shown in Figure 7.3.

Figure 7.3 Overview of the research's third phase approach.

The study in Chapter 5 applied a research method (RM3-i) analyzing bloggers' influence styles at blogger-level to provide an in-depth understanding of the manner in which bloggers exert influence. Influence style was defined in terms of engagement style, persuasion style, and persona style, whereby the sentiment analysis method proposed in Chapter 4 was used to determine the sentiments expressed in the blog posts and the persona style of the bloggers.

The results of the study in Chapter 5 showed that bloggers could differ from the overall blog site and other bloggers in influence style. For example, in the Engagement style, the level of participation and focus on topics could differ between the individual bloggers. Being able to differentiate the participation level and topic focus of individual bloggers allows the bloggers' depth and scope of knowledge in respective topics to be determined. Bloggers could also exhibit similar influence styles due to the limited sharing of other blogs' content because of copy-rights constraints, as well as similar persuasion style based on the genre of the blog site. For example, a product review blog site which tends to be subjective versus news blog sites that are usually objective. Likewise, popular bloggers tend to have positive persona with in-link bloggers expressing agreement towards them. Nonetheless, there were also controversial bloggers with both agreement and disagreement opinions from their linking bloggers.

By analyzing the influence styles of individual bloggers, this study was able to provide a fine-grained description of the manner in which influence is exerted by the linked bloggers. This knowledge on individual blogger's influence style was used in the subsequent study in Chapter 6 to detect influence diffusion between linked bloggers.

7.2.4 Study on influence diffusion detection

The study in Chapter 6 applied the research methods (RM3-i) and (RM3-ii) to develop an influence diffusion model based on the Independent Cascade (IC) model and Linear Threshold (LT) model. While phase one of this study found similar sentiment and agreement values to be significantly correlated to influence diffusion, the study fulfilled research objective (RO3) by incorporating the influence styles of

bloggers as additional features into developing the Influence Diffusion Detection (IDD) model to improve the performance of influence diffusion detection between linked bloggers. An overview of the IDD model is shown in Figure 7.4.

Figure 7.4 Overview of the IDD model.

The IDD model performed well compared to the in-degree and sentiment values baseline approaches, as it maps the influence characteristics of the bloggers to the links between bloggers. In comparison, the IDD-IS-Sent (IC) model had higher precision, while the IDD-IS-Sent (LT) model had higher recall. This could be because the IDD-IS-Sent (LT) model considers influence probability from all the active neighbors of a blogger resulting in higher probability of activating the neighbor-blogger pair. On the other hand, the IDD-IS-Sent (IC) model evaluates only a single

neighbor at each step of the process to provide more precise results. The IDD-IS-Sent (LT) model conceptualizes a more realistic situation in a blogosphere network as bloggers could be influenced by a combination of their neighboring bloggers instead of only one blogger. Nevertheless, the IDD-IS-Sent(IC) model provides better clarity on the influence diffusion between a single pair of bloggers.

Agreement towards the linked bloggers was included as a feature in the IDD-IS-Sent-Agree (IC) and IDD-IS-Sent-Agree (LT) models. The results showed that the presence of agreement expressions is a clear indication of influence as seen in the improved precision performance for the models. Between the two models, IDD-IS-Sent-Agree (IC) had a higher precision performance as the effect of the agreement feature was confined to the link between two bloggers. On the other hand, by considering all neighboring bloggers, the IDD-IS-Sent-Agree (LT) model might have mitigated the effect that a single linking blogger had while expressing agreement towards the linked blogger.

In addition to detecting influence diffusion, the IDD model further describes the manner and style in which influence is diffused through the network path by providing the bloggers' influence styles. The detailed analysis of influence style between linked bloggers provided by the IDD model gives a comprehensive description of the influence exerted, and improved influence diffusion detection performance achieving the overall research goal of developing an influence diffusion detection model to automatically detect the influence flow between two linked bloggers.

7.3 Comparison with Related Work

Previous studies (Herring et al., 2005a; Agarwal and Liu, 2008; Adar and Adamic, 2005) that used blog features to detect influence were mainly graph-based. However, analyzing simple blog features such as the number of in-links and out-links, length of posts, etc. could not give highly accurate influence detection performance as influence itself is a complex concept that could be detected through further in-depth analysis. For example, a linking blog post opposing the ideas and opinions of a high authority blog post is displaying non-influenced behavior, while further analyzing the sentiments expressed in the linking blog post would have provided a clearer evidence of influence. The first phase of this dissertation study explored the blog features that could be used to detect influence and the statistical results showed that sentiments expressed between the linked bloggers are useful indicators of influence propagation between the bloggers.

The second phase of this study proposed a sentiment analysis method that leverages typed dependencies rules and additional semantic rules to further consider the sentiments expressed in the blog posts to improve sentiment detection performance. Previous studies (Wilson et al., 2005; Moilanen and Pulman, 2007) studied sentiment analysis at clause or phrase level to provide a more refined analysis as compared to a document-level or sentence-level approach (Riloff and Wiebe, 2003; Turney, 2002). Recent studies (Shaikh et al., 2008; Thet et al., 2010) focused on the functional relations of words using typed dependency parsing, which provides a detailed analysis on the grammar and semantics of textual data. These studies used a heuristic approach to determine the typed dependency pattern rules. However, a heuristic method might not comprehensively identify all possible rules due to the limitations of manual observation. In this study, both linguistic and machine learning methods were

combined to automatically detect the sentiment polarity of bigrams in specific typed dependencies. This study applied the Class Sequential Rules (CSR) to derive the typed dependencies rules from the sequential patterns within the word bigrams and further took into account the complex relationships between words to yield better performance versus the heuristic approach.

More recent studies (Leskovec et al, 2010; Cai et al, 2011; Li et al., 2011) similarly identified sentiment indicators as useful in detecting influence and analyzed the sentiments between links to determine the influence flow within a network. However, these studies focused on a single notion in detecting influence, that is, whether influence exists between the linked bloggers, and had not studied in details the manner in which influence is exerted. The third phase of this study further analyzed the influence styles of bloggers as additional features to the similar sentiments on common topics between blog posts and agreement expressed towards the linked bloggers to detect influence diffusion based on the Linear Threshold (LT) model and Independent Cascade (IC) model approach. To the best of the knowledge available, this research is the first study that used influence styles as features in an influence diffusion model to improve the performance of influence diffusion detection.

Previous studies (Gruhl et al, 2004; Barbieri et al., 2012; Garg et al., 2012) similarly measured blog feature metrics and further applied the metrics on the Threshold model and Cascade model to detect influence within social networks. Instances of blog features used were based on the probabilities of a blogger linking to or reading a blog post, and the studies had not considered the manner in which influence is diffused. On the other hand, this study identified the influence styles of bloggers to provide a clearer understanding of the influence being propagated, which was shown to yield good performance. In addition, the profiling of bloggers' influence

styles provides details on the manner in which influence is propagated through the linked paths in the blog network.

7.4 Contributions

This study proposed a novel approach of detecting influence diffusion between two linked bloggers through the additional use of bloggers' influence styles as features in the developed Influence Diffusion Detection (IDD) model. In addition to improving influence detection performance, the IDD model provides further description of the bloggers' influence style. Knowing the manner in which bloggers exert influence through their influence styles gives a better understanding on how influence is diffused across the linked bloggers. Bloggers could then know themselves better on the style they exert influence. From here, influence styles that are influential could be identified and fostered to exert more influence. For example, a blogger could express more subjectivity in their blog posts if linked bloggers are better influenced through persuasion. Improving on their influence styles allow bloggers to increase their influence in the blogosphere and in turn reach out to a wider group of readers.

This dissertation study contributes to research and practice in three dimensions. First, the research study performed quantitative analysis of features found in the blogs, as well as the sentiment and agreement analysis on the blog postings content, and community identification to determine the usefulness of each feature in detecting influence diffusion in the blogosphere. Previous studies (Adar and Adamic, 2005; Agarwal and Liu, 2008) focused on the use of blog features and had not fully considered the sentiments expressed in the blog post content in detecting blogosphere influence. Knowing the important blog features and the effect of content analysis in

detecting influence among blog posts allows development of a more efficient influence diffusion model for determining the main chain of influence propagation in the blogosphere. The results from the study showed that sentiments on common topics are highly correlated with influence diffusion and can be used to detect influence propagation between linked blog posts. Unlike most of the previous content analysis studies (Matsumura et al., 2008; Song et al., 2007) that measure document similarity (i.e. common terms) between linked blog posts, this research study applies sentiment analysis to consider the sentiments expressed within the whole blog post content.

The second contribution of this research is to establish a sentiment analysis method that leverages a detailed evaluation of the linguistic and semantic significance of phrases to take into consideration the sentiments expressed in the blog posts. This study introduces semantic rules into the sentiment analysis process to handle the complex relationships that exists between words. It was observed that using polarity rules provide results that are on-par, if not better compared to POS-polarity rules. This study improves the sentiment analysis process by removing the need to identify the part-part-of-speech (POS) in the rules derivation. It is noted, the word-term and POS are still required in the semantic analysis to appreciate the meaning of the words. The proposed sentiment analysis method based on the CSR and additional semantic rules improves sentiment analysis performance as compared to the previous heuristic approach (Thet et al, 2010) through identifying more comprehensive rules to handle the complex relationships between words.

The final contribution of this research is the novel approach to analyzing bloggers' influence styles, and using the influence styles, in addition to similar sentiment and agreement values as features in the proposed Influence Diffusion Detection (IDD) model to improve the performance of influence diffusion detection

performance between linked bloggers. Identification of bloggers' influence style would give a better description and understanding of the influence exerted by the bloggers through their engagement style, persuasion style, and persona. For example, influencing bloggers could be persuasive or objective, and could post actively or be usually dormant, but would post few highly relevant postings. The study revealed through an empirical approach that influence style is specific to individual bloggers, and the blog site influence style may not be representative of all bloggers. The proposed IDD model provides a more detailed description in which influence could be diffused across linked bloggers as compared to previous studies which relied on simple blog features or direct sentiment polarities between links. In addition to detecting influence diffusion, the IDD model further describes the manner and style in which influence is diffused in the influence network path through providing the subsequent linked bloggers' influence styles. From here, the influence styles of each blogger along the influence path could be plotted to visualize the pattern of influence styles in the bloggers' network. A list of published papers based on this research study is given in Appendix I.

7.5 Limitations of Study

The blog features study in Chapter 3 evaluated community identity to be useful for influence detection. However, community identity was not considered in the proposed IDD model as the scope of this dissertation study is limited to the use of sentiments on common topics, agreement towards the linked bloggers, and influence styles for influence diffusion detection. Identifying community is also challenging for the product review domain as bloggers sometimes do not easily identify themselves

with any community. For example, a blogger could write on both Android and Apple topics, and does not identify himself or herself with either community. Hence, the evaluation of the community identity for the product review dataset was not done in this study.

It is noted that in the proposed Sentiment Analysis model using the CSR approach, the execution priority of overlapping rules from different classes poses a problem. A generic but higher priority rule could override a specific but lower priority rule from another class resulting in a wrongly classified polarity by the generic rule. As with all learning algorithms, the discovery of rules using CSR is dependent on the dataset used where the distribution and frequency of the polarity cases determine the rules to be derived, which in turn affects the performance. Hence, as with any rule-based approach, the ordering of rules has to be considered carefully. The part-of-speech labels used in the semantic analysis of words was determined using the Stanford Parser. It was observed that there was a small percentage of error of less than 5% in the part-of-speech tagging, for example, AMOD(regard:VB, be:VB), which was controlled using the support and confidence threshold values. In building the lexicons for the semantic analysis of words, the comprehensiveness of the lexicons is limited by the training dataset and current knowledge-base on the observed complex semantic words and phrases. The quality of the lexicons is also subjected to the domain and context in which the lexicon words are built. For example, the word “Apple” could refer to a fruit or a technology company depending on the context used. Any limitation on the subjectivity lexicons would have an impact on the sentiment analysis performance when subjective words or phrases are not detected. While previous studies (Somasundaran and Wiebe, 2009; Ruppenhofer et al., 2008) adopted a discourse-level relationship analysis approach to provide better performance, this study

limits its scope to using the connective terms identified through the phrase-level typed dependencies to identify the discourse relations in a sentence. Consequently, the performance of the Sentiment Analysis model affects the influence style profiling outcome for the bloggers' persona style identification as well as the influence detection performance.

This study has assumed equal weight for the influence styles in the influence diffusion probability (κ) computation, which may not be the case in a blogosphere environment. For example, readers could be more influenced by a subjective persuasion style blogger who blogs in a less engaging style as compared to a blogger who blogs actively but is objective in style. In this case, persuasion style could be given more weightage over the engagement style in the IDD model to improve the influence diffusion detection performance. Influence styles could vary across bloggers from different blog domains as seen in a highly regarded political blogger who blogs infrequently compared to an influential product review blogger who actively blogs. Political blog readers could also be influenced easily by a subjective persuasion style blogger, while readers in a product review domain is influenced by an objective persuasion style blogger. Nevertheless, by identifying the influence styles of bloggers in each domain, the IDD model could be generalized across different domains.

Bloggers' influence styles were derived using a quantitative analysis of the selected blog features based on deterministic thresholds to provide an objective measurement to the influence styles evaluation. However, there is no gold standard used in practice to confirm the influence styles, hence the study could only compare the influence styles of the bloggers relative to each other. This study limits the detection of persona style through analyzing the sentiments on common topics between the linking and linked blog posts. Agreement analysis was not used as there

are not many explicitly mentioned agreement or disagreement cases between linked blog posts. Nevertheless, detecting agreement towards the linked bloggers could have improved the analysis on the persona style of bloggers.

The results of the bloggers' influence style analysis are subjected to the period and duration of data collection, where a different period of analysis would have given a different set of measurement values resulting in different influence styles being identified. In this study, the influence diffusion detection is confined to between direct-linked bloggers due to the constraint in the number of coded data. Nevertheless, influence diffusion across a large blogosphere network could be inferred through analyzing the influence diffusion through subsequent linked bloggers.

7.6 Future Work

Though this study evaluated community identity to be useful for influence detection, community identity was not considered in the proposed IDD model. Future work could include community identity as a possible feature in the influence detection model by using a dataset that contains bloggers with a clear indication of their community.

In the sentiment analysis evaluation, polarity conflict cases were systematically processed according to the priority of the generated rules. Future work could perform a separate semantic analysis on the bigrams' words and part-of-speech to resolve polarity conflicts through identifying semantic relationships of the bigram words to improve performance. The performance of the Sentiment Analysis model is also subjected to the context of the sentences. For example, the word "*little*" is considered negative in the subjectivity lexicon, but is neutral in context in the sentence "*put up*

a *little* gallery with a bunch of the template...” Future work could look into the detection of contextual words to identify the correct polarity for the phrase. For example, detecting the words “*put up a*” before the word “*little*” could identify the phrase “*put up a little*” as neutral. This study limits its scope to the use of phrase-level typed dependencies to identify the discourse relations in a sentence. Future work can further analyze clause-level typed dependencies patterns to improve sentiment analysis performance. In the Sentiment Analysis model, unmatched bigrams were classified as in the neutral class by default. Further investigation into the possible classification for unmatched bigrams could have given a better performance.

Results from this study found that there are clear differences in influence styles between the bloggers in the dataset. However, there is no real gold standard or ground truth to compare the analyzed influence styles against. Future work could determine a method for validating this type of analysis. The bloggers studied came from multi-blogger sites, and it is possible that they were governed or responding to a “house style” which might increase similarities within a site. In addition, bloggers could be constrained in their style based on the blogosphere domain, for example, political bloggers could be more subjective in persuasion style as compared to the product review bloggers who tend to be more objective in their persuasion style. An examination of the influence style differences within a blog site and across different blog sites could be done in future work. This study has assumed equal weight for the influence styles while computing the influence diffusion probability in the IDD model. Future work could further analyze the importance of each influence style to determine the optimal performance for detecting influence diffusion.

The IDD model only measures the overall influence styles within the analyzed time period. Future work could look into the influence style time dependency to

explore possible factors, for example, a change of sentiment polarity towards a product or a reduced number of postings over a period of time that could affect a blogger's influence style. This study focused on the analysis of bloggers' engagement, persuasion, and persona influence styles, future work could be done to study bloggers' emotions to provide further details on the bloggers' influence styles. As this study focused only on the influence diffusion detection between two linked bloggers, future work could explore the possible influence style patterns that indicate influence flow in a large network to better describe the manner in which influence is propagated. The results of this study also reinforced the notion that influence diffusion is topic specific. This study has used a manual pre-defined topic scope. However, future work could include a wider range of topic-related terms through using an automatic topic detection approach, for example, possibly based on the Latent Dirichlet Allocation (Blei et al., 2003) method.

7.7 Concluding Remarks

The Influence Diffusion Detection (IDD) was developed in this dissertation study to fulfill the overall objective of developing an influence diffusion model to automatically detect the influence flow between two linked bloggers. The study explored the relevant blog features that could be used to detect influence flow in response to the research gap where the use of simple blog features may not yield accurate results. Similar sentiments expressed between the linking and linked bloggers was determined to be significantly correlated to influence diffusion. Subsequently, a sentiment analysis approach was further established, in consideration of the complex

relationships between words, through the use of typed dependencies rules and semantic rules.

This study introduced the novel approach of analyzing bloggers' influence styles through empirical means and this allows a better understanding of how influence is propagated between linked bloggers. Results of the study showed that the use of bloggers' influence styles in addition to similar sentiments and agreement values as features in the proposed IDD model give good performance and further provide visualization of the influence style flow in the network path. This dissertation study has provided a means to effectively detect influence diffusion across linked bloggers automatically through the proposed IDD model.

REFERENCES

- Adar, E., & Adamic, L. A. (2005). Tracking information epidemics in blogspace. *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, Compiegne, France (pp. 207-214). NY: IEEE Computer Society.
- Agarwal, N. (2008). A study of communities and influence in blogosphere. *Proceedings of the 2nd SIGMOD PhD Workshop on Innovative Database Research*, Vancouver, Canada (pp. 19-24). NY: ACM Press.
- Agrawal, R., & Srikant, R. (1994). Fast Algorithms for Mining Association Rules in Large Databases. In J. B. Bocca, M. Jarke, C. Zaniolo (Eds.), *Proceedings of the 20th Conference on Very Large Databases*, Santiago de Chile, Chile (pp.487-499), CA: Morgan Kaufmann.
- Agarwal N., Liu H., Tang L., & Yu P. S. (2008). Identifying the influential bloggers in a community. In M. Najork, A. Z. Broder, S. Chakrabarti (Eds.), *Proceedings of the International Conference on Web Search and Web Data Mining* Palo Alto, California, USA (pp.207-218). NY: ACM.
- Agarwal, N., & Liu, H. (2008). Blogosphere: Research issues, tools, and applications. *SIGKDD Explorations Newsletter*, 10(1), 18-31.
- Agarwal, N. & Liu, H. (2009). *Modeling and Data Mining in Blogosphere*, CA: Morgan & Claypool.
- Aman, S., & Szpakowicz, S. (2007). Identifying Expressions of Emotion in Text. In V. Matousek & P. Mautner (Eds.), *Proceedings of the 10th International Conference on Text, Speech and Dialogue*, Pilsen, Czech Republic (pp. 196-205). Berlin: Springer-Verlag.
- Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The Berkeley FrameNet Project. In C. Boitet, P. Whitelock (Eds.), *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Montreal, Quebec, Canada* (pp. 86-90). Stroudsburg, PA: ACL.
- Bailey, N. T. J. (1975). *The Mathematical Theory of Infectious Diseases and its Applications* (2nd edition). Griffin, London.
- Barbieri, N., Bonchi, F., & Manco, G. (2012). Topic-Aware Social Influence Propagation Models. In. M. J. Zaki, A. Siebes, J. X. Yu, B. Goethals, G. I. Webb & X. Wu (Eds.), *Proceeding of the IEEE 12th International Conference on Data Mining* Brussels, Belgium (pp. 81-90). NY: IEEE Computer Society
- Bass, F. M. (1969). A new product growth for model consumer durables. *Management Science*, 15(5), 215–227.

- Berger, E. (2001). Dynamic Monopolies of Constant Size. *Journal of Combinatorial Theory Series B*, 83, 191-200.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. & Lafferty, J. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(4-5), 993-1022.
- Brehm, J. W. (1966). A theory of psychological reactance. Academic Press.
- Cai, K. K., Spangler, S., Chen, Y., & Zhang, L. (2008). Leveraging Sentiment Analysis for Topic Detection. In C. Zhang, & N. Cercone (Eds.), *Proceedings of the International Conference on Web Intelligence and Intelligent Agent Technology*, Sydney, Australia (pp. 265-271). NY: IEEE Computer Society.
- Cai, K. K., Bao, S. H., Yang, Z., Tang, J., Ma, R., Zhang, L., & Su, Z. (2011). OOLAM: An Opinion Oriented Link Analysis Model for influence persona discovery. In I. King, W. Nejdl, & H. Li (Eds.), *Proceedings of the 4th International Conference on Web Search and Data Mining*, Hong Kong, China (pp. 645-654). NY: ACM Press.
- Cer, D., Marneffe, M -C., Jurafsky, D., & Manning, C. D. (2010). Parsing to Stanford Dependencies: Trade-offs between speed and accuracy. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner & D. Tapias (Eds.), *Proceedings of the 7th International Conference on Language Resources and Evaluation*, Malta. LREC: European Language Resources Association.
- Chen, W., Wang, C., & Wang, Y. (2010a). Scalable influence maximization for prevalent viral marketing in large-scale social networks. In B. Rao, B. Krishnapuram, A. Tomkins, Q. Yang (Eds.), *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, Washington DC, USA* (pp. 1029-1038). NY: ACM.
- Chen, W., Yuan, Y., & Zhang, L. (2010b). Scalable influence maximization in social networks under the linear threshold model. In G. I. Webb, B. L. 0001, C. Zhang, D. Gunopulos & X. Wu (Eds.), *Proceedings of the 2010 International Conference on Data Mining*, Las Vegas, Nevada, USA (pp. 88-97). NY: IEEE Computer Society.
- Cialdini, R. B. (2001), *Influence: Science and Practice*, 4th edition, Allyn & Bacon, Needham Heights, MA.
- Cialdini, R. B., & Goldstein, N. J. (2004). Social Influence: Compliance and conformity. *Annual Review of Psychology*, 55, 591-621.
- Cialdini, R. B., & Trost, M. R. (1998). Social influence: Social norms, conformity and compliance. In D.T. Gilbert, & S.T. Fiske (Eds.), *The handbook of social psychology* (4th ed.), 2, 151-192. NY: McGraw-Hill.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), pp. 37-46.

- Costa, P. T. Jr., & McCrae, R. R. (1992). Normal personality assessment in clinical practice: The NEO Personality Inventory. *Psychological Assessment*, 4(1), 5-13.
- Cowan, R., & Jonard, N. (2004). Network structure and the diffusion of knowledge. *Journal of Economic Dynamics and Control*, 28(8), 1557 – 1575.
- Ding X., Liu B., & Yu P. S. (2008). A holistic lexicon-based approach to opinion mining. In M. Najork, A. Z. Broder, S. Chakrabarti (Eds.), *Proceedings of the Conference on Web Search and Web Data Mining*, Palo Alto, California, USA (pp.231-240). NY: ACM.
- Dodds, P., & Watts, D. (2004). Universal behavior in a generalized model of contagion. *Physical Review Letters*, 92(21).
- Domingos, P. (2005). Mining social networks for viral marketing. *IEEE Intelligent Systems*, 20(1), 80-93.
- Easingwood, C. J., Mahajan, V., & Muller, E. (1983). A Nonuniform Influence Innovation Diffusion Model of New Product Acceptance, *Marketing Science*, 2, 273-296.
- Esuli, A., & Sebastiani, F. (2006). SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation* (pp. 417-422).
- Fellbaum, C. (1998). Wordnet: An Electronic Lexical Database. Bradford Books.
- Garg, P., King, I., & Lyu, M. R. (2012). Information propagation in social rating networks. In X.W. Chen, G. Lebanon, M. Zaki, & G.X. Wang (Eds), *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, Maui, HI, USA (pp. 2279-2282). NY: ACM.
- Ghosh, R., & Lerman, K. (2008). Community detection using a measure of global influence. In M. M. Gaber, R. R. Vatsavai, O. A. Omitaomu, J. Gama, N. V. Chawla, A. R. Ganguly (Eds.), *Proceedings of the 2nd SNA-KDD Workshop on Social Network Mining and Analysis*, Las Vegas, NV, USA. Springer LNCS.
- Girvan, M., Callaway, D. S., Newman, M. E. J., & Strogatz, S. H. (2002). A simple model of epidemics with pathogen mutation. *Physical Review E*.
- Goldenberg, J., Libai, B., & Muller, E. (2001). Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth. *Marketing Letters*, 12(3), 211-223.
- Gomez, R. M., Leskovec, J., & Krause, A. (2010). Inferring networks of diffusion and influence. In *Proceedings of Transactions on Knowledge Discovery from Data* (pp. 1019–1028). NY: ACM.
- Goyal, A., Bonchi, F. & Lakshmanan, L. V. (2010). Learning influence probabilities in social networks. In B. D. Davison, T. Suel, N. Craswell, B. Liu

- (Eds.), *Proceedings of the third ACM international conference on Web search and data mining*, New York, USA (pp. 241-250). NY: ACM.
- Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, 78(6), 1360–1380.
- Granovetter, M. S. (1978). Threshold models of collective behavior. *American Journal of Sociology*, 83(6), 1420-1443.
- Gruhl, D., Guha, R., Liben-Nowell, D., & Tomkins, A. (2004). Information diffusion through blogspace. In S. Feldman, & M. Uretsky (Eds.), *Proceedings of the 13th International Conference on World Wide Web*, Manhattan, NY, USA (pp. 491-501). NY: ACM.
- Guadagno, R. E., Okdie, B. M., & Eno, C. A. (2008). Who blogs? Personality predictors of blogging. *Computers in Human Behavior*, 24(5), 1993-2004.
- Guha, R., Kumar, R., Raghavan, P. & Tomkins, A. (2004). Propagation of trust and distrust. In S. Feldman, & M. Uretsky (Eds.), *Proceedings of the 13th International Conference on World Wide Web*, Manhattan, NY, USA (pp. 403-412). NY: ACM.
- Hassan, A., Qazvinian, V., & Radev, D. R. (2010). What's with the attitude? A study of participant attitude in multi-party online discussions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Massachusetts, USA (pp. 1245-1255). PA: ACL.
- Hatzivassiloglou, V., & McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In P. R. Cohen, W. Wahlster (Eds.), *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics* (pp. 174-181). PA: ACL.
- Herring, S.C., Scheidt, L.A., Wright, E. & Bonus, S. (2005a). Weblogs as a bridging genre. *Information Technology & People*, 18(2), 142-171.
- Herring, S. C., Kouper, I., Paolillo, J. C., Scheidt, L. A., Tyworth, M., Welsch, P., Wright, E., & Yu, N. (2005b). Conversations in the blogosphere: An analysis “from the bottom up.” In Sprague, R.H. (Ed): *Proceedings of the Thirty-Eighth Hawaii International Conference on System Sciences*, Big Island, HI, USA (pp. 107b). Los Alamitos: IEEE Press.
- Herring, S. C. (2009). Web content analysis: Expanding the paradigm. In Hunsinger, J., Allen, M. & Klastrup, L. (Eds.), *The International Handbook of Internet Research*, Berlin: Springer Verlag.
- Hofmann, T. (2002). Probabilistic latent semantic indexing. In F. Grey, M. Hearst, R. Tong (Eds.), *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval*, Tampere, Finland (pp. 50–57). NY: ACM.

- Hogg, M. A., & Vaughan, G. M. (2005). *Social psychology*. Harlow: Pearson/Prentice Hall.
- Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In W. Kim, R. Kohavi, J. Gehrke, W. DuMouchel (Eds.), *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 168-177). Seattle, WA, USA (pp. 168-177). NY: ACM.
- Iribarren, J. L., & Moro, E. (2011). Affinity paths and information diffusion in social networks. *Social Networks*, 33(2), 134-142.
- Jakob, N., Weber, S. H., Muller, M. C., & Gurevych, I. (2009). Beyond the stars: exploiting free-text user reviews to improve the accuracy of movie recommendations. In B. Yu, M. Jiang (Eds.), *Proceedings of the International Conference on Information and Knowledge Management workshop on TSA*. Hong Kong, China (pp.57-64). NY: ACM.
- Java, A., Kolari, P., Finin, T., & Oates, T. (2006). Modeling the spread of influence on the blogosphere. In E. Adar, N. Glance, M. Hurst (Eds.), *Proceedings of the WWW 2006 Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics*.
- Joshi, M., & Penstein-Rosé, C. (2009). Generalizing dependency features for opinion mining. In K. Y. Su, J. Su, J. Wiebe (Eds.), *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, Singapore (pp. 313-316). PA: ACL.
- Kale, A., Karandikar, A., Kolari, P., Java, A. & Joshi, A. (2007). Modeling trust and influence in the blogosphere using link polarity. In N. S. Glance, N. Nicolov, E. Adar, M. Hurst, M. Liberman, F. Salvetti (Eds.), *Proceedings of the First International Conference on Weblogs and Social Media*, Boulder, Colorado, USA (26-28).
- Kang, K.-H., Lim, S.-H., Kim, S.-W., Jang, M.-H., & Jeong, B.-S. (2012). A topic-oriented analysis of information diffusion in a blogosphere. *Database and Expert Systems Applications, Lecture Notes in Computer Science*, (7447), 129-140.
- Kelman, H.C. (1958). Compliance, identification, and internalization: Three Processes of attitude change. *Journal of Conflict Resolution*, 2(1), 51-60.
- Kempe, D., Kleinberg, J., & Tardos, E. (2003). Maximizing the spread of influence through a social network. In L. Getoor, T. E. Senator, P. Domingos, C. Faloutsos (Eds.), *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Washington, D.C. (pp. 137-146). NY: ACM.
- Kim, S.-M. & Hovy, E. (2006). Identifying and analyzing judgment opinions. In A. Sarkar, C. P. Rose, S. Stoyanchev, U. Germann, C. Shah (Eds.), *Proceedings of*

the Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Boulder, Colorado, USA (pp. 200-207). NJ: ACL.

- Kimura, M., & Saito, K. (2006). Tractable models for information diffusion in social networks. In J. Fürnkranz, T. Scheffer, M. Spiliopoulou (Eds.), *Proceedings of the 10th European conference on Principle and Practice of Knowledge Discovery in Databases* (pp. 259-271). Berlin: Springer-Verlag.
- Krackhardt, D. (1992). The strength of strong ties: The importance of philos in organizations, (216–239). Harvard Business School Press, Boston, MA.
- Kwon, N., Zhou L., Hovy, E., & Shulman, S. W. (2007). Identifying and classifying subjective claims. In J. B. Cushing, T. A. Pardo (Eds.), *Proceedings of the 8th Annual International Conference on Digital Government Research: Bridging Disciplines and Domains* (pp. 76-81). PA: Digital Government Society of North America.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159-174.
- Lee, C., Lee, G. G & Jang, M. (2007). Dependency structure language model for topic detection and tracking. *Information Processing and Management.*, 43(5), 1249-1259.
- Leskovec, J., McGlohon, M., Faloutsos, C., Glance, N., & Hurst, M. (2007a). Cascading behavior in large blog graphs. In: V. Kumar, C. Apte, & D. Skillicorn (Eds.): *Proceedings of the 7th Society for Industrial and Applied Mathematics International Conference on Data Mining*.
- Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., & Glance, N. (2007b). Cost-effective outbreak detection in networks. In P. Berkhin, R. Caruana, X. Wu (Eds.), *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, San Jose, CA, USA (pp. 420-429). NY: ACM.
- Leskovec, J., Huttenlocher, D., & Kleinberg, J. (2010). Predicting positive and negative links in online social networks. In M. Rappa, P. Jones, J. Freire, & S. Chakrabarti (Eds.): *Proceedings of the 19th International Conference on World Wide Web* (pp. 641-650). New York, NY: ACM Press.
- Li, X., & Croft, W. B. (2006). Improving novelty detection for general topics using sentence level information patterns. In V. Tsotras, E. Fox, B. Liu (Eds.), *Proceedings of the 15th ACM international conference on Information and knowledge management*, Arlington, VA, USA (pp. 238-247). NY: ACM.
- Li, H., Bhowmick, S. S., & Sun, A. (2011). CASINO: Towards conformity-aware social influence analysis in online social networks. In B. Berendt, A. Vires, W. Fan, C. Macdonald, I. Ounis, I. Ruthven (Eds.), *Proceedings of the 20th ACM*

- International Conference on Information and Knowledge Management* (pp. 1007-1012). NY: ACM Press.
- Lim, S.-H., Kim S.-W., Kim, S., & Park, S. (2011). Construction of a blog network based on information diffusion. In W. C. Chu, W. E. Wong, M. J. Palakal, C. C. Hung (Eds.), *Proceedings of the 2011 ACM Symposium on Applied Computing* (pp. 937-941). NY: ACM
- Lin, X., Mei, Q., Han, J., Jiang, Y., & Danilevsky, M. (2011). The joint inference of topic diffusion and evolution in social communities. In D. J. Cook, J. Pei, W. Wang, O. R. Zaiane, X. Wu (Eds.), *Proceedings of the International Conference on Data Mining*, Vancouver, Canada (pp. 378-387). DC: IEEE.
- Liu, B. S., Madhavan, R., & Sudharshan, D. (2005). DiffuNET: The impact of network structure on diffusion of innovation. *European Journal of Innovation Management*, 8(2), 240–262.
- Liu, B. (2006) *Web Data Mining: Exploring Hyperlinks, Contents and Usage Data*, (37-54). Berlin: Springer Heidelberg.
- Liu, L., Tang, J., Han, J., Jiang, M., & Yang, S. (2010). Mining topic-level influence in heterogeneous networks. In J. Huang, N. Koudas, G. Jones, X. Wu, K. C. Thompson, A. An (Eds.), *Proceedings of the 19th ACM international conference on Information and knowledge management*, Toronto, Canada (pp. 199-208). NY: ACM.
- Luu, D., Lim, E., Hoang, T., & Chua, F. (2012). Modeling Diffusion in Social Networks Using Network Properties. In J. Breslin, N. Ellison, J. G. Shanahan, Z. Tufekci. (Eds.), *Proceedings of the International AAAI Conference on Weblogs and Social Media* (pp. 218-225). CA: AAAI Press.
- Mackie, D. M. (1987). Systematic and nonsystematic processing of majority and minority persuasive communications. *Journal of Personality and Social Psychology*, 53, 41-52.
- Mahajan, V., Muller, E., & Bass, F. (1995). Diffusion of new products: Empirical generalizations and managerial uses. *Marketing Science* 14(3), 79–88.
- Manning, C. D., & Schutze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.
- Marneffe, M.-C., & Manning, C. D. (2008). The Stanford typed dependencies representation. In A. Copestake (Ed.) *Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation* (pp. 1-8). PA: ACL.
- Matsumura, N., Yamamoto, H. & Tomozawa, D. (2008), Finding influencers and consumer insights in the blogosphere. In E. Adar, M. Hurst, T. Finin, N. S. Glance, N. Nicolov, B. L. Tseng (Eds.), *Proceedings of International Conference on Weblogs and Social Media* (pp. 76-83). CA: AAAI Press

- McPherson, M., Smith-Lovin, L., & Cook, J. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1), 415–444.
- Meade, N., & Islam, T. (2006). Modelling and forecasting the diffusion of innovation - a 25-year review. *International Journal of Forecasting* 22.
- Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology*, 67, 371-378.
- Mitchell T. M. (1997). *Machine Learning* (McGraw Hill, 1997)
- Mohammad, S., Dunne, C., & Dorr, B. (2009). Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. In P. Koehn, R. Mihalcea (Eds.), *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2, Singapore* (pp. 599-608). PA: ACL.
- Moilanen, K., & Pulman, S. (2007). Sentiment Composition. In R. Mitkov, G. Angelova (Eds.), *Proceedings of the Recent Advances in Natural Language Processing*, Borovets, Bulgaria (pp.378-382). PA: John Benjamins.
- Morris, S. (2000). Contagion. *Review of Economic Studies*, 67.
- Na, J.-C. Sui, H., Khoo, C., Chan, S. & Zhou, Y. (2004). Effectiveness of simple linguistic processing in automatic sentiment classification of product reviews. In I. C. McIlwaine (Ed.), *Proceedings of the Conference of the International Society for Knowledge Organization (ISKO)* (pp. 49–54). Berlin: Ergon Verlag.
- Na, J.-C., Thet, T. T., & Khoo, C. (2010). Comparing sentiment expression in movie reviews from four online genres. *Online Information Review*. 33(2), 317-338.
- Nivre, J. (2005). Dependency grammar and dependency parsing. *Technical Report MSI report 05133*, Växjö University: School of Mathematics and Systems Engineering.
- Osman, D., Yearwood, J., & Vamplew, P. (2009). Weblogs for market research: Finding more relevant opinion documents using system fusion. *Online Information Review*, 33(5), 873-888.
- Pang, B., Lee, L. & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine-learning techniques. In J. Haijie, Y. Matsumoto (Eds.), *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing* (pp. 79–86). PA: ACL.
- Pang, B., & Lee, L. (2004). A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In D. Scott, W. Daelemans, M. Walker (Eds.), *Proceedings of the Association for Computational Linguistics*, Barcelona, Spain (pp.271-278). PA: ACL.
- Pang, B. & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In K. Knight, H. T. Ng, K. Oflazer

- (Eds.), *Proceedings of the Association for Computational Linguistics*, Michigan USA (pp. 115–124). PA: ACL.
- Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2), 1-135.
- Polanyi, L., & Zaenen, A. (2006). Contextual valence shifters. In J.G. Shanahan, Y. Qu, Wiebe, J. (Eds.), *Computing Attitude and Affect in Text: Theory and Applications* (pp. 1-10). Berlin: SpringerLink.
- Popescu A. M. & Etzioni, O. (2005). Extracting product features and opinions from reviews. In R. J. Mooney, C. Brew, L. F. Chien, K. Kirchhoff (Eds.), *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, Canada* (pp. 339-346). PA: ACL.
- Prasad, R., Dinesh, N., Lee, A., & Joshi, A. (2006). Annotating attribution in the Penn Discourse TreeBank. In M. Gamon, A. Aue (Eds.), *Proceedings of the COLING Workshop on Sentiment and Subjectivity in Text*, Sydney, Australia (pp. 31-38). PA: ACL.
- Qian, T., Durme, B. V., & Schubert, L. (2009). Building a semantic lexicon of English nouns via bootstrapping. In U. Germann, C. Shah, S. Stoyanchev, C. P. Rose (Eds.), *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Student Research Workshop and Doctoral Consortium, Colorado, USA* (pp. 37-42). PA: ACL.
- Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. Longman.
- Riloff, E., & Wiebe, J. (2003). Learning extraction patterns for subjective expressions. In M. Collins, M. Steedman (Eds.), *Proceedings of the Empirical Methods in Natural Language Processing*, Sapporo, Japan (pp.105-112). PA: ACL.
- Robertson S. E. & Jones K. S. (1976). Relevance weighting of search terms, *Journal of the American Society for Information Science*, 27(3), 129-146
- Ruppenhofer, J., Somasundaran, S., & Wiebe, J. (2008). Finding the Sources and Targets of Subjective Expressions. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, D. Tapias (Eds.), *Proceedings of the International Conference on Language Resources and Evaluation*, Marrakech, Morocco. Paris: ELRA.
- Saito, K., Nakano, R., & Kimura, M. (2008). Prediction of Information Diffusion Probabilities for Independent Cascade Model. *Knowledge-Based Intelligent Information and Engineering Systems_Lecture Notes in Computer Science*, 5179, 67-75.

- Shaikh, M. A. M., Prendinger, H., & Ishizuka, M. (2008). Sentiment Assessment of Text By Analyzing Linguistic Features And Contextual Valence Assignment. *Applications of Artificial Intelligence*, 22(6), 558-601.
- Shaikh, N. I., Rangaswamy, A., & Balakrishnan, A. (2010). Modeling the diffusion of innovations using small-world networks. Working paper, Penn State University.
- Somasundaran, S., & Wiebe, J. (2009). Recognizing Stances in Online Debates. In K. Y. Su, J. Su, J. Wiebe (Eds.), *Proceedings of the Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, Singapore (pp. 226-234). PA: ACL.
- Somasundaran, S., & Wiebe, J. (2010). Recognizing stances in ideological on-line debates. In D. Inkpen, & C. Strapparava (Eds.): *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text* (pp. 116-124). PA: ACL.
- Song, X., Chi, Y., Hino, K. & Tseng, B. (2007). Identifying opinion leaders in the blogosphere. In M. J. Silva, A. H. F. Laender, R. A. Baeza-Yates, D. L. McGuinness, B. Olstad, O. O. Olsen, A. O. Falcao (Eds.), *Proceedings of the 16th ACM Conference on Information and Knowledge Management* (pp. 971-974). NY: ACM.
- Stauffer, D. (2002). Sociophysics: the sznajd model and its applications. *Computer Physics Communications*, 146(1), 93–98.
- Stone P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). The General Inquirer: A Computer Approach to Content Analysis.
- Strapparava, C., Valitutti, A. (2004). WordNet-Affect: an affective extension of WordNet. In T. L. Maria, F. X. Maria, F. Fatima, C. Rute, S. Raquel (Eds), *Proceedings of the 4th International Conference on Language Resources and Evaluation* (pp. 1083-1086). ELRA.
- Tan, L. K. W., Na, J.-C., & Theng, Y.-L. (2011). Influence Detection between Blog Posts through Blog Features, Content Analysis, and Community Identity. *Online Information Review*, 35(3), 425-442.
- Tan L. K. W., Na, J.-C, Theng, Y.-L., & Chang K.-Y. (2012). Phrase-Level Sentiment Polarity Classification Using Rule-Based Typed Dependencies and Complex Phrases Consideration. *Journal of Computer Science and Technology*, 27(3), 650-666.
- Tan L. K. W., & Na, J.-C. (2013). Analysis of Bloggers'Influence Style within Blogs. *Journal of Information Science Theory and Practice*, 1(2), 36-57.
- Tang, J., Sun, J., Wang, C., & Yang, Z. (2009). Social influence analysis in large-scale networks. In J. Elder, F. Fogelman-Soulié, P. Flach & M. Zaki

- (Eds.), *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 807-816). NY: ACM.
- Thet, T. T., Na, J.-C. & Khoo, C. S. G. (2010). Aspect-Based Sentiment Analysis of Movie Reviews on Discussion Boards. *Journal of Information Science*, 36(6), 823-848.
- Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In I. Pierre (Ed.), *Proceedings of the Association for Computational Linguistics* (pp. 417-424). PA: ACL.
- Turney, P. D., & Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems* (TOIS), 21(4), 315–346.
- Weng, J., Lim, E. P., Jiang, J., & He, Q. (2010). TwitterRank: finding topic-sensitive influential twitterers. In B. D. Davison, T. Suel, N. Craswell, B. Liu (Eds.), *Proceedings of the third ACM international conference on Web search and data mining* New York, USA (pp. 261-270). NY: ACM.
- Wiebe, J. (1994). Tracking point of view in narrative. *Computational Linguistics*, 20(2), 233-287.
- Wilson, T., Wiebe, J. & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In R. J. Mooney, C. Brew, L. F. Chien, K. Kirchhoff (Eds.), *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Vancouver, Canada (pp. 347-354). PA: ACL.
- Wilson, T., Wiebe, J., & Hwa, R. (2006). Recognizing Strong and Weak Opinion Clauses. *Computational Intelligence*, 22(2), 73-99.
- Wilson, T., Wiebe, J., & Hoffmann, P. (2009). Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3), 399-433.
- Xiang, R., Neville, J., & Rogati, M. (2010). Modeling relationship strength in online social networks. In M. Rappa, P. Jones (Eds.), *Proceedings of the 19th international conference on World Wide Web, Raleigh, NC, USA* (pp. 981-990). NY: ACM.
- Yang, C., Lin, K. H. Y., & Chen, H. H. (2007). Building emotion lexicon from weblog corpora. In S. Ananiadou (Ed.), *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions* (pp. 133-136). PA: ACL.
- Yang, J., & Leskovec, J. (2010). Modeling Information Diffusion in Implicit Networks. In G. I. Webb, B. Liu, C. Zhang, D. Gunopulos, X. Wu (Eds.), *Proceedings of the 2010 IEEE International Conference on Data Mining, Sydney, Australia* (pp. 599-608). DC: IEEE Computer Society.

- Yarkoni, T. (2010). Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, 44(3), 363-373.
- Yi, J., Nasukawa, T., Bunescu, R., & Niblack, W. (2003). Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques. In J. Shavlik, X. Wu, A. Tuzhilin (Eds.), *Proceedings of the 2003 IEEE International Conference on Data Mining, Florida, USA* (pp. 427-434). DC: IEEE Computer Society.
- Zafarani, R., & Liu, H. (2009). Social Computing Data Repository at ASU Arizona State University, School of Computing, Informatics and Decision Systems Engineering.
- Zhang, C., Zeng, D., Li, J., Wang, F.-Y. & Zuo, W. (2009). Sentiment analysis of Chinese documents: from sentence to document level. *Journal of the American Society for Information Science and Technology*, 60(12), 2474-2487.

APPENDIX A: Additional samples of terms found in lexicon

Examples of phrases in subjective phrase lexicon.

Subjective Phrases	Polarity
More or less	Neutral
On par	Neutral
Open source	Neutral
First person	Neutral
New york	Neutral
Hour long	Neutral
Game plan	Neutral
Give and take	Neutral
Go hand in hand	Neutral
Fed up	Negative
Can only hope	Negative
Set you back	Negative
Pulled off	Negative
Run rampant	Negative
Rip off	Negative
Smells fishy	Negative
Green with envy	Negative
Greenhorn	Negative
Back seat	Negative
Cutting edge	Positive
Ground breaking	Positive
Stand the test	Positive
In demand	Positive
Fell off the chair	Positive
In the same league	Positive
Red carpet	Positive
Ahead of the pack	Positive
Get the picture	Positive

Examples of terms found in modifier lexicons.

S/N	Modifier Types	Terms
1.	Intensifiers	Many, more, massive, big, increasing, increase, strong, so, deeply, profoundly, enormously, tremendously, hugely, staggeringly, far, greatly, heartily, cordially, warmly, highly, intensely, much, practically, severely, hard, so, strongly, powerfully, terribly, awfully, easily, lots, widely quite, etc.
2.	Maximisers	Strikingly, unquestionably, unimpeachably, decidedly, emphatically, definitely, absolutely, altogether, wholly, all, whole, completely, entirely, exclusively, solely, only, extremely, highly, exceedingly, super, fully, amply, perfectly, strong, unquestionably, most, utterly, thoroughly, exhaustively, totally, very, least of all, etc.
3.	Minimisers	Little, small, trivial, scarcely, never, in the least, even a little, the least bit, in the slightest, at all, a bit, etc.
4.	Mitigators	Few, small, minimal, decreasing, decrease, weak, possibly, still, rather, sufficiently, slightly, somewhat, mildly, gently, partially, partly, rather, slightly, only, merely, simply, just, kind of, sort of, all but, in some respects, to some extent, a bit, a little, a trifle, etc.
5.	Negators	Different, else, neither, unlikely, seldom, versa, no, versa, not, non, none, nor, unlikely, vice, it's far from, on the other hand, etc.

Aspect and Features Examples

Aspect	Features
Connectivity	Reception, Network, Wifi, Thunderbolt, etc.
Camera	Cameras, Photos, Lens, Images, etc.
Design	Casing, Length, Body, Size, Enclosure, Chassis, Interface, etc.
Support	Issues, Break Downs, Crash, Malfunction, etc.
Display	Screen, Screen Size, Graphics, Pixel, HD, Retina Display, etc.
Apps	Safari, iTunes, Siri, Facebook, Twitter, etc.
OS	IOS, Lion, Android, Jellybean, Ice Cream Sandwich, IOS 4, IOS 5, IOS 6, etc.
Sound	Earphones, Music, MP3, MP4, Acoustic, Digital music, etc.
Hardware	Rams, Drives, Components, Processors, Chips, A5, A6, Batteries, etc.

APPENDIX B: Software Specifications

SPSS Statistical Software

SPSS Statistics version 17.0

Developed by:
SPSS Inc. (Acquired by IBM in 2009)

Application: Used in research study for analyzing correlation between blog features and influence diffusion.

The Stanford NLP Toolkit

Stanford Parser version 2.0.3

Developed by:
The Natural Language Processing Group at Stanford University

Application: Parsing of sentences to generate typed dependencies of words.

Python Programming Language

Python version 2.7.3

Developed by:
Python Software Foundation

Application: Coding and development of sentiment analysis model, influence style analysis model, and influence diffusion detection model.

APPENDIX C: Sample coding for typed dependencies rules.

Appended below is the sample coding for the AMOD rule stated in Table 4.19

```
### AMOD rule ###  
  
if typedDependency == 'amod':  
  
    ### Positive D and Positive G ###  
    if FD>=0 and FG>=0:  
        outputPolarity = G+((1-G)*D)  
  
    ### Negative D and Positive G ###  
    if FD<0 and FG>=0:  
        outputPolarity = FD  
  
    ### Positive D and Negative G ###  
    if FD>=0 and FG<0:  
        outputPolarity = -1.0*(G+((1-G)*D))  
  
    ### Negative D and Negative G ###  
    if (FD<0 and FG<0):  
        outputPolarity = -1.0*(G+((1-G)*D))
```

Appended below is the sample coding for the ADVMOD rule stated in Table 4.20

```
### ADVMOD ###  
  
if typedDependency == 'advmod':  
  
    ### Positive D and Positive G ###  
    if FD>=0 and FG>=0:  
        outputPolarity = G+((1-G)*D)  
  
    ### Negative D and Positive G ###  
    if FD<0 and FG>=0:  
        outputPolarity = -1.0*D  
  
    ### Positive D and Negative G ###  
    if FD>=0 and FG<0:  
        outputPolarity = -1.0*(G+((1-G)*D))  
  
    ### Negative D and Negative G ###  
    if (FD<0 and FG<0):  
        outputPolarity = -1.0*(G+((1-G)*D))
```

APPENDIX D: Bloggers Influence Network

Appended diagram below shows the influence network of the 107 bloggers found in the blog dataset used in the influence diffusion detection study. Maximum path length is 6 nodes and less than 10 bloggers-chain paths exceed a length of 5 nodes, indicating the sparseness of interconnected bloggers.

Figure D1. Bloggers Influence Network.

APPENDIX E: Samples of manually coded data for influence diffusion detection.

Influence Example

Blog posts tagged as having influence flow based on the similar negative sentiments expressed on the common “Apple” topic.

Linking blog post	Linked blog post
This isn't the first time that Apple co-founder Steve Wozniak has given the company a hard time, and it probably won't be his last, but today the co-founder said that Apple has fallen somewhat behind in the smartphone market with their iPhone, which is allowing competitors - namely Samsung - to catch up. The comments come from an interview with German magazine, in which Wozniak discusses the importance of branding. Wozniak said that Apple is "somewhat behind with features in the smartphone business," and "others have caught up." Specifically, Wozniak mentions Samsung, who notes that the company is "currently making great products." However, Wozniak is known for standing in line for every single Apple product launch just like any other customer, and he says that if Apple "made lousy products," he wouldn't be standing in line, which essentially means that while Apple still makes great products, other companies are slowly catching up to the Cupertino-based company. Of course, all of this was translated from German, and Wozniak has been misquoted in the past due to translation errors from other foreign publications, but these quotes specifically seem pretty cut and dry. Then again, Wozniak isn't afraid to speak the truth on what he thinks about his former company, so his thoughts certainly aren't surprising. [via MacRumors]	Ben Lovejoy Apple co-founder Steve Wozniak said in an interview yesterday that Apple has fallen somewhat behind in the smartphone field allowing competitors like Samsung to catch up. The comment was made as Wozniak spoke about the importance of brand (translated from the original German report by wiwo.de) at the Best Brand Awards. I am proud that we have such loyal fans. But this loyalty is not given the need to have the best products is always there. Currently we are in my opinion somewhat behind with features in the smartphone business. Others have caught up. Samsung is a big competitor. But precisely because they are currently making great products. The comments were however made in the context of his great pride in Apple and its products. Wozniak also explained why he stands in line at Apple stores to buy new products on launch days despite remaining an employee of the company to this day. The launch of a new Apple product is for me like a big major concert which you absolutely must attend personally. It's history. Sure I could order online or call the store manager so that they smuggled me in the back door. But that would be nothing. No I will have the experience like any normal customer. But one thing is certain. If Apple made lousy products I would not be in line. Wozniak has never been shy in speaking about Apple and while he has only very limited contact with the company today he still draws significant attention at his appearances at events at Apple product launches. While he remains a fan of Apple and the iPhone he has long been open to alternatives and has noted several times that he carries a number of iPhones and Android devices with him.

Non-influence Example

Blog posts tagged as non-influence based on the dissimilar sentiments expressed on the common “iOS App Store” topic.

Linking blog post (Negative)	Linked blog post (Neutral)
Apple has removed the offending apps, the company told me via email. Holiday downloaders are now safe from this attempted scam. A couple of new apps that just hit the iOS App Store ahead of the annual App Store holiday freeze are not what they appear to be. Masquerading as official Halo 4 ports for Apple's mobile devices, these are actually a pretty lame racing game and an adequate chess simulator, which have somehow managed to be reborn with new names, descriptions and screenshots, but appear to have kept their reviews intact. The resulting illusion is of highly rated Halo 4 iOS titles (though one-star reviews are starting to percolate as people get wise), one universal and one iPhone-specific, with price tags that, while a steal were they actually Halo 4, are actually stealing from you instead. Nothing in the description of either of these games would lead anyone to believe that these aren't official ports (besides a few instances of dodgy sentence structure), and each in fact advertises iOS-specific features like Game Center integration. The illusion even permeates down to the developer and support links for each app, which takes you to a page for Halo 4 that even claims a 343 Industries copyright at the bottom. That's not like the scams we've seen in the past where developers have provided so-called "guides" for Minecraft that lift assets and look like the game itself in screenshots and description, but have a small disclaimer noting they aren't actually a game. This is a full-blown scam, without any possible defense, and it's a shame to see it hit the App Store right when downloads usually pop over the holidays, and when Apple is in a position where it might not be able to do anything to reverse it until the freeze lifts December 28. I've opted not to link them here because I'd rather not contribute to their heinous plot, but check out AppShopper if you're curious and want to see the elaborate ruse in full . One thing's for sure: whoever Dung Bui and Toan Tran (the developers listed for these apps) are, they won't be getting back on the App Store anytime soon once Apple takes steps to resolve this mess. Speaking of resolutions, we've reached out to Apple to let them know about the offending apps and will update if we hear back. For now though, spread the word, since friends don't let friends buy fake iOS Halo games	Apple has sent reminders to App Store developers that the annual shutdown of iTunes Connect is scheduled to take place from Friday, December 21, 2012 to Friday, December 28th, 2012. iTunes Connect is the portal for developers to submit app updates and changes. As a result, no new apps, app updates or price changes will be reflected in the App Store during that time. We strongly recommend that you do not schedule any pricing changes in iTunes Connect that would take effect between Friday, December 21, 2012 and Friday, December 28, 2012. Pricing changes scheduled to take effect during this date range will not be reflected on the App Store and your app or In-App Purchase will become unavailable for purchase until after the shutdown. We also recommend that you do not schedule any apps to go live during the shutdown. Releases scheduled with a start date between Friday, December 21, 2012 and Friday, December 28, 2012 will not go live on the App Store until after the shutdown. As a result of the shutdown deadline, a large number of App developers have been cutting prices today as they compete for top ranking on the App Store Top Charts during the lucrative Christmas sales period.

APPENDIX F: Instructions given to coders for inter-coder reliability testing.

Appended is a case sample of the test to be conducted. Each tester would be required to read through both the linking blog post and linked blog post and (Step 1) to tag the influence (YES/NO) flow from the linked blog post towards the linking blog post. (Step 2) to rank the features by their ability to indicate influence flow for the given features values. Each case is estimated to take about 20mins to conduct.

The following table explains how the blog post level and blog level features were measured:

S/N	Features	Definition
Blog post level data between the linked and linking blog posts		
1.	<i>Sentiment values between the two blog posts</i>	Tally of the positive or negative sentiments expressed on the common topics discussed by each linked and linking posts. A same resultant polarity would indicate same sentiment and conversely, different polarity indicates different sentiment. A balanced polarity means neutral or unknown sentiment value.
2.	<i>Agreement on linked blogger by linking blogger</i>	Agreement expressed on the linked blogger by linking blogger in the linking posts content.
3.	<i>Copied Sentences count</i>	
	(a) <i>Number of sentences copied from linked post to linking post</i>	The number of sentences explicitly copied from the linked post by the linking blogger on the linking post.
	(b) <i>Copied sentence ratio in linked blog post / linking blog post</i>	Number of copied sentences divided by number of sentences found in each post.
4.	<i>Time difference between postings of linked and linking blog</i>	The difference in timing between the linked and linking blog.
Blog level data between the linked and linking blogs		
1.	<i>Similar sentiment ratio between linked and linking blog</i>	Total number of similar sentiment posts divided by the total number of posts excluding the neutral posts.
2.	<i>Agreement ratio of linking blog on linked blog</i>	Total number of agreement posts divided by the total number of posts excluding the neutral cases.
3.	<i>Average copied sentences over total number of sentences.</i>	Total copied sentences divided by the total number of sentences found between the two linked blogs.
4.	<i>Average time difference over all the cases</i>	Total time difference in postings between the two linked blogs divided by the total number of cases.
5.	<i>Number of cases where authority of linked blog is greater than linking blog</i>	Number of posts where Auth(Linked Blog) > Auth(Linking Blog) divided by the number of cases.
6.	<i>Community value between the two blog posts</i>	If the two blog posts are in the same community, the community value is tagged "same", otherwise "different".

Case: 1RS-TP10

POST A (Linking blog post)

06.24.09 -- 12:57PM // [link](#) | [RECOMMEND \(56\)](#)

From the Alternate Universe

Erick Erickson, writing yesterday at [RedState](#):

First, we need to be clear on the facts -- not the media speculation:

-- Sanford did tell his staff and family where he was going.

-- Because he was traveling without a security detail, it was in his best interests that no one knew he was gone.

-- His political enemies -- Republicans at that -- ginned up the media story.

-- When confronted by a pestering media, things went downhill.

-- Again though, at all times there was no doubt that Sanford's staff and family knew where he was.

(Thanks to TPM Reader *DK* for the tip.)

--David Kurtz

POST B (Linked blog post)

The Lessons of Mark Sanford's Hike

Posted by [Erick Erickson](#) ([Profile](#))

Tuesday, June 23rd at 10:23AM EDT

102 Comments

Sanford's lies spread through his office and out to the rest of us. The left is linking to this post to laugh at it. What they are missing is that most of us tend to give people the benefit of the doubt — even people like John Edwards. We live in a fallen world and we ourselves are fallen. I am disappointed in Sanford, but not angry. The default for politicians seems to be unchaste. All we can do is work for ideas and try to find men of good character to fight for those ideas. I think one thing I have noticed in the past five years is that Democrats and Republicans tend to elevate politicians to such a level that there is no accountability. It is insular. There is no support group, no small group of friends, and no authority that can guide, admonish, and correct politicians privately. We have each other. I'm off in a bit to hang out with friends from my Bible Study. Sanford probably has none of that. I'm sure John Edwards did not. Nor Bill Clinton. Nor John Ensign. What Mark Sanford did was wrong. He needs to go in a dark hole somewhere where no one can see him or hear him and rehabilitate himself. On the bright side, I doubt his indiscretions will affect the FisCon movement. The left is going to spend the next week making Sanford into the second coming of James Dobson to smear real marriage advocates and social conservatives — positions Sanford was rarely vocal on. Blessed is the Lord God Jehovah who brings forth bread from heaven, water from rocks, and men like Mark Sanford from the dust of the earth. His will be done.

STEP 1

- Please read the content of the Linking blog post and Linked blog post.
- Is the Linking Blogger influenced by the Linked Blogger? (YES) / (NO)

STEP 2 (Machine generated attributes provided)

- a) Please tag in order of priority for which the listed features could detect influence between the linked blog posts.

	Attribute Value	Priority
Post Level		
Sentiment values between the two blog posts	Different	
Agreement on linked blogger by linking blogger	No	
Number of sentences copied from linked post to linking post	6	
Copied sentence ratio in linked blog post / linking blog post	0.171	
Time difference between postings of linked and linking blog	24hrs	
Blog Level		
Similar sentiment ratio between linked and linking blog	0	
Agreement ratio of linking blog on linked blog	0	
Average copied sentences over total number of sentences.	0.0357	
Average time difference over all the cases	23.25	
Number of cases where authority of linked blog is greater than linking blog	No	
Community value between the two blog posts	Different	

APPENDIX G: Samples of Clause-Level Typed Dependencies

Table G1. Categorization of Clause-Level Typed Dependencies (Marneffe and Manning, 2008)

S/N	Typed Dependency	Definition	Examples
Adverbial Clause-Level			
1.	advcl: adverbial clause modifier	An adverbial clause modifier of a VP or S is a clause modifying the verb (temporal clause, consequence, conditional clause, etc.).	“The accident happened as the night was falling” advcl(happened, falling) “accident happened” is the verb phrase, while “night was falling” is the adverbial clause.
Subject Clause-Level			
2.	nsubj: nominal subject	A nominal subject is a noun phrase which is the syntactic subject of a clause.	“Clinton defeated Dole” nsubj (defeated, Clinton)
3.	nsubjpass: passive nominal subject	A passive nominal subject is a noun phrase which is the syntactic subject of a passive clause.	“Dole was defeated by Clinton” nsubjpass(defeated, Dole)
4.	csubj: clausal subject	A clausal subject is a clausal syntactic subject of a clause, i.e. the subject is itself a clause.	“What she said makes sense” csubj (makes, said) “what she said” is the subject.
5.	csubjpass: clausal passive subject	A clausal passive subject is a clausal syntactic subject of a passive clause.	“That she lied was suspected by everyone” csubjpass(suspected, lied) “that she lied” is the subject.
6.	rel: relative	A relative of a relative clause is the head word of the WH-phrase introducing it. This analysis is used only for relative words which are not the subject nor the object of the relative clause. Relative words which act as the subject of a relative clause are analyzed as a nsubj, relative words which acts as the object of a relative clause are analyzed as a dobj.	“I saw the man whose wife you love” rel (love, wife)

Table G1. Categorization of Clause-Level Typed Dependencies (Marneffe and Manning, 2008) (cont.)

S/N	Typed Dependency	Definition	Examples
7.	xcomp: open clausal complement	An open clausal complement (xcomp) of a VP or an ADJP is a clausal complement without its own subject, whose reference is determined by an external subject.	“He says that you like to swim xcomp(like, swim)
8.	xsubj: controlling subject	A controlling subject is the relation between the head of an open clausal complement (xcomp) and the external subject of that clause.	“Tom likes to eat fish” xsubj (eat, Tom)
Verbial Clause- Level			
9.	aux: auxiliary	An auxiliary of a clause is a non-main verb of the clause, e.g. modal auxiliary, “be” and “have” in a composed tense.	“He should leave” aux(leave, should)
10.	auxpass: passive auxiliary	A passive auxiliary of a clause is a non-main verb of the clause which contains the passive information.	“Kennedy has been killed” auxpass(killed, been) aux(killed,has)
11.	parataxis: parataxis	The parataxis relation (from Greek for “place side by side”) is a relation between the main verb of a clause and other sentential elements, such as a sentential parenthetical, or a clause after a “:” or a “;”.	“The guy, John said, left early in the morning” parataxis(left , said)
12.	ccomp: clausal complement	A clausal complement of a verb or adjective is a dependent clause with an internal subject which functions like an object of the verb, or adjective.	“He says that you like to swim” ccomp(says, like) “I am certain that he did it” ccomp(certain, did)
13.	rmod: relative clause modifier	A relative clause modifier of an NP is a relative clause modifying the NP. The relation points from the head noun of the NP to the head of the relative clause, normally a verb.	“I saw the book which you bought” rmod(book,bought)
Clause Separator			
14.	complm: complementizer	A complementizer of a clausal complement (ccomp) is the word introducing it. It will be the subordinating conjunction “that” or “whether”.	“He says that you like to swim” complm(like, that)

Table G1. Categorization of Clause-Level Typed Dependencies (Marneffe and Manning, 2008) (cont.)

S/N	Typed Dependency	Definition	Examples
15.	conj: conjunct	A conjunct is the relation between two elements connected by a coordinating conjunction, such as \and", \or", etc.	“Bill is big and honest” conj (big, honest)
16.	mark: marker	A marker of an adverbial clausal complement (advcl) is the word introducing it. It will be a subordinating conjunction different from “that” or “whether”: e.g. “because”, “when”, “although”, etc.	“Forces engaged in fighting after insurgents attacked” mark(attack, after)
17.	pcomp: prepositional complement	The prepositional complement of a preposition is the head of a clause following the preposition.	“We have no information on whether users are at risk” pcomp(on, are)
18.	pobj: object of a preposition	The object of a preposition is the head of a noun phrase following the preposition, or the adverbs “here” and “there”.	“I sat on the chair” pobj (on, chair)
19.	prep: prepositional modifier	A prepositional modifier of a verb, adjective, or noun is any prepositional phrase that serves to modify the meaning of the verb, adjective, or noun.	“I saw a cat in a hat” prep(cat, in)
20.	prepc: prepositional clausal modifier	a prepositional clausal modifier of a verb, adjective, or noun is a clause introduced by a preposition which serves to modify the meaning of the verb, adjective, or noun.place.	“He purchased it without paying a premium” prepc without(purchased, paying)
21.	purpcl: purpose clause modifier	A purpose clause modifier of a VP is a clause headed by (in order) to specifying a purpose. At present the system only recognizes ones that have “in order to”. It can also recognize “to” clauses introduced by “be VBN”.	“He talked to him in order to secure the account” purpcl(talked, secure)
22.	ref: referent	A referent of the head of an NP is the relative word introducing the relative clause modifying the NP.	“I saw the book which you bought” ref (book, which)

APPENDIX H: Experiments to determine the IDD-IS-Sent (IC) and IDD-IS-Sent (LT) models' threshold values.

The scores for the IDD-IS-Sent (IC) and IDD-IS-Sent (LT) models at respective threshold values are shown in Figure H1 where it can be seen that IDD-IS-Sent (LT) has an overall higher recall value. The corresponding threshold values for the highest F1-scores of the IDD-IS-Sent (IC) model and IDD-IS-Sent (LT) model are 0.025 and 0.015 respectively. These numbers were subsequently used as the threshold values in the evaluation of the models based on the best performance given for the models. The threshold values range of {0.01, 0.1} was chosen because there was no significant improvement in the F1-scores of the models beyond this range after extensive experiments.

Figure H1. Scores for the IDD-IS-Sent (IC) and IDD-IS-Sent (LT) model at respective threshold values

APPENDIX I: List of Publications

Blog Features Analysis

- Tan, L. K. W., Na, J.-C., and Theng, Y.-L. (2011). Influence Detection between Blog Posts through Blog Features, Content Analysis, and Community Identity. *Online Information Review*, 35(3), 425-442.
- Tan L. K. W. (2012). Influence Detection Within The Blogosphere. In H. H. Chen, G. Chowdhury (Eds.), *Proceedings of the International Conference on Asia-Pacific Digital Libraries Workshop on Global Collaboration of Information Schools, Doctoral Student Consortium*, Taipei, Taiwan. Berlin: Springer.

Sentiment Analysis

- Tan L. K. W., Na, J.-C, Theng, Y.-L., Chang K.-Y. (2011). Phrase-Level Sentiment Polarity Classification Using Rule-Based Typed Dependencies. In F. Diaz, E. Hovy, I. King, J. Li, D. Metzler, M. F. Moens, J. Tang, L. Zhang (Eds.), *Proceedings of the ACM SIGIR 3rd Workshop on Social Web Search and Mining*, Beijing, China. NY: ACM.
- Tan L. K. W., Na, J.-C, Theng, Y.-L., Chang K.-Y. (2011). Sentence-Level Sentiment Polarity Classification Using a Linguistic Approach. In C. Xing, F. Crestani, A. Rauber (Eds.), *Proceedings of the International Conference on Asia-Pacific Digital Libraries*, Beijing, China (pp. 77-87). Berlin: Springer.
- Tan L. K. W., Na, J.-C, Theng, Y.-L., Chang K.-Y. (2012). Phrase-Level Sentiment Polarity Classification Using Rule-Based Typed Dependencies and Complex Phrases Consideration. *Journal of Computer Science and Technology*, 27(3), 650-666.

Influence Style Analysis

- Tan L. K. W., Na, J.-C, Theng, Y.-L., Chang K.-Y. (2012). Blog Site Profiling through Influence Style Detection. In H. H. Chen, G. Chowdhury (Eds.), *Proceedings of the International Conference on Asia-Pacific Digital Libraries*, Taipei, Taiwan (pp. 329-332). Berlin: Springer.
- Tan L. K. W., & Na, J.-C. (2013). Analysis of Bloggers' Influence Style within Blogs. *Journal of Information Science Theory and Practice*, 1(2), 36-57.
- Tan L. K. W., Na, J.-C, Theng, Y.-L. (2013). Influence Diffusion Detection Using Blogger's Influence Style. In S. R. Urs, J.-C. Na, G. Buchanan (Eds.), *Proceedings of the International Conference on Asia-Pacific Digital Libraries*, Bangalore, India (pp. 132-142). Berlin: Springer.
- Tan L. K. W., Na, J.-C, Ding Y. (2014). Influence Diffusion Detection Using the Influence Style Model. *Journal of the American Society for Information Science and Technology*. (Accepted Mar 2014).