


December 2014

# Three Essays on Opinion Mining of Social Media Texts

Shuyuan Deng

*University of Wisconsin-Milwaukee*

Follow this and additional works at: <https://dc.uwm.edu/etd>

 Part of the [Databases and Information Systems Commons](#), and the [Finance and Financial Management Commons](#)

---

## Recommended Citation

Deng, Shuyuan, "Three Essays on Opinion Mining of Social Media Texts" (2014). *Theses and Dissertations*. 679.  
<https://dc.uwm.edu/etd/679>

This Dissertation is brought to you for free and open access by UWM Digital Commons. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of UWM Digital Commons. For more information, please contact [open-access@uwm.edu](mailto:open-access@uwm.edu).

**THREE ESSAYS ON OPINION MINING  
OF SOCIAL MEDIA TEXTS**

by

Shuyuan Deng

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Management Science

at

The University of Wisconsin-Milwaukee

December 2014

**ABSTRACT**

**THREE ESSAYS ON OPINION MINING**

**OF SOCIAL MEDIA TEXTS**

by

Shuyuan Deng

The University of Wisconsin-Milwaukee, 2014

Under the Supervision of Professor Atish Sinha and Professor Huimin Zhao

This dissertation research is a collection of three essays on opinion mining of social media texts. I explore different theoretical and methodological perspectives in this inquiry. The first essay focuses on improving lexicon-based sentiment classification. I propose a method to automatically generate a sentiment lexicon that incorporates knowledge from both the language domain and the content domain. This method learns word associations from a large unannotated corpus. These associations are used to identify new sentiment words. Using a Twitter data set containing 743,069 tweets related to the stock market, I show that the sentiment lexicons generated using the proposed method significantly outperforms existing sentiment lexicons in sentiment classification. As sentiment analysis is being applied to different types of documents to solve different problems, the proposed method provides a useful tool to improve sentiment classification.

The second essay focuses on improving supervised sentiment classification. In previous work on sentiment classification, a document was typically represented as a collection of single words. This method of feature representation suffers from severe ambiguity, especially in classifying short texts, such as microblog messages. I propose the use of dependency features in sentiment classification. A dependency describes the relationship between a pair of words even when they are distant. I compare the sentiment classification performance of dependency features with a few commonly used features in different experiment settings. The results show that dependency features significantly outperform existing feature representations.

In the third essay, I examine the relationship between social media sentiment and stock returns. This is the first study to test the bidirectional effects in this relationship. Based on theories in behavioral finance research, I speculate that social media sentiment does not predict stock return, but rather that stock return predicts social media sentiment. I empirically test a set of research hypotheses by applying the vector autoregression (VAR) model on a social media data set, which is much larger than those used in previous studies. The hypotheses are supported by the results. The findings have significant implications for both theory and practice.

© Copyright by Shuyuan Deng, 2014  
All Rights Reserved

# TABLE OF CONTENTS

CHAPTER 1 .....	1
CHAPTER 2 .....	5
Essay 1: Adapting Sentiment Lexicons to Domain-Specific Social Media Texts.....	5
2.1. Introduction.....	5
2.2. Literature Review.....	9
2.3. Proposed Method .....	13
2.4. Evaluation .....	21
2.5. Conclusion and Future Directions .....	28
CHAPTER 3 .....	29
Essay 2: Resolving Ambiguity of Sentiment -The Quest for Sentiment Indicators .....	29
3.1. Introduction.....	29
3.2. Literature Review.....	31
3.3 Method .....	36
3.4 Evaluation .....	37
3.5 Conclusion .....	54
CHAPTER 4 .....	56
Essay 3: To predict or to be predicted? An Empirical Study on Social Media Sentiment and Stock Return.....	56
1. Introduction.....	56
2. Literature Review.....	60
3. Theoretical Background and Hypotheses .....	64
4. Data and Measurement .....	66
5. Model Specification.....	69
6. Results.....	70
7. Discussion.....	77
8. Conclusion .....	81
REFERENCES .....	82

## LIST OF FIGURES

<b>Figure 1. The Proposed Lexicon Generation Method .....</b>	<b>16</b>
<b>Figure 2. Procedure for Recognizing Sentiment Words from Candidates.....</b>	<b>20</b>
<b>Figure 3. Sentiment Classification using Dependency Features.....</b>	<b>36</b>
<b>Figure 4. IRF Plot of Model 1 .....</b>	<b>72</b>
<b>Figure 5. IRF Plot for Model 2 .....</b>	<b>75</b>

## LIST OF TABLES

<b>Table 1. Agreements and Conflicts between Existing Lexicons .....</b>	<b>23</b>
<b>Table 2. Positive and Negative Entries in Existing Lexicons .....</b>	<b>24</b>
<b>Table 3. Experiment Results .....</b>	<b>26</b>
<b>Table 4. Representations of Sentence 1 .....</b>	<b>33</b>
<b>Table 5. Unigram + Dependency vs. Unigram + Bigram .....</b>	<b>40</b>
<b>Table 6. Unigram + Dependency vs. Unigram + Bigram (Feature Selected) .....</b>	<b>42</b>
<b>Table 7. 1G+DEP vs. 1G+2G+3G .....</b>	<b>44</b>
<b>Table 8. 1G+DEP vs. 1G+2G+3G (Feature Selected) .....</b>	<b>46</b>
<b>Table 9. 1G+DEP vs. POS .....</b>	<b>47</b>
<b>Table 10. 1G+DEP vs. POS (Feature Selected) .....</b>	<b>49</b>
<b>Table 11. Dependency vs. Bi-gram .....</b>	<b>51</b>
<b>Table 12. DEP vs. 2G (Feature Selected) .....</b>	<b>53</b>
<b>Table 13. Related Studies on Social Media Sentiment and Stock Return .....</b>	<b>60</b>
<b>Table 14. Summary Statistics of Variables (N = 1259) .....</b>	<b>68</b>
<b>Table 15. Coefficients of Model 1 .....</b>	<b>71</b>
<b>Table 16. Granger Causality Test for Model 1 .....</b>	<b>71</b>
<b>Table 17. FEVD for Model 1 (in %) .....</b>	<b>71</b>
<b>Table 18. Coefficients of Model 2 .....</b>	<b>74</b>
<b>Table 19. Granger Causality Test for Model 2 .....</b>	<b>74</b>
<b>Table 20. FEVD for Model 2 (in %) .....</b>	<b>74</b>



## ACKNOWLEDGEMENTS

I cannot express enough thanks to my dissertation committee co-chairs, Dr. Atish Sinha and Dr. Huimin Zhao, for their tremendous help and encouragement over these years. I have learned a great deal from them, not only in conducting research but also in pursuing an academic career. I will always be grateful for their interest in seeing me succeed.

I would also like to recognize the continuous guidance Dr. Zhijian Huang provided on my research. I am also thankful to Dr. Srite and Dr. France for their invaluable constructive criticism and illuminating views. I would also like to extend my thanks to my fellow doctoral students, in particular Dr. Austin Kwak, for his help and encouragement. My special thanks also go to StockTwits for the data they provided.

Finally, to my caring and supportive wife, Jiao Wu: my deepest gratitude. It was a great comfort and relief when she was looking after the children. Thanks to my parents-in-law as well. This dissertation would take a much longer time to complete without their selfless caring for my family. I would also like to acknowledge the financial and emotional support from my parents, who encouraged me to pursue a Ph.D. degree at the first place and have always been there when I needed them.

# CHAPTER 1

## INTRODUCTION

Social media refers to the virtual places where people share information with their connections. Wikipedia categorizes social media into six categories (Wikipedia, 2014):

1. Collaborative projects (e.g., Wikipedia)
2. Blogs and microblogs (e.g., Twitter)
3. Content communities (e.g., YouTube)
4. Social networking sites (e.g., Facebook)
5. Virtual game-worlds (e.g., World of Warcraft)
6. Virtual social worlds (e.g., Second Life)

Many social media websites have been operating for less than a decade. Yet, the use of social media has already become the number one activity on the Internet over the last a few years. Twitter, for instance, generated 340 million messages and handled 1.6 billion search queries per day by 2012. Facebook had 1.28 billion monthly active users by March, 2014 (Wikipedia, 2014). In early 2012, YouTube reported 4 billion videos streamed per day. The vast amount of user-generated contents on social media has turned out to be one of the most important types of “big data” businesses are pursuing today. For businesses, it is always important to stay aware of the changing commercial environment. Social media captures every bit of the social-economic change that is of interest to both businesses and their customers.

How to effectively and efficiently extract useful information from social media? This remains a challenging question in the frontier of business analytics. Social media data are

typically unstructured. Finding structure from social media data requires advanced techniques from natural language processing and information retrieval. Associating social media data with business activities also requires advanced text mining techniques. Yet, there is no one-size-fits-all solution for the question raised at the beginning of the paragraph. One reason is that business problems are highly domain dependent. The types of useful information vary across different business domains. Another reason is that social media texts are also domain dependent. Extracting relevant information from relevant messages is a very intricate task.

Using information extracted from social media to predict human activities has also become an important research area in recent years. As the largest source of public opinion, social media texts are believed to represent the “wisdom of the crowd”. Both researchers and practitioners have successfully used social media to predict a variety of social and economic activities. Yet, the predictive value of social media still remains largely unexplored. There is also a lack of a theoretical foundation on the relationship between social media opinion and other human activities.

This dissertation advances the theory and techniques for social media analytics by developing predictive models and testing hypotheses. It consists of three research essays. The first essay improves sentiment classification of social media texts by proposing a method to automatically construct a domain-aware sentiment lexicon. The second essay proposes the use of dependency structures in sentiment classification. The third essay examines the relationship between social media sentiment and stock market activities.

### **Essay 1: Adapting Sentiment Lexicons to Domain-Specific Social Media Texts**

The application of sentiment analysis to social media texts has great potential, but faces great challenges because of domain issues. Sentiment orientation of words varies by content domain, but learning context-specific sentiment in social media domains continues to be a major challenge. The language domain poses another challenge since the language used in social media today differs significantly from that used in traditional media. To address these challenges, we propose a method to adapt existing sentiment lexicons for domain knowledge using a domain-specific corpus and a dictionary. We have evaluated our method using a large developing corpus containing 743,069 tweets related to the stock market and five existing sentiment lexicons as seeds and baselines. The results support the usefulness of our method, showing significant improvement in sentiment classification performance.

### **Essay 2: Resolving Ambiguity of Sentiment -The Quest for Sentiment Indicators**

Sentiment analysis has become popular in Business Intelligence and Analytics applications because of the great need for learning insights from the vast amounts of user generated content on the Internet. One major challenge of sentiment analysis, like most text classification tasks, is finding structure from unstructured texts. Existing sentiment analysis techniques employ the supervised learning approach and the lexicon scoring approach, both of which largely rely on the representation of a document as a collection of words and phrases. The semantic ambiguity of single words and the sparsity of phrases have negatively affect the robustness of sentiment analysis, especially in the context of short social media texts. In this study, we propose to represent texts as graphs based on a dependency grammar. We use this representation in both the supervised learning and the

lexicon scoring approaches for sentiment analysis. We compare our method with the current standard practice using a labeled data set containing 353,228 social media messages. The combination of the unigram features and dependency features significantly outperforms the common practice.

### **Essay 3: To Predict or to be Predicted? Empirical Study on Social Media Sentiment and Stock Return**

User sentiment extracted from social media texts has been used to predict a variety of social and economic activities, such as election results and product sales. It also has been used to predict stock return by both researchers and investors in recent years. However, there is a lack of consensus on the predictability of social media sentiment. This essay investigates the relationship between social media sentiment and stock returns by examining multiple theories from psychology and behavioral finance. We test hypotheses using vector autoregression (VAR) and a data set containing 18 million social media messages. The empirical results support our hypothesis that social media sentiment does not predict stock return but, instead, is heavily influenced by stock return. We also reveal the strong negative sentiment response on social media and its dynamics. These results should give caution to data scientists who use social media information for developing predictive models. Our findings also have strong implications on research in social media, e-commerce, and behavioral finance.

## CHAPTER 2

### Essay 1: Adapting Sentiment Lexicons to Domain-Specific

#### Social Media Texts

##### 2.1. Introduction

Social media has experienced exponential growth in the past few years and has become the number one activity on the Internet today (Wikipedia 2013). The vast amount of user-generated content has made social media the largest data source of public opinion (Bifet and Frank 2010; Yu et al. 2013). Such a data source is invaluable for business intelligence and analytics since opinion is a key predictor of human behavior (Liu 2012). Despite the tremendous effort in influencing customers through marketing campaigns on social media (Divol et al. 2012), extracting public opinion from social media is still in its infancy (Chen et al. 2012; Yu et al. 2013). Both business practitioners and researchers are still in search for more effective tools to derive value from social media data. Social media data include profiles, networks, pictures, videos, and textual content. Compared to other types of data, text data are more popular and dynamic because they can be generated in almost all circumstances (Chau and Xu 2012). Moreover, user-generated text data usually contain opinions and are more likely to influence other users than traditional media (Bickart and Schindler 2001). As a result, text data in social media have the best potential to keep businesses informed in real-time, if appropriate techniques are employed (Sakaki et al. 2010).

Sentiment analysis, as a class of techniques to extract and assess opinions in texts, has been used to analyze text data in social media (Abbasi et al. 2011; Chen and Zimbra 2010).

Sentiment analysis has been used to solve a variety of business problems. Aggarwal et al. (2012a) found that the sentiment of blogs significantly influences the financial performance of ventures. Mishne and Glance (2006) analyzed the correlation between movie sales and the sentiment in blog postings related to the movie. They found that the sentiment is correlated more with the sales than with the volume of the relevant blogs. Bollen et al. (2011) measured tweet sentiment in different dimensions and found certain dimensions of sentiment predict the Dow Jones Industrial Average (DJIA) index. Oh and Sheng (2011) found that sentiment adds predictive power to existing text mining models when using microblogs to predict individual stock prices.

Such applications of sentiment analysis and their findings have demonstrated the tremendous opportunities for understanding the public opinion through social media. However, major challenges remain to be addressed because the effectiveness of sentiment analysis largely depends on the language being analyzed (Boiy and Moens 2009). In the past decade, many studies have applied sentiment analysis to online reviews and have obtained satisfactory results (Abbasi et al. 2011; Hu and Liu 2004; Ngo-Ye and Sinha 2012; Pang et al. 2002). However, textual social media data raise new challenges for sentiment analysis.

Sentiment analysis commonly employs techniques from text mining and natural language processing (NLP) (Chen and Zimbra 2010). Most research and applications implement the task as classifying the directional states of sentiment, e.g., positive, negative, or neutral (Liu 2012; Pang et al. 2002). The two most popular approaches for sentiment classification are supervised learning and lexicon scoring. The supervised learning approach trains a machine learning classifier using a large annotated corpus in which the sentiment of each

document is recognized by experts (Riloff and Wiebe 2003). This approach has been shown to be more accurate than the lexicon approach when the training data and testing data are from the same domain (Chaovalit and Zhou 2005). The power of machine learning resides in the training data. However, a large and high-quality training data set demands tremendous human effort and time. Such corpora of social media texts are rarely available. This raises a bigger challenge for analyzing social media texts because social media texts are generally short and heterogeneous. It requires a much larger training set than traditional media texts to build an effective classifier.

In the absence of sufficient training data, the lexicon scoring approach has been widely used as a convenient and effective alternative by most researchers and practitioners. This approach searches for sentiment indicators in the document to be classified based on a lexicon (Liu 2012). The overall sentiment of the document is determined by the dominant polarity (i.e., positive or negative) among the indicators (Fu et al. 2012). The performance of this approach relies on an accurate and abundant sentiment lexicon. Nevertheless, the effectiveness of the existing lexicons is quite limited when applied to new problems (O'Leary 2011).

The existing sentiment lexicons consist of mainly words that are deemed to carry sentiment. It is well known that the sentiment orientation of a word varies by its context (Ding et al. 2008; Liu 2012; Pang and Lee 2008). Considering the following tweet commenting on the stock prices of Apple Inc. (\$AAPL) and Caterpillar Inc. (\$CAT):

*I'm seeing a red close on both \$AAPL and \$CAT today.*



The word, *red*, implies a pessimistic opinion about the price movements of \$AAPL and \$CAT because price decreases are usually quoted in red in the US. However, this information cannot be captured without knowing the context. In general contexts, *red* would be understood as a color, which does not carry any sentiment. Previous research has also found that using general sentiment lexicons for domain-specific tasks can lead to a severe misunderstanding of the information (Loughran and McDonald 2011). In this essay, we refer to such contexts of discourse as content domains. There has been some research to derive the contextual polarity of words using annotated data for sentiment classification (Wilson et al. 2009). However, given the unavailability of sufficient training data for social media, learning context-specific sentiment continues to be a major challenge (Raina et al. 2007). It is much easier to obtain unlabeled data than labeled data.

The language domain poses another challenge that sentiment analysis needs to address when applied to social media. This essay adopts the definition of language domain as the lexical and syntactical choices of language. As our language is constantly evolving, social media users keep finding new expressions for their emotions. The language used in social media today differs significantly from that used in traditional media. For instance, word lengthening is shown to bear strong sentiment (e.g., coooooool) (Brody and Diakopoulos 2011). Without incorporating the latest language elements of social media, the power of sentiment analysis in this language domain would be quite limited. Unfortunately, none of the existing sentiment lexicons being used is based on social media language.

In this study, we address these domain challenges of conducting sentiment analysis on social media texts. We propose an automated approach to generating a sentiment lexicon

that integrates elements from both content domain and language domain. To our knowledge, none of the existing methods addresses these two challenges together.

In the following section, we review the lexicon scoring approach to sentiment classification, as well as lexicon generation methods. Next, we describe our approach to generating a domain lexicon and evaluate our method with respect to existing lexicons and other sentiment analysis tools. In the final section, we summarize our contributions and discuss prominent future research directions.

## **2.2. Literature Review**

### **Classification Using Supervised Learning**

The supervised learning approach for sentiment classification uses machine learning classifiers to learn the associations between the sentiment class and various features from a training corpus. Features represent useful information in the documents. The bag-of-words representation, where each document is modeled as a vector containing the frequency of words or phrases, has been the most commonly used. Binary feature representation, indicating the presence or absence of words, have also been used for short texts. Assigning higher weight to less frequent features in a corpus has been shown to be more effective in some domains. Part-of-speech (POS) tags have been also attached to words in some studies to differentiate words with multiple syntactical properties (Hu and Liu 2004). More complex features have also been proposed, but they do not consistently outperform plain word features (Abbott et al. 2011). Bag-of-words features usually lead to a large feature set and are likely to cause overfitting (Abbasi et al. 2011). Various feature

selection methods have been proposed to alleviate the overfitting problem (Abbasi et al. 2011; Chou et al. 2010).

Supervised learning methods for sentiment analysis require a large training corpus in order for the classifier to learn the numerous ways that sentiment is expressed. The requirement for social media is even higher since social media texts are generally short and highly heterogeneous. This has raised a big obstacle for both business practitioners and researchers. However, to our knowledge, no extant study has used a manually labeled training corpus containing more than a few thousand instances. Major social media platforms can generate many more messages than others in just a few seconds. This indicates that the training data sets used in previous studies are not utilizing big data and are not large enough to capture the characteristics of the language used on social media platforms, such as Twitter. In such circumstances, the lexicon approach provides a good alternative to sentiment analysis.

### **Classification Using Lexicons**

A sentiment lexicon, also called opinion lexicon (Liu 2012), is a collection of words or phrases that are commonly used to express feelings (Pang and Lee 2008). Some of the most widely used sentiment lexicons are General Inquirer (GI) (Stone et al., 1966), Multi-Perspective Question Answering (MPQA) (Wilson et al., 2005a), SentiWordNet (SWN) (Esuli and Sebastiani 2006), and Opinion Lexicon (OL) (Hu and Liu 2004). Each entry in the lexicon is associated with a sentiment score. In most lexicons, the score just indicates the direction of the sentiment, i.e., positive, negative, or neutral. Some lexicons use a continuous scale to reflect the strength of sentiment. However, so far no research study has found a significant increase in performance when using continuous sentiment scores. When

classifying the sentiment in a document, each word in the document is checked against the sentiment lexicon and sentiment scores are recorded as matched words are found. The document-level sentiment is calculated using both the positive score and the negative score. Most studies use the difference between the score obtained from the positive words and that from the negative words (Fu et al. 2012; Hu and Liu 2004). Some studies impose more weights on certain words. For example, (Das and Chen 2007) assigned higher weights to the scores obtained from matching adjectives and adverbs, which are believed to be more likely to bear sentiment. In some sentiment lexicons, a POS tag is attached to each word for disambiguation of word with multiple POS tags (Wilson et al. 2005b). For example, “good” as a noun usually does not carry any positive or negative feelings. But when it is used as an adjective, it most likely indicates positive feelings.

The lexicon approach is more favorable than the machine learning approach in the absence of a large training data set. It is believed to work well on short texts, which is a major characteristic of social media texts (Bermingham and Smeaton 2010). It is also suitable for real-time sentiment classification given its relatively lower computation requirement (Chaovalit and Zhou 2005). General lexicons are robust across different domains (Liu 2012). However, their performance usually suffers from two aspects, insufficiency and inaccuracy (Brody and Diakopoulos 2011).

### **Sentiment Lexicon Generation Methods**

The sentiment orientation of a word in a sentiment lexicon is usually assigned either manually or using an automatic method. The two most popular sentiment lexicons, General Inquirer (Stone et al. 1966) and Multi-Perspective Question Answering (MPQA) Subjectivity Lexicon (Wilson et al., 2005b), are manually compiled by experts. Due to the

large investment in expert time and effort required, the manual approach is not efficient for developing sentiment lexicons in new domains. Besides, sentiment words in the developing corpus usually follow a long-tail distribution. Most of the expert time is spent on scanning the most frequent sentiment words again and again. The less frequent sentiment words could be easily overlooked. This imposes a size limit on manually developed lexicons. In fact, the largest manually developed lexicon, MPQA, has only 7,630 entries.

Automatic methods usually utilize one of two types of language resources, dictionary or corpus (Liu 2012). In the dictionary-based approach, new sentiment words are identified by their relationships with a small set of handpicked sentiment words, known as the seed lexicon. The dictionary being used defines these relations. For instance, the Liu's Opinion Lexicon (Hu and Liu 2004) includes sentiment adjectives recognized by synonyms and antonyms of a seed lexicon. The effectiveness of this method is highly dependent on the synonym and antonym entries of the dictionary used. Although these relations between entries are highly accurate, these lexicons do not contain any domain-specific information.

In the corpus-based approach, new sentiment words are recognized based on their relationships with seed words in the corpus. Hatzivassiloglou and McKeown (1997) identified sentiment words using conjunctions in a 21-million-word WSJ (Wall Street Journal) corpus. For example, in the sentence, "This approach is nice and easy," "nice" and "easy" should have the same polarity since they are used to express the same opinion toward the same topic. In another case, "this monitor is nice but expensive," "expensive" should have the opposite polarity as "nice". In their algorithm (Hatzivassiloglou and McKeown 1997), they did not use a seed lexicon. Instead, they relied on linguistic

observations to determine the polarity of a word. In general, they believed that unmarked words are more likely to bear positive sentiments than marked words.

Other corpus-based approaches also utilize seed words. Turney (2002) introduced a Pointwise Mutual Information and Information Retrieval (PMI-IR) approach to identify the polarity of phrases. This method is not designed to develop a sentiment lexicon. Instead, it directly identifies the polarity of the potential sentiment words in the documents being classified. PMI (Church and Hanks 1990) is used to measure the co-occurrence between two words/phrases. The polarity of a word is the difference of its PMI with a positive word and that with a negative word. The two seed words chosen in (Turney 2002) are “excellent” and “poor”, which are believed to be the most frequently used in online reviews.

Wiebe & Riloff (2005) used an iterative approach to automatically annotate a developing corpus and then extract subjective expressions from it. To create the training corpus, they used a seed lexicon and two high-precision low-recall classifiers, one to identify subjective sentences and the other to identify objective sentences. Then they used several patterns to extract subjective and objective words. However, this method does not identify the polarity of subjective words.

### **2.3. Proposed Method**

As discussed earlier, the manual approach for developing sentiment lexicon is most accurate when the sentiment words are picked directly by human experts (i.e., high precision). However, this approach is only able to identify a limited number of sentiment words within a reasonable time frame (low recall). The dictionary-based approach is able to discover much more sentiment indicators if the dictionary contains sufficient synonyms

and antonyms. However, the words identified are generally less accurate than the manual approach, without human supervision. Besides, the relations in dictionaries are not updated frequently and, therefore, cannot incorporate new elements into the language in a timely fashion.

The corpus-based approach has the best potential to address our research challenge. First, it provides the flexibility for trading off between precision and recall of the identified sentiment words. When a seed lexicon is used, loose relations between the seeds and the candidates can identify more sentiment words but with lower precision. In contrast, strict word relations increase the precision of new sentiment words at the cost of recall. The control over word relations allows the corpus-based algorithms to adapt to different needs. Second, the developing corpus used in this approach can incorporate the latest information from both the content domain and the language domain. As no annotation is needed, such a developing corpus can be easily crawled or streamed for immediate use. Using such a corpus has the obvious advantage over human labor and dictionaries when sentiment analysis is applied to new domains.

### **Method Description**

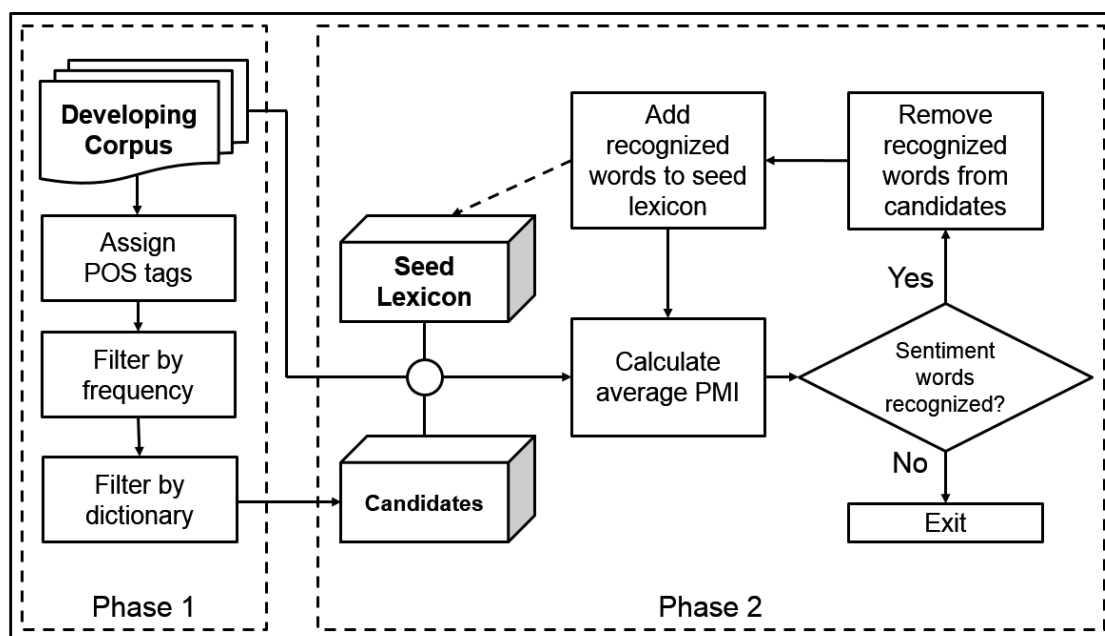
We propose a corpus-based lexicon generation method that learns sentiment words based on both content domain and language domain. This method utilizes three language resources: a developing corpus, a seed lexicon, and a dictionary. The use of a developing corpus for generating the sentiment lexicon provides the necessary resources for domain-relevant sentiment indicators. This corpus should be highly relevant to both the content domain and the language domain. The seed sentiment lexicon is used to recognize the polarity of new sentiment words once the relation between the seed and the candidate is

established. The dictionary is used to filter out idiosyncratic noise, as spelling mistakes are common in social media texts. To prevent the loss of potential sentiment words that are specifically used in social media, the dictionary needs to contain sufficient entries from social media language. For instance, Wiktionary (Wiktionary 2013) is an acceptable choice for this task. Compared to other dictionaries, Wiktionary is actively updated and contains more terms used in social media.

Figure 1 illustrates our method, which consists of two phases: candidate extraction and sentiment recognition. The following is an outline of the steps in our method:

- 1) Extract candidates from the developing corpus.
- 2) Find the relations between the candidate words and the sentiment words in a seed lexicon.
- 3) Determine the sentiment orientation of the candidates and add the recognized sentiment words to the seed lexicon.
- 4) Repeat steps 2 – 3 until no more new words are added.





**Figure 1. The Proposed Lexicon Generation Method**

### Extracting the Candidates

The first step of our method is to extract candidates for sentiment words from the developing corpus. Not all words in the developing corpus are equally likely to be sentiment words. Certain filtering needs to be applied. We use the POS tags as the first filter of candidates. In some previous work on developing sentiment lexicons, all types of words are retained (Velikovich et al. 2010). Most previous work favors adjectives and adverbs, which are by nature more likely to carry sentiments (Liu et al. 2005). Several extant studies on sentiment lexicon creation consider only adjectives and adverbs as sentiment clues (Hatzivassiloglou and McKeown 1997; Hu and Liu 2004; Qiu et al. 2009; Turney 2002). Nouns and verbs are less frequently used since many of them do not carry any sentiments. But they are also commonly used to describe objects and events. As a matter of fact, SWN and MPQA both include a fair number of nouns and verbs (Esuli & Sebastiani, 2006; Wilson et al. 2005b). Intuitively, automatically learning nouns and verbs

should increase the noise of a lexicon, i.e., introducing many words that do not bear sentiment. Including these two types of words is likely to increase the chance of adding non-sentiment words to the lexicon, i.e., reducing the precision of the lexicon. This is the reason why most previous studies have not used them. However, we believe it would also add many domain sentiment words that would otherwise not be recognized, thereby increasing the recall. To allow us to make a trade-off between precision and recall, we propose to use three sets of candidates: the first set includes all four types of words (ARVN, for adjective, adverb, verb, and noun), the second candidate set excludes nouns (ARV), and the third candidate set excludes nouns and verbs (AR). In the ARVN set, proper nouns are excluded since they rarely bear sentiments.

The second filter we use is an English dictionary. Many tokens in social media texts are not really words. Spelling mistakes are prevalent. The dictionary is used to retain tokens that are indeed words. We use the English Wiktionary, a dictionary that is updated frequently and contains a large amount of language being used on the Web. This dictionary also provides an application programming interface (API) that can be readily used in any algorithm.

The third filter we apply on the candidates is based on word frequency in the developing corpus. Irregular spellings in social media are common. However, many irregular spellings are accepted and have become popular among social media users. For those words that are not recognized by the dictionary but are being used repeatedly, we believe they are likely to be such tokens. Thus, if a word cannot be found in the English Wiktionary but has occurred more than a certain number of times in the developing corpus, we retain it in the candidate set.

### Identifying Polarity

The second step in our method is to use a seed sentiment lexicon to identify sentiment polarity of the candidate words. We use PMI (Church and Hanks 1990) to measure the association between the candidate words and the seed words. The PMI between two words  $w_1$  and  $w_2$  is

$$PMI(w_1, w_2) = \log_2 \frac{p(w_1, w_2)}{p(w_1)p(w_2)}$$

where  $p(w_i)$  is the probability that word  $w_i$  occurs in a document and  $p(w_1, w_2)$  is the probability that  $w_1$  and  $w_2$  co-occur in a document. The sentiment polarity of each candidate  $c$  is determined by

$$Polarity(c) = \frac{1}{N_p} \overset{\circ}{\bar{a}}_{i=1}^{N_p} PMI(c, p_i) - \frac{1}{N_n} \overset{\circ}{\bar{a}}_{i=1}^{N_n} PMI(c, n_i)$$

where  $\{p_i, i = 1, 2, \dots, N_p\}$  and  $\{n_i, i = 1, 2, \dots, N_n\}$  are the sets of positive sentiment words and negative sentiment words, respectively.

We interpret polarity as the difference between a candidate's average association with all positive seed words and all negative seed words. Specifically, it is the logarithm of the average conditional probability of the candidate given a set of positive words, divided by that of the candidate given a set of negative words. We use a positive parameter,  $H$ , to control the threshold of the difference. A candidate with a polarity score greater than  $H$  is added to the seed lexicon as a positive word. A candidate with a polarity score smaller than  $-H$  is added to the seed lexicon as a negative word. A smaller value of  $H$  leads to more

words (and more noise) in the final lexicon. This procedure is repeated until no more words are added to the lexicon.

It is possible that a candidate set contains some of the seed words. In that case, the sentiment polarity of the seed word is recalculated using the above procedure. By doing so, the algorithm corrects the polarity of the seed word using domain knowledge.

Figure 2 presents our algorithm. The only parameter that needs to be tuned is the sentiment benchmark threshold,  $H$ . When  $H = 2^x$ , a candidate is considered as positive if it co-occurs with positive seeds on average  $x$  times more than with negative seeds, and vice versa. A high value of  $H$  would recognize too few sentiment words. A low value of  $H$  would add too much noise.

Let  $C = \{c_1, c_2, \dots, c_n\}$  denote the set of candidates. Let  $PW = \{pw_1, pw_2, \dots, pw_A\}$  and  $NW = \{nw_1, nw_2, \dots, nw_B\}$  denote the positive and negative seed word sets, respectively.

For each  $c_i$  in  $C$ ,

{

Let  $APMI(c_i, PW)$  denote the average PMI between  $c_i$  and the elements in  $PW$ , i.e.,

$$APMI(c_i, PW) = \frac{1}{K} \sum_{j=1}^K PMI(c_i, pw_j)$$

where  $K = \text{count}(PMI(c_i, pw_j) \neq 0)$ .

Let  $APMI(c_i, NW)$  denote the average PMI between  $c_i$  and the elements in  $NW$ , i.e.,

$$APMI(c_i, NW) = \frac{1}{L} \sum_{j=1}^L PMI(c_i, nw_j)$$

where  $L = \text{count}(PMI(c_i, nw_j) \neq 0)$ .

Let  $SS(c_i)$  denote the sentiment score of  $c_i$ , i.e.,

$$SS(c_i) = APMI(c_i, PW) - APMI(c_i, NW)$$

Let  $H$  denote the threshold to benchmark  $SS(c_i)$ ,  $H > 0$ .

If  $SS(c_i) > H$ ,  $\text{Polarity}(c_i) = \textit{positive}$  and move  $c_i$  from  $C$  to  $PW$ ;

If  $SS(c_i) < -H$ ,  $\text{Polarity}(c_i) = \textit{negative}$  and move  $c_i$  from  $C$  to  $NW$ .

**Figure 2. Procedure for Recognizing Sentiment Words from Candidates**

## **2.4. Evaluation**

We evaluated our method on a specific combination of language domain and content domain, which has both academic and practical impact. For the language domain, we used postings from Twitter, the most popular microblogging platform (Alexa 2013). As of early 2013, Twitter had over 200 million active users creating over 400 million tweets each day (Twitter 2013). This data source has great value for mining public opinion. For the content domain, we chose tweets related to the stock market. Recent studies have shown the usefulness of using tweet sentiment to predict stock price movements (Bollen et al. 2011; Oh and Sheng 2011; Xu 2012; Yu et al. 2013), but the treatment of domain characteristics for sentiment analysis in extant literature has been minimal.

### **Seed Lexicons and Baselines**

We used four general sentiment lexicons and one domain lexicon as both seed lexicons and baselines. The general lexicons are General Inquirer (GI), MPQA Subjectivity Lexicon (MPQA), Opinion Lexicon (OL), and SentiWordNet (SWN). The domain lexicon we used is published by (Loughran and McDonald 2011), coded as LM. This lexicon contains positive and negative words picked by the authors after reviewing a large amount of financial reports.

The GI lexicon (Stone et al., 1966) refers to the positive and negative word lists among its various dictionaries for content analysis. It contains 1915 positive words and 2291 negative words. These words were manually picked from psychological dictionaries by experts. It is frequently used by both business researchers (e.g., Tetlock et al. 2008; Das & Chen 2007) and computational linguistic researchers (e.g., Wilson et al. 2009; Jiang et al. 2012).

The MPQA sentiment lexicon was also compiled manually from multiple resources, including GI (Wilson et al., 2005a). Each entry contains both the polarity (i.e., positive, negative, or neutral) and the strength of subjectivity (i.e., strong or weak). A word that is neutral in polarity can be strongly subjective, for example, “absolute”. The entries are word-tag pairs, that is, each word and its POS tag. The polarity of a word can change depending on the tag. The lexicon contains a much larger number of negative entries than positive ones, 4,912 against 2,718. It has also been used in OpinionFinder (OF), a software package that is used by business researchers to analyze sentiment in social media texts (Bollen et al. 2011) and news (Schumaker et al. 2012).

SWN contains sentiment-bearing synsets automatically identified from WordNet, a dictionary widely used for word sense disambiguation (WSD) (Esuli and Sebastiani 2006). Each synset represents a word sense, which may contain a group of words that have this sense. Each synset is also associated with a POS tag. The same word can fit in different synsets, and thus, can have different sentiment orientations. Each entry is assigned a positive score and a negative score, which range between 0 and 1. The objective score of the entry is subsequently calculated as  $1 - (\text{positive score} + \text{negative score})$ . The use of SWN is not that straightforward since the entries are not words and the same words may have conflicting entries. Since WSD is no trivial task and rarely done in sentiment analysis, the synset sentiment scores have to be converted to word scores. Fahrni & Klenner (2008) and Fu et al. (2012) did such conversion by averaging the sentiment score of a word across synsets. (Zhang et al. 2012) and (Balasubramanyan et al. 2011) converted the continuous scale to discrete values. (Thet et al. 2010) adopted the highest sentiment score of each word. (Taboada et al. 2011) used both the first sense of each word and the average score and

found that the average-sense method generally performs better than the first sense method. We used a conversion similar to that of (Fu et al. 2012). That is, for each word in SWN, we use the net average score of its different senses.

The OL lexicon contains 4783 negative and 2006 positive words. The polarity of words in the lexicon takes two values, positive and negative. The majority of the words were identified using a dictionary-based algorithm (Hu and Liu 2004). The dictionary used is also WordNet.

Among these baseline lexicons, only MPQA and SWN contain POS tags for their entries. To make them consistent with other lexicons, we used their entries without the tags and removed words which had different polarities with different tags.

**Table 1. Agreements and Conflicts between Existing Lexicons**

Lexicon	Size	Agreements				Conflicts			
		MPQA	SWN	OL	GI	MPQA	SWN	OL	GI
MPQA	7,630								
SWN	39,680	5,402				1,260			
OL	6,789	5,418	4,782			50	1,245		
GI	4,206	2,542	2,563	2,149		60	611	26	
LM	2,690	951	1,232	1,043	508	26	292	9	15

Table 1 summarizes the agreements and conflicts among these lexicons. The agreements are counts of common words having the same polarity between two lexicons. The conflicts are counts of common words having different polarities between two lexicons. The non-trivial existence of the conflicting sentiment words supports our claim that the same word can have a different polarity in a different context, which is one of the motivating factors



for this study. The agreement between most lexicon pairs indicates that the sentiment words included in each individual lexicon are far from complete. This also confirms the necessity of expanding an existing lexicon. The agreements between pairs containing LM are lower than the rest. Because LM has been developed for the finance domain, while other lexicons do not pertain to specific content domains, the low agreements indicate that general sentiment lexicons severely lack domain-specific entries. This is precisely the area that our method aims to improve.

We also created another baseline lexicon (coded as Combined4) by combining all of the lexicons mentioned above, except SWN. Conflicts across lexicons were removed. SWN was excluded from the combined sentiment lexicon because it contains much more entries than any other lexicon. If included, it will dominate the combined lexicon.

**Table 2. Positive and Negative Entries in Existing Lexicons**

<b>Lexicon</b>	<b>Positive</b>	<b>Negative</b>	<b>Total</b>
MPQA	2,718	4,912	7,630
GI	1,915	2,291	4,206
OL	2,006	4,783	6,789
LM	353	2,337	2,690
Combined4	3,649	7,231	10,880
SWN	18,386	21,294	39,680

### **Experiment Procedure**

To construct the developing corpus, we queried the tweets stream using the stock symbols of the Standard and Poor’s 500 (S&P 500) index, e.g., \$AAPL and \$GOOG. Since we focused on analyzing English texts in this study, we filtered out non-English tweets using

the English Wiktionary. If more than half of the words in a tweet are not in the dictionary, the tweet was removed. The final developing corpus contains 743,069 tweets. We randomly sampled 500 tweets to validate the data. Among the sample, only 26 tweets are written in a non-English language and 11 tweets are not related to the stock market. This supports the usability of our developing corpus.

To extract candidates, we first POS tagged the developing corpus. We trained the Stanford POS Tagger using a customized training set. The tagger achieved over 97.5% tagging accuracy on a sample of tweets. This is much higher than the default tagger model, which was only able to achieve 82% tagging accuracy. After the POS tagging, 99,917 unique tokens were obtained. The three candidate sets, ARVN, ARV, and AR, were subsequently created. After removing words that occurred less than 3 times in the developing corpus, words that are not in Wiktionary, stop words, numbers, and hash marks, the three candidate sets contain 20,198, 12,431, and 5,359 candidates, respectively. We compared ARVN with all baseline lexicons and found that many potential sentiment words, for example, *earnings*, *profits*, *rally*, *swing*, *drops*, and *lol*, were not contained in these lexicons. This lends further support to the motivation for our study.

We used the lexicon scoring approach for sentiment classification: if a tweet contains more negative words than positive words, then it is classified as negative; otherwise, it is classified as positive. Through some preliminary experiments, we set the sentiment benchmark parameter  $H$  to 8, meaning that a candidate is considered as belonging to one class if it co-occurs with the seeds of the same class on average three times more than with the seeds of the other class. This value prevents our method from identifying too many or too few sentiment words.

We asked three doctoral students in a business school to label a random sample of tweets as having positive, negative, or neutral sentiment toward the mentioned stocks. The evaluation of our proposed method was based on its performance ( $F$ -measure and accuracy) in classifying these tweets. Since our domain lexicon generation method focuses on distinguishing between positive and negative words, we used only positive and negative tweets for testing. The final testing set contains 584 positive tweets and 584 negative tweets.

## Results

Table 3 presents the results of our experiments. The expanded lexicons are denoted as  $\langle \text{baseline lexicon} \rangle_{\langle \text{candidate set} \rangle}$ . The results show that the expanded lexicons generally outperformed the baseline lexicons. The Combined4 lexicon outperformed all individual lexicons, showing the usefulness of combining multiple lexicons. Expanding the Combined4 lexicon with ARVN candidates further significantly improved the performance, demonstrating the value of our proposed method beyond combining multiple lexicons. A few examples of the newly recognized sentiment words are *mish-mash* – negative, *dumbs* – negative, *hit-and-miss* – negative, *illuminate* – positive, *frontrunning* – positive, and *strong-arm* – positive. These words are closely related to our evaluation domain.

**Table 3. Experiment Results**

Lexicon	Pos.	Neg.	F-measure (%)	Accuracy (%)	Lexicon	Pos.	Neg.	F-measure (%)	Accuracy (%)
Combined4_ARVN	3,625	8,035	78.02	78.99	LM_AR	355	2,801	60.28	62.52
Combined4_ARV	3,630	7,733	75.61	76.76	LM	353	2,337	59.54	63.38
Combined4	3,649	7,231	72.78	74.36	LM_ARV	362	3,505	55.67	59.01
Combined4_AR	3,634	7,427	72.59	74.19	LM_ARVN	365	4,350	50.78	57.29

<b>GI_AR</b>	1,386	2,283	51.78	58.32	<b>MPQA_ARV</b>	2,286	4,635	70.73	72.13
<b>GI_ARV</b>	1,393	3,275	49.56	57.89	<b>MPQA_ARV N</b>	2,286	4,922	70.23	71.44
<b>GI</b>	1,915	2,291	47.12	56.35	<b>MPQA_AR</b>	2,288	4,329	65.99	68.78
<b>GI_ARVN</b>	1,394	4,481	43.31	53.77	<b>MPQA</b>	2,718	4,912	64.55	67.84
<b>OL_ARV</b>	1,999	5,620	68.31	69.47	<b>SWN_ARVN</b>	54,595	44,802	57.33	58.58
<b>OL_AR</b>	1,995	5,107	67.40	69.21	<b>SWN_ARV</b>	30,756	29,168	56.98	58.06
<b>OL</b>	2,006	4,783	66.28	68.87	<b>SWN</b>	18,386	21,294	55.62	57.89
<b>OL_ARVN</b>	1,998	6,105	65.87	67.75	<b>SWN_AR</b>	22,509	24,522	55.62	57.89

We believe this result is dependent on the large coverage of the sentiment words the Combined4 lexicon contains. When an individual lexicon was used as the seed, the ARVN could not produce the best result. This is caused by the insufficient coverage of the individual lexicons. Among the original lexicons, GI had the worst performance. This is as we expected since the lexicon was created half a century ago. Expanding GI using our method improved the *F*-measure by over 4%. However, expanding LM did not lead to much improvement in sentiment classification. We believe this is caused by the imbalance between positive and negative words in this lexicon. The original LM contains 2,348 negative words and only 355 positive words. The overwhelming amount of negative words dominated the lexicon generation procedure as the expanded lexicon contains 4,673 negative words and 932 positive words. Such an unbalanced lexicon caused strong bias toward the negative class. OL and MPQA achieved better performance individually since both of them have larger initial coverage of sentiment words. These results indicate that the seed lexicon is crucial in lexicon expansion. With a high quality seed lexicon, more domain sentiment words can be learned.

## **2.5. Conclusion and Future Directions**

In this essay, we proposed a lexicon expansion method to improve sentiment classification by learning domain knowledge. We tested our method with a specific combination of language domain and content domain. By using a large unannotated developing corpus, our method expands a seed lexicon by adding more domain-specific sentiment words. The evaluation results show that the expanded lexicons improve the sentiment classification performance significantly compared to the seed lexicons. Based on the findings, the proposed method will benefit a variety of business applications using sentiment analysis. An unannotated developing corpus needed for this method can be easily obtained for a target domain and a domain-specific sentiment lexicon can be quickly learned. This is especially useful in the age of “big data”.

This study opens up several avenues for future research. A better seed lexicon could be used to further improve the performance of the expanded lexicon. In this study, we used PMI to measure the association between words. Other measurements, such as context similarity, could also be useful. Besides single words, bigrams and trigrams may be useful to include in the candidate set. Certain patterns need to be developed to filter these candidates. Furthermore, domain-specific language resources, other than dictionary and developing corpus, could be explored.

## CHAPTER 3

### Essay 2: Resolving Ambiguity of Sentiment –The Quest for Sentiment

#### Indicators

##### 3.1. Introduction

###### Social media and sentiment analysis

Social media has become the number one activity on the Internet (Wikipedia 2013). The tremendous volume of social media texts contains rich user opinions, providing businesses with a great opportunity to monitor their environments in real time (Bifet and Frank 2010; Yu et al. 2013). Sentiment analysis, also known as opinion mining, has emerged to be a useful tool to extract subjective information from texts. It has become an important part of the modern business intelligence and analytics solutions.

###### Research gap

Sentiment analysis typically classifies the directional emotions in texts into several categories, e.g., positive, negative, and neutral (Abbasi et al. 2011; Chen et al. 2012). It relies on natural language processing (NLP) and text mining techniques. Since most text data are unstructured, the biggest challenge of sentiment analysis is finding effective structures from texts. The standard practice in this field is to represent texts as a collection of words and/or phrases. However, single words are ambiguous. Their semantics vary by context. Social media texts are typically short, making the contextual information even less available. Phrases are much less ambiguous compared to single words. However, they lack flexibility since they only account for fixed word permutations.

**Research objective**

This study addresses the ambiguity issue in sentiment analysis by introducing dependency features as sentiment indicators. Dependencies are pairwise word relations. We argue that dependency representation of texts is more effective than phrases in sentiment analysis. Using dependencies has two advantages over using single words and phrases. First, dependencies incorporate contextual information by using word relations. Second, a word relation can still be established even if two words are not adjacent. In this study, we propose dependency features in supervised sentiment classification.

We compare the classification correctness of dependencies with the current standard practice, n-grams and part-of-speech tagged words, on a large data set.

The remainder of the essay is organized as follows: The second part reviews standard practices in sentiment analysis and contrast dependency features with commonly used features. The third part describes the sentiment classification method using dependency features. The fourth part evaluates the dependency features in comparison to standard practices and discusses the results. The last part addresses the limitations and identifies future research directions.

### **3.2. Literature Review**

There are two major approaches to sentiment analysis, supervised learning and lexicon scoring (Liu 2012; Pang and Lee 2008). The supervised learning approach represents a document as a set of linguistic features and trains a machine learning classifier using a large annotated corpus, in which the sentiment category of each document is known. The trained model is subsequently used to classify the sentiment of other texts (Hu and Liu 2004; Pang et al. 2002). The most common method used to represent texts is the word n-gram model. Unigram models present a document as a vector of word frequencies (i.e., vector space model). This is also known as the bag-of-words (BOW) model. Term frequency-inverse document frequency (TFIDF), which assigns more weights to words that only occur in a few documents, has been used as an improvement to plain frequency (Chou et al. 2010). For short documents, such as social media text, the binary value of word presence has also been used (i.e., exists or does not exist). As an effort to resolve word semantics, part-of-speech (POS) tagged words have also been used. The major drawback of the BOW model is the strong assumption that the word order (i.e., syntax) does not matter. To incorporate syntactical information, existing studies have also attempted to use phrase patterns, bi-grams, and tri-grams. However, they have not found consistent performance improvement by using these features.

#### **Syntax**

One type of important information that previous research has failed to effectively capture is syntax. Syntax refers to the principles of constructing sentences (Chomsky 1965). In natural language processing, there are two types of representations of sentence structures, the constituency grammar and the dependency grammar (Covington 2001). Constituency



grammar, also known as phrase structure grammar, describes a sentence as a set of constituency relations (Chomsky 2002). Single words (i.e., leaves) are the constituents of phrases, which, in turn, are constituents of more complicated phrases, the eventual constituents of the sentence (i.e., root). The phrase patterns used in previous text mining research are a simplified case of constituency features. Constituents are not suitable for use as features in text mining. It does not take much effort to see that higher level and lower level constituents are highly correlated. Using constituency representation would generate a large amount of redundant features, which are no more useful than BOW.

Dependency grammar provides the flexibility to capture the relationship between non-adjacent words. It represents a sentence as a set of pairwise-word relations. The structure is flat compared to constituency grammar. In a dependency relation, one word is called the head and the other is called the dependent. The head usually plays a more important role in determining the behavior of the relation. The dependent typically is the subject, the object, or the modifier of the head.

We show the advantage of dependency grammar through a motivating example. Sentence 1, *AAPL stealing the thunder*, is excerpted from a major microblogging website. The author of this posting reported positive sentiment. Obviously, the sentiment indicator in this text is *stealing the thunder*. The different representations of this sentence are shown in Table 4.

Without loss of generality, we suppose a classifier is trained using Sentence 1 and is used to classify two more sentences, both of which are real world examples:

Sentence 2: *Apple keeps stealing Samsung's thunder*.

Sentence 3: *Apple stealing user information via Face Time*.

We show different representations of both sentences in Table 5. In these representations, if an element has appeared in Sentence 1, we display it in bold.

It is clear that Sentences 1 and 2 both have positive sentiment towards Apple, while Sentence 3 expresses negative sentiment. Among the different types of feature representations, only unigram, POS-tagged words, and dependency relations can find similarity between Sentences 1 and 2. However, unigram and POS-tagged words still find some similarity between Sentence 1 and 3. Eventually, dependency relation is the only feature representation that can distinguish the two sentiment classes in our example. We illustrate the similar elements of different representations for the three sentences in Table 6.

**Table 4. Representations of Sentence 1**

<b>Representation</b>	<b>Elements</b>
Unigram	AAPL, stealing, the, thunder
POS Tagged	AAPL/NNP, stealing/VBG, the/DT, thunder/NN
Bigram	AAPL stealing, stealing the, the thunder
Trigram	AAPL stealing the, stealing the thunder
Constituency (excluding leaves)	(ROOT (S (NP (NNP AAPL)) (VP (VBG stealing) (NP (DT the) (NN thunder)))) (S (NP (NNP AAPL)) (VP (VBG stealing) (NP (DT the) (NN thunder)))) (NP (NNP AAPL)) (VP (VBG stealing) (NP (DT the) (NN thunder))) (NP (DT the) (NN thunder))
Dependency	root(ROOT, stealing) nsubj(stealing, apple) dobj(stealing, thunder) det(the, thunder)

**Table 5. Representations of Sentences 2 and 3**

Representation	Elements
1) Apple keeps stealing Samsung's thunder	
Unigram	Apple, keeps, <b>stealing</b> , Samsung, 's, <b>thunder</b>
POS Tagged	Apple/NNP, keeps/VBZ, <b>stealing/VBG</b> , Samsung/NNP, 's/POS, <b>thunder/NN</b>
Bigram	apple keeps keeps stealing stealing samsung samsung 's 's thunder
Trigram	apple keeps stealing keeps stealing samsung stealing samsung 's samsung 's thunder
Constituency (excluding leaves)	(ROOT (S (NP (NNP Apple)) (VP (VBZ keeps) (VP (VBG stealing) (NP (NP (NNP Samsung) (POS 's)) (NN thunder)))))) (S (NP (NNP Apple)) (VP (VBZ keeps) (VP (VBG stealing) (NP (NP (NNP Samsung) (POS 's)) (NN thunder)))))) (NP (NNP Apple)) (VP (VBZ keeps) (VP (VBG stealing) (NP (NP (NNP Samsung) (POS 's)) (NN thunder)))) (VP (VBG stealing) (NP (NP (NNP Samsung) (POS 's)) (NN thunder))) (NP (NP (NNP Samsung) (POS 's)) (NN thunder)) (NP (NNP Samsung) (POS 's))
Dependency	nsubj(keeps, apple) dep(keeps, stealing) <b>dobj(stealing, thunder)</b> poss(thunder, samsung)
2) Apple stealing user information via Face Time	
Unigram	Apple, <b>stealing</b> , user, information, via, face, time
POS Tagged	Apple/NNP, <b>stealing/VBG</b> , user/NN, information/NN, via/PPREP, Face/NN, Time/NN
Bigram	Apple stealing, stealing user, user information, information via, via Face, Face Time
Trigram	apple stealing user, stealing user information, user information via, information via Face, via Face Time

Constituency (excluding leaves)	(ROOT (S (NP (NNP Apple)) (VP (VBG stealing) (NP (NN user) (NN information)) (PP (IN via) (NP (NNP Face) (NNP Time)))))) (S (NP (NNP Apple)) (VP (VBG stealing) (NP (NN user) (NN information)) (PP (IN via) (NP (NNP Face) (NNP Time)))))) (NP (NNP Apple)) (VP (VBG stealing) (NP (NN user) (NN information)) (PP (IN via) (NP (NNP Face) (NNP Time)))) (NP (NN user) (NN information)) (PP (IN via) (NP (NNP Face) (NNP Time))) (NP (NNP Face) (NNP Time))
Dependency	nsub(stealing, Apple) dobj(stealing, information) nn(information, user) nn(time, face) prep_via(stealing, time)

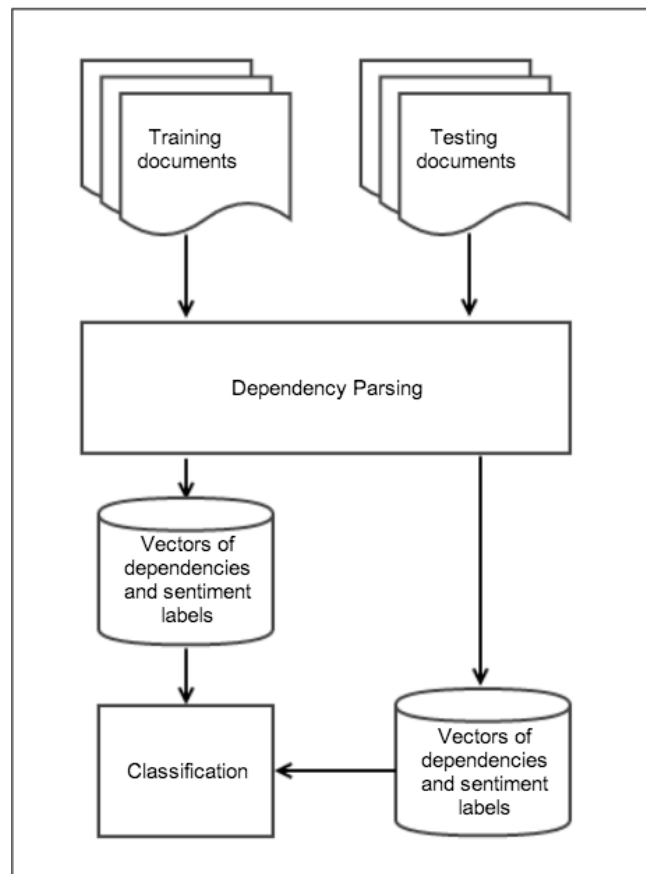
**Table 6. Similar Representation Elements in the Three Example Sentences**

Representation	Sentence 1	Sentence 2	Sentence 3
Unigram	stealing, thunder	stealing, thunder	Stealing
POS Tagged	stealing/VBG, thunder/NN	stealing/VBG, thunder/NN	stealing/VBG
Bi-gram			
Tri-gram			
Constituency (Excluding leaves)			
Dependency	dobj(stealing, thunder)	dobj(stealing, thunder)	

### 3.3 Method

In this study, we propose the use of dependency features to improve supervised sentiment classification. We compare proposed dependency features with n-gram features in terms of classification correctness.

To use the dependency as features, a document is parsed into dependencies. The occurrence of each dependency in the corpus is measured for each document to generate a vector space model. Then, a training data set containing dependency vectors and their sentiment categories are used to build a classifier.



**Figure 3. Sentiment Classification using Dependency Features**

### 3.4 Evaluation

To evaluate the effectiveness of the dependency features, we conducted sentiment classification on a large social media data set. The data set contains all user messages from Stocktwits between July 2009 and April 2014. Stocktwits is a leading social media platform for investors to share opinions about the financial market. Similar to tweets, Stocktwits messages are also limited to 140 characters. Instead of officially supporting Hashtags, Stocktwits uses Cashtags to track stocks and other financial assets mentioned in a message. On Stocktwits, users can mark the sentiment of their postings as either bullish or bearish. In our data set, there are 87,776 bearish and 265,452 bullish postings in the data set. The experimental task is to classify each message as bullish or bearish.

To minimize the effect of random guesses, we sampled equal number of bullish and bearish messages for cross validation. We first substituted mentions, Cashtags, and URLs as USER, CASHTAG, and URL. Then, we removed messages containing only these terms because such messages provide no information of user opinions. Then, all messages were tokenized. We counted the number of tokens in each message, excluding USER, CASHTAG, and URL. The token counts ranged from 1 to 67, with a mean of 13.9 and a standard deviation of 7.4. To match bullish message with bearish messages, we divided each group into 8 bins based on token count. In each bin, we randomly sampled the same number of bullish messages to match the number of bearish messages. Eventually, 170,874 message are retained, half of which were bullish, and the other half, which were bearish.

To generate the dependency features, we first used the CMU Ark Tweet POS Tagger to tag the Stocktwits messages. Then, we used the Stanford Parser to parse the tagged messages

into dependencies. We substituted terms containing numbers as <num>. Capitalization was not retained.

For each model, we performed 10 times 10-fold cross validation. All experiments were run using the Scikit-Learn package in Python. The following classification metrics were used. All metrics, except accuracy, were calculated for each class. Next, the weighted average was calculated as the overall performance of a model.

$$\textit{Precision} = \frac{\textit{number of correctly classified documents in the class}}{\textit{total number of documents classified as the class}}$$

$$\textit{Recall} = \frac{\textit{number of correctly classified documents for the class}}{\textit{total number of documents in the class}}$$

$$\textit{Accuracy} = \frac{\textit{number of correctly classified documents}}{\textit{total number of documents}}$$

$$\textit{F - measure} = \frac{2 \cdot \textit{precision} \cdot \textit{recall}}{\textit{precision} + \textit{recall}}$$

Four different experiments were conducted. Experiment 1 compared the combination of unigram features and dependency features (coded as 1G+DEP) against the combination of unigrams and bigrams (coded as 1G+2G). Experiment 2 compared the 1G+DEP against the combination of unigrams, bigrams, and trigrams (coded as 1G+2G+3G). Experiment 3 compares the 1G+DEP against POS-tagged features (coded as POS). Experiment 4 compares dependency features (coded as DEP) against bigrams (coded as 2G) without using any unigrams.

To show the robustness of the results, each experiment was conducted using different classifiers, both frequency and presence quantifiers, with and without inverse document

frequency. The two classifiers used are Naïve Bayes (NB) and Linear Support Vector Machine (SVM). We also incorporated Chi-square feature selection to test the sensitivity of the results to feature selection. The features with the Top 10% largest Chi-square values are used to train and test the model in each run.

### **Experiment 1: 1G+DEP vs. 1G+2G**

Table 5 shows the F-measure, accuracy, precision, and recall of unigram features combined with dependencies and bigrams, respectively. All settings using the dependency and unigram features (1G+DEP) achieved an improvement over those using bigram and unigram features (1G+2G). The improvement ranges from 0.47% to 1.96%. We conducted t-tests and found that all differences are significant at the 5% level. The results from the NB classifier are slightly better than those from SVM. The binary-IDF quantification achieved the best result among all settings.

### **1G+DEP vs. 1G+2G (with feature selection)**

Table 6 exhibits the results of the same settings with feature selection. The improvement of using 1G+DEP is even larger after feature selection. Based on t-tests, all improvements in F-measure and accuracy are significant at the 5% level. The largest improvement in accuracy is 2.34%, using NB and the binary-IDF measures. However, the overall results did not improve, compared to those of the experiment without feature selection. We believe this is because our Top 10% Chi-square feature selection approach is relatively simplistic and both NB and SVM are already fairly robust to high dimensional data.



Table 5. Unigram + Dependency vs. Unigram + Bigram

<b>F-measure</b>					
<b>Classifier</b>	<b>Measure</b>	<b>IDF</b>	<b>1G+DEP</b>	<b>1G+2G</b>	<b>Improvement</b>
NB	Frequency	No	0.7909	0.7872	0.47%
		Yes	0.7894	0.7861	0.42%
	Binary	No	0.7904	0.7769	1.74%
		Yes	0.8082	0.7927	1.96%
SVM	Frequency	No	0.7896	0.7834	0.79%
		Yes	0.7887	0.7842	0.57%
	Binary	No	0.7902	0.7767	1.74%
		Yes	0.8072	0.7926	1.84%
<b>Accuracy</b>					
<b>Classifier</b>	<b>Measure</b>	<b>IDF</b>	<b>1G+DEP</b>	<b>1G+2G</b>	<b>Improvement</b>
NB	Frequency	No	0.7919	0.7858	0.78%
		Yes	0.7894	0.7847	0.60%
	Binary	No	0.7896	0.7748	1.91%
		Yes	0.8067	0.7905	2.05%
SVM	Frequency	No	0.7912	0.7839	0.93%
		Yes	0.7889	0.7836	0.68%
	Binary	No	0.7893	0.7745	1.91%
		Yes	0.8065	0.7906	2.01%
<b>Precision</b>					
<b>Classifier</b>	<b>Measure</b>	<b>IDF</b>	<b>1G+DEP</b>	<b>1G+2G</b>	<b>Improvement</b>
NB	Frequency	No	0.7946	0.782	1.61%

		Yes	0.7892	0.7809	1.06%
	Binary	No	0.7873	0.7696	2.30%
		Yes	0.8017	0.7844	2.21%
SVM	Frequency	No	0.7956	0.7853	1.31%
		Yes	0.7896	0.782	0.97%
	Binary	No	0.7867	0.7692	2.28%
		Yes	0.8043	0.7851	2.45%
<b>Recall</b>					
<b>Classifier Measure IDF 1G+DEP 1G+2G Improvement</b>					
NB	Frequency	No	0.7873	0.7925	-0.66%
		Yes	0.7897	0.7914	-0.21%
	Binary	No	0.7936	0.7845	1.16%
		Yes	0.8148	0.8012	1.70%
SVM	Frequency	No	0.7836	0.7815	0.27%
		Yes	0.7877	0.7864	0.17%
	Binary	No	0.7938	0.7844	1.20%
		Yes	0.8101	0.8003	1.22%

Table 6. Unigram + Dependency vs. Unigram + Bigram (Feature Selected)

F-measure					
Classifier	Measure	IDF	1G+DEP	1G+2G	Improvement
NB	Frequency	No	0.7876	0.7798	1.00%
		Yes	0.7855	0.7804	0.65%
	Binary	No	0.78	0.7733	0.87%
		Yes	0.8066	0.7903	2.06%
SVM	Frequency	No	0.7863	0.7779	1.08%
		Yes	0.7846	0.7782	0.82%
	Binary	No	0.7793	0.7729	0.83%
		Yes	0.8062	0.7900	2.05%
Accuracy					
Classifier	Measure	IDF	1G+DEP	1G+2G	Improvement
NB	Frequency	No	0.7885	0.7785	1.28%
		Yes	0.7860	0.7780	1.03%
	Binary	No	0.7780	0.7684	1.25%
		Yes	0.8054	0.7870	2.34%
SVM	Frequency	No	0.7874	0.7774	1.29%
		Yes	0.7856	0.7775	1.04%
	Binary	No	0.7773	0.7680	1.21%
		Yes	0.8050	0.7868	2.31%
Precision					
Classifier	Measure	IDF	1G+DEP	1G+2G	Improvement
NB	Frequency	No	0.7908	0.7752	2.01%
		Yes	0.7875	0.7721	1.99%

	Binary	No	0.7731	0.7574	2.07%
		Yes	0.8019	0.7785	3.01%
SVM	Frequency	No	0.7904	0.7761	1.84%
		Yes	0.7882	0.7757	1.61%
	Binary	No	0.7724	0.7572	2.01%
		Yes	0.8013	0.7784	2.94%
<b>Recall</b>					
<b>Classifier Measure IDF 1G+DEP 1G+2G Improvement</b>					
NB	Frequency	No	0.7846	0.7849	-0.04%*
		Yes	0.7834	0.7889	-0.70%
	Binary	No	0.7873	0.7907	-0.43%
		Yes	0.8114	0.8024	1.12%
SVM	Frequency	No	0.7825	0.7801	0.31%
		Yes	0.7811	0.7809	0.03%*
	Binary	No	0.7868	0.7901	-0.42%
		Yes	0.8113	0.8019	1.17%

\* Insignificant at 5%.

### **Experiment 2: 1G+DEP vs. 1G+2G+3G**

To further verify the effectiveness of the dependency features, we add trigram features to the baseline. Table 7 exhibits the metrics comparison between 1G+DEP and 1G+2G+3G. Using trigrams, in addition to unigrams and bigrams, slightly improved classification accuracy. However, the 1G+DEP feature set still outperforms this baseline by up to 1.77% in terms of accuracy. All differences in F-measure and accuracy are significant at the 5% level using a t-test. This experiment was run again with feature selection. The

results are shown in Table 8. Once again, the 1G+DEP feature set improves the baseline by over 2% in accuracy. All results are significant at the 5% level.

### Experiment 3: 1G+DEP vs. POS

Table 9 compares the classification performance between 1G+DEP and POS-tagged words (POS). The former improved classification accuracy by up to 2.54%. All differences are significant at 5%. However, with feature selection, the improvement is less consistent across different settings. In some settings, the F-measure and accuracy dropped. This is mainly due to the increase in the recall of POS after feature selection.

**Table 7. 1G+DEP vs. 1G+2G+3G**

<b>F-measure</b>					
<b>Classifier</b>	<b>Measure</b>	<b>IDF</b>	<b>1G+DEP</b>	<b>1G+2G+3G</b>	<b>Improvement</b>
NB	Frequency	No	0.7909	0.7875	0.43%
		Yes	0.7894	0.7862	0.41%
	Binary	No	0.7904	0.7809	1.22%
		Yes	0.8082	0.7956	1.58%
SVM	Frequency	No	0.7896	0.7848	0.61%
		Yes	0.7887	0.7839	0.61%
	Binary	No	0.7902	0.7809	1.19%
		Yes	0.8072	0.7954	1.48%
<b>Accuracy</b>					
<b>Classifier</b>	<b>Measure</b>	<b>IDF</b>	<b>1G+DEP</b>	<b>1G+2G+3G</b>	<b>Improvement</b>
NB	Frequency	No	0.7919	0.7863	0.71%
		Yes	0.7894	0.7846	0.61%

	Binary	No	0.7896	0.7786	1.41%
		Yes	0.8067	0.7927	1.77%
SVM	Frequency	No	0.7912	0.7853	0.75%
		Yes	0.7889	0.7833	0.71%
	Binary	No	0.7893	0.7787	1.36%
		Yes	0.8065	0.7929	1.72%
<b>Precision</b>					
<b>Classifier Measure IDF 1G+DEP 1G+2G+3G Improvement</b>					
NB	Frequency	No	0.7946	0.7831	1.47%
		Yes	0.7892	0.7805	1.11%
	Binary	No	0.7873	0.7728	1.88%
		Yes	0.8017	0.7848	2.15%
SVM	Frequency	No	0.7956	0.7867	1.13%
		Yes	0.7896	0.7818	1.00%
	Binary	No	0.7867	0.7732	1.75%
		Yes	0.8043	0.7859	2.34%
<b>Recall</b>					
<b>Classifier Measure IDF 1G+DEP 1G+2G+3G Improvement</b>					
NB	Frequency	No	0.7873	0.7919	-0.58%*
		Yes	0.7897	0.7919	-0.28%
	Binary	No	0.7936	0.7893	0.54%
		Yes	0.8148	0.8067	1.00%
SVM	Frequency	No	0.7836	0.7829	0.09%
		Yes	0.7877	0.7861	0.20%
	Binary	No	0.7938	0.7888	0.63%
		Yes	0.8101	0.8052	0.61%

\* Insignificant at 5%.

**Table 8. 1G+DEP vs. 1G+2G+3G (Feature Selected)**

<b>F-measure</b>					
<b>Classifier</b>	<b>Measure</b>	<b>IDF</b>	<b>1G+DEP</b>	<b>1G+2G+3G</b>	<b>Improvement</b>
NB	Frequency	No	0.7876	0.7825	0.65%
		Yes	0.7855	0.7810	0.58%
	Binary	No	0.7800	0.7660	1.83%
		Yes	0.8066	0.7934	1.66%
SVM	Frequency	No	0.7863	0.7807	0.72%
		Yes	0.7846	0.7789	0.73%
	Binary	No	0.7793	0.7656	1.79%
		Yes	0.8062	0.7931	1.65%
<b>Accuracy</b>					
<b>Classifier</b>	<b>Measure</b>	<b>IDF</b>	<b>1G+DEP</b>	<b>1G+2G+3G</b>	<b>Improvement</b>
NB	Frequency	No	0.7885	0.7802	1.06%
		Yes	0.7860	0.7784	0.98%
	Binary	No	0.7780	0.7625	2.03%
		Yes	0.8054	0.7905	1.88%
SVM	Frequency	No	0.7874	0.7795	1.01%
		Yes	0.7856	0.7777	1.02%
	Binary	No	0.7773	0.7623	1.97%
		Yes	0.8050	0.7904	1.85%
<b>Precision</b>					
<b>Classifier</b>	<b>Measure</b>	<b>IDF</b>	<b>1G+DEP</b>	<b>1G+2G+3G</b>	<b>Improvement</b>
NB	Frequency	No	0.7908	0.7745	2.10%

<b>Recall</b>					
<b>Classifier Measure IDF 1G+DEP 1G+2G+3G Improvement</b>					
SVM	Binary	Yes	0.7875	0.7719	2.02%
		No	0.7731	0.7551	2.38%
	Frequency	Yes	0.8019	0.7824	2.49%
		No	0.7904	0.7767	1.76%
	Binary	Yes	0.7882	0.7748	1.73%
		No	0.7724	0.7551	2.29%
Frequency	Yes	0.7882	0.7748	1.73%	
	No	0.7904	0.7767	1.76%	
NB	Binary	Yes	0.8114	0.8048	0.82%
		No	0.7873	0.7776	1.25%
	Frequency	Yes	0.7834	0.7902	-0.86%
		No	0.7846	0.7908	-0.78%
SVM	Binary	Yes	0.8113	0.8036	0.96%
		No	0.7868	0.7768	1.29%
	Frequency	Yes	0.7811	0.7830	-0.24%
		No	0.7825	0.7848	-0.29%

Table 9. 1G+DEP vs. POS

<b>F-measure</b>					
<b>Classifier Measure IDF 1G+DEP POS Improvement</b>					
NB	Binary	Yes	0.8082	0.7885	2.50%
		No	0.7904	0.7766	1.78%
	Frequency	Yes	0.7894	0.7717	2.29%
		No	0.7909	0.7765	1.85%



SVM	Frequency	No	0.7896	0.7760	1.75%
		Yes	0.7887	0.7718	2.19%
	Binary	No	0.7902	0.7764	1.78%
		Yes	0.8072	0.7885	2.37%
<b>Accuracy</b>					
<b>Classifier Measure IDF 1G+DEP POS Improvement</b>					
NB	Frequency	No	0.7919	0.7766	1.97%
		Yes	0.7894	0.7716	2.31%
	Binary	No	0.7896	0.7742	1.99%
		Yes	0.8067	0.7867	2.54%
SVM	Frequency	No	0.7912	0.7765	1.89%
		Yes	0.7889	0.7720	2.19%
	Binary	No	0.7893	0.7739	1.99%
		Yes	0.8065	0.7870	2.48%
<b>Precision</b>					
<b>Classifier Measure IDF 1G+DEP POS Improvement</b>					
NB	Frequency	No	0.7946	0.7769	2.28%
		Yes	0.7892	0.7713	2.32%
	Binary	No	0.7873	0.7684	2.46%
		Yes	0.8017	0.7818	2.55%
SVM	Frequency	No	0.7956	0.7777	2.30%
		Yes	0.7896	0.7726	2.20%
	Binary	No	0.7867	0.7680	2.43%
		Yes	0.8043	0.7830	2.72%
<b>Recall</b>					
<b>Classifier Measure IDF 1G+DEP POS Improvement</b>					
NB	Frequency	No	0.7873	0.7761	1.44%

		Yes	0.7897	0.7721	2.28%
	Binary	No	0.7936	0.7850	1.10%
		Yes	0.8148	0.7953	2.45%
SVM	Frequency	No	0.7836	0.7743	1.20%
		Yes	0.7877	0.7709	2.18%
	Binary	No	0.7938	0.7851	1.11%
		Yes	0.8101	0.7940	2.03%

**Table 10. 1G+DEP vs. POS (Feature Selected)**

<b>F-measure</b>					
<b>Classifier</b>	<b>Measure</b>	<b>IDF</b>	<b>1G+DEP</b>	<b>POS</b>	<b>Improvement</b>
NB	Frequency	No	0.7876	0.7718	2.05%
		Yes	0.7855	0.7677	2.32%
	Binary	No	0.78	0.786	-0.76%
		Yes	0.8066	0.7888	2.26%
SVM	Frequency	No	0.7863	0.771	1.98%
		Yes	0.7846	0.7671	2.28%
	Binary	No	0.7793	0.7857	-0.81%
		Yes	0.8062	0.7885	2.24%
<b>Accuracy</b>					
<b>Classifier</b>	<b>Measure</b>	<b>IDF</b>	<b>1G+DEP</b>	<b>POS</b>	<b>Improvement</b>
NB	Frequency	No	0.7885	0.7718	2.16%
		Yes	0.7860	0.7672	2.45%
	Binary	No	0.7780	0.7813	-0.42%
		Yes	0.8054	0.7852	2.57%
SVM	Frequency	No	0.7874	0.7714	2.07%

		Yes	0.7856	0.7677	2.33%
	Binary	No	0.7773	0.7804	-0.40%
		Yes	0.8050	0.7853	2.51%
<b>Precision</b>					
<b>Classifier Measure IDF 1G+DEP POS Improvement</b>					
NB	Frequency	No	0.7908	0.7717	2.48%
		Yes	0.7875	0.7661	2.79%
	Binary	No	0.7731	0.7694	0.48%
		Yes	0.8019	0.7758	3.36%
SVM	Frequency	No	0.7904	0.7722	2.36%
		Yes	0.7882	0.7691	2.48%
	Binary	No	0.7724	0.7674	0.65%
		Yes	0.8013	0.7768	3.15%
<b>Recall</b>					
<b>Classifier Measure IDF 1G+DEP POS Improvement</b>					
NB	Frequency	No	0.7846	0.7721	1.62%
		Yes	0.7834	0.7694	1.82%
	Binary	No	0.7873	0.8033	-1.99%
		Yes	0.8114	0.8023	1.13%
SVM	Frequency	No	0.7825	0.7699	1.64%
		Yes	0.7811	0.7652	2.08%
	Binary	No	0.7868	0.8049	-2.25%
		Yes	0.8113	0.8005	1.35%

#### Experiment 4: DEP vs. 2G

To compare dependency directly against bigrams, our last experiment uses each type of features without unigram for sentiment classification. Table 11 shows that dependency features alone can improve classification accuracy by up to 1.62% over bigram features. All differences in F-measure and accuracy are significant at the 5% level. The improvement is more pronounced after feature selection.

**Table 11. Dependency vs. Bi-gram**

F-measure					
Classifier	Measure	IDF	DEP	2G	Improvement
NB	Frequency	No	0.7581	0.7553	0.37%
		Yes	0.7589	0.7546	0.57%
	Binary	No	0.7588	0.7562	0.34%
		Yes	0.7595	0.7554	0.54%
SVM	Frequency	No	0.7452	0.7406	0.62%
		Yes	0.7613	0.7518	1.26%
	Binary	No	0.7454	0.7406	0.65%
		Yes	0.7619	0.7517	1.36%
Accuracy					
Classifier	Measure	IDF	DEP	2G	Improvement
NB	Frequency	No	0.7610	0.7556	0.71%
		Yes	0.7600	0.7547	0.70%
	Binary	No	0.7452	0.7358	1.28%
		Yes	0.7613	0.7492	1.62%
SVM	Frequency	No	0.7606	0.7550	0.74%

		Yes	0.7597	0.7543	0.72%
	Binary	No	0.7450	0.7355	1.29%
		Yes	0.7608	0.7492	1.55%
<b>Precision</b>					
<b>Classifier Measure IDF DEP 2G Improvement</b>					
NB	Frequency	No	0.7659	0.7544	1.52%
		Yes	0.7612	0.7534	1.04%
	Binary	No	0.7449	0.7273	2.42%
		Yes	0.7601	0.7445	2.10%
SVM	Frequency	No	0.7659	0.7543	1.54%
		Yes	0.7614	0.7537	1.02%
	Binary	No	0.7448	0.7266	2.50%
		Yes	0.7595	0.7441	2.07%
<b>Recall</b>					
<b>Classifier Measure IDF DEP 2G Improvement</b>					
NB	Frequency	No	0.7519	0.758	-0.80%
		Yes	0.7577	0.7574	0.04%*
	Binary	No	0.7459	0.7545	-1.14%
		Yes	0.7636	0.759	0.61%*
SVM	Frequency	No	0.7506	0.7564	-0.77%
		Yes	0.7565	0.7557	0.11%
	Binary	No	0.7456	0.7553	-1.28%
		Yes	0.7632	0.7597	0.46%

\* Insignificant at 5% level.

Table 12. DEP vs. 2G (Feature Selected)

F-measure					
Classifier	Measure	IDF	DEP	2G	Improvement
NB	Frequency	No	0.7483	0.7327	2.13%
		Yes	0.7487	0.7328	2.17%
	Binary	No	0.7465	0.7376	1.21%
		Yes	0.7635	0.7549	1.14%
SVM	Frequency	No	0.748	0.7321	2.17%
		Yes	0.7481	0.7322	2.17%
	Binary	No	0.7443	0.7379	0.87%
		Yes	0.7629	0.7546	1.10%*
Accuracy					
Classifier	Measure	IDF	DEP	2G	Improvement
NB	Frequency	No	0.7511	0.7352	2.16%
		Yes	0.7511	0.7352	2.16%
	Binary	No	0.7409	0.7316	1.27%
		Yes	0.7600	0.7430	2.29%
SVM	Frequency	No	0.7510	0.7347	2.22%
		Yes	0.7507	0.7349	2.15%
	Binary	No	0.7396	0.7308	1.20%
		Yes	0.7593	0.7426	2.25%
Precision					
Classifier	Measure	IDF	DEP	2G	Improvement
NB	Frequency	No	0.7569	0.7404	2.23%
		Yes	0.7561	0.7401	2.16%
	Binary	No	0.7308	0.723	1.08%
		Yes	0.7526	0.7214	4.32%

SVM	Frequency	No	0.7572	0.74	2.32%
		Yes	0.756	0.7403	2.12%
	Binary	No	0.7314	0.7207	1.48%
		Yes	0.7516	0.7211	4.23%
<b>Recall</b>					
<b>Classifier Measure IDF DEP 2G Improvement</b>					
NB	Frequency	No	0.7404	0.7289	1.58%
		Yes	0.7417	0.7291	1.73%
	Binary	No	0.7635	0.7592	0.57%
		Yes	0.7747	0.7916	-2.13%
SVM	Frequency	No	0.7394	0.7281	1.55%*
		Yes	0.7407	0.7278	1.77%
	Binary	No	0.7588	0.7629	-0.54%*
		Yes	0.7745	0.7913	-2.12%

\* Insignificant at 5% level.

### 3.5 Conclusion

In this study, we proposed the use of dependency features in supervised sentiment classification. Dependency's capability to provide context for words and flexibility to allow distant relationships enable it to achieve better performance over multi-gram features and POS features. The results are robust for both large and small feature sets. The effectiveness of dependency largely relies on the parsing accuracy.

Based on our findings in this study, we have identified a few future research directions. This study tested the use of dependency feature using short social media texts. We are interested to see how dependency features influence the classification of longer texts. Moreover, this study achieved improved sentiment classification results by combining

dependency and unigram features. Such a feature set has certain redundancies since many dependencies can be reduced to single words. Developing an effective feature selection method for dependencies is a potential avenue for future research. While this study focused on the supervised sentiment classification, the lexicon-based sentiment classification approach could also benefit from the use of dependency information. Future effort can be directed to developing a sentiment lexicon of dependencies.



## CHAPTER 4

### Essay 3: To Predict or to be Predicted? An Empirical Study on Social Media Sentiment and Stock Return

#### 1. Introduction

Social media messages, as one of the major types of big data, are believed to be the largest data source for public opinion (Bifet and Frank 2010; O'Leary 2011; Tsytsarau and Palpanas 2012; Yu et al. 2013). Individuals are driven to express their emotions and opinions on social media (Aggarwal et al. 2012b). The boom of mobile platforms in recent years has removed most limits on where and when they want to post a message. With public opinion flows generating at such high volume and velocity, organizations are given the unprecedented opportunity to monitor their customers and competitors (Jacobs 2009). Moreover, it even seems promising to use the “crowd wisdom” on social media to predict economic activities. For instance, Tumasjan et al. (2011) found that party mentions on Twitter predicted German federal election results. Several studies also found that social media activities helped predict product sales (Dewan and Ramaprasad 2014; Mishne and Glance 2006; Rui et al. 2013).

In recent years, social media has been studied for its predictive power in the stock market. One particular type of information extracted from social media messages, *sentiment*, has received tremendous attention. Sentiment refers to the emotional state of users, such as happy, angry, nervous, etc. Most of these states can be categorized into two polarities, positive and negative. Sentiment analysis, as a major application of opinion mining, has been used to extract and assess the sentiment from text data (Abbasi et al. 2011; Chen and

Zimbra 2010). Institutional investors are now looking at sentiment reports of social media data to support their investment decisions (Fan and Gordon 2013). Hedge funds, such as Cayman Atlantic, have been established to execute trading strategies based on social media sentiment analysis. Bollen et al. (2011) were among the first researchers to explore the predictive power of social media sentiment on stock return. Both hedge funds and analytics providers claim that their models are based on Bollen et al. (2011)'s findings. Other studies also emerged to either develop predictive models to predict stock return using social media sentiment or test the hypothesis of predictability.

### **Research Questions**

While one may think it is promising to predict stock returns using social media sentiment, it doesn't take much effort to find out that related studies are fragmented and have contradictory results. Different data sets spanning different periods have been used. Some of them achieved amazingly high prediction accuracy while others found the predictive power of social media sentiment to be insignificant. Moreover, there has been lack of consensus on this theoretical foundation on how social media sentiment and stock return are associated. In this study, we examine the theoretical foundation and characterize this relationship. In particular, we address three research questions:

- 1) How is social media sentiment related to stock return?
- 2) Are positive sentiment and negative sentiment associated with stock return in the same way?
- 3) What are the dynamics for the above relationships?

Intuitively, social media users post messages related to the events that influence them.

Investors should in general react to the changes in stock prices in a consistent manner.

Surprisingly, this relationship has never been examined in related work. We believe that the seemingly significant correlation between social media sentiment and stock return is mainly driven by this relationship.

This study intends to resolve this confusion by investigating the interaction between social media sentiment and stock return. We examine this relationship at the daily level. We propose hypotheses based on theories from psychology and finance and test them using vector autoregression (VAR). VAR allows us to model social media sentiment and stock return in a dynamic system and capture the potential causality loop between variables.

### **Findings**

Using a data set containing 18,226,067 microblog messages that span over 5 years, we find that: 1) investor sentiment on social media does not predict stock return; 2) stock return influences investor sentiment on social media to a significant degree; 3) stock return has a much stronger influence on negative sentiment than on positive sentiment; 4) this influence reaches its maximum on the next day and decreases rapidly after that.

### **Contributions**

Our research makes several contributions. First, through rigorous testing on a data set that is much larger than those used in any related work, we show that investor sentiment is largely driven by stock return rather than being a driver. This serves to caution social media analysts when they develop predictive models using social media information. Second, we show that stock returns affect positive sentiment and negative sentiment differently. Based on these findings, public company management should execute

different strategies in managing public relations in social media when their stock price increases and decreases. The results are of significant interest to information systems researchers, social psychologists, and behavioral finance researchers.

### **Outline**

The rest of the chapter is organized as follow. In Section 2, we review related work. In Section 3, we discuss the theoretical background and formulate research hypotheses. In Section 4, we describe the data set and measurement. In Section 5, we describe the VAR model and other econometric analysis procedures used in this study. In Section 6, we present the results and supplement them with a test for robustness. In Section 7, we discuss the results, and the implications for research and practice. In Section 8, we discuss the limitations of this study and identify a set of future research directions.

## 2. Literature Review

**Table 13. Related Studies on Social Media Sentiment and Stock Return**

Study	Social media	Data size	Period	Length	Stocks	Type	Findings
Yu et al. 2013	Blogs, forum messages, and tweets	52746	07/01/2011 - 09/30/2011	3 months	824 individual stocks	Hypothesis testing	Forum and blog sentiment predicts stock return, but tweet sentiment doesn't.
Oliveira et al. 2013	StockTwits	364457*	06/01/2010 - 10/31/2012	5 months	5 individual stock, and SPX	Predictive modeling	StockTwits sentiment doesn't predict stock return.
Oh & Sheng 2011	StockTwits	72221	05/11/2010 - 08/07/2010	3 months	1909 individual stocks	Predictive modeling	StockTwits sentiment has strong predictive power on stock return.
Luo et al. 2013	Blogs	not reported	2007/08/01 - 2009/07/31	2 years	9 individual stocks	Hypothesis testing	Blog sentiment significantly predict stock return, but only explain a small proportion of variation.
Bollen et al. 2011	Tweets	342255**	02/28/2008 - 12/19/2008	10 months	DJIA	Both	86.7% direction accuracy on predicting DJIA. However, positive and negative sentiments do not predict DJIA return.
Nann et al. 2013	Tweets and forum messages	2971381	06/01/2011 - 11/30/2011	6 months	SPX	Predictive modeling	60.38% directional accuracy.
* Highest among the stocks studied. ** The number of messages used in the Granger causality test.							

Bollen et al. (2011) were among the first to explore the predictive power of social media sentiment on stock return. They used OpinionFinder (Wilson et al. 2005a) to classify tweets as either positive or negative. Using Granger causality test on a sample of 342,255 tweets, they found neither positive nor negative sentiment predicts the return of the Dow

Jones Industrial Average (DJIA). However, in another experiment, where they classified tweet sentiment into six emotion dimensions, they found the calm dimension significantly predicts DJIA return. Nonetheless, there has never been a theoretical explanation on why this predictive power exists. In a cross validation using a testing set of 13 trading days, they achieved an accuracy of 86.7% in predicting DJIA price movement direction. This accuracy is, in fact, amazing, considering finance researchers have always been debating if stock return is predictable at all (Ang and Bekaert 2007; Campbell and Thompson 2008). Since Bollen et al. (2011)'s used a small sample, it is likely that these results are merely by chance.

Oh and Sheng (2011) used sentiment extracted from Stocktwits, an investor-focused microblogging website, to predict individual stock return. They achieved F-measure values over 80% in F-measure when predicting price directions. However, their results also suffer from small sample bias. The data set they used contain 72,221 messages for 1,909 tickers over a 3-month period. That is on average less than half a message per day for each ticker. Since opinions on social media represents "wisdom of the crowd", there is no reason to believe that an individual social media message can have such high predictive power on a dependent variable that is so difficult to predict. Nann et al. (2013) developed a predictive model by aggregating sentiment from multiple social media sources and achieved over 60% directional accuracy in predicting S&P 500 return. However, this study also suffers from the small sample bias as in other related research. More recent studies found contradictory results. For instance, using econometric models on a panel data set over a 3-month period, Yu et al. (2013) found tweet sentiment does not significantly predict the price on individual stock returns. However, they found that

forum sentiments have significant predictive power. This contradicts the results from Das and Chen (2007), who found exactly the opposite. Oliveira et al. (2013) developed a predictive model using Stocktwits sentiment and compared it with traditional equity models using fundamental and technical indicators in predicting individual stock returns. They concluded that the sentiment model is not useful. Luo et al. (2013) used sentiment on blogs to predict individual stock return and found significant predictive power. They also found theoretical support for this result. However, the blogs used in the study are retrieved from news websites such as Engadget and Mashable. Authors of these blogs are usually professional writers rather than average social media users. Thus, these blogs have more similarity to news than to the crowd wisdom on social media. It has been well established in the finance literature that news does predict stock return (Tetlock 2007). Moreover, in their results, blog sentiment only accounts for 2.75% of the variance of the stock return. This magnitude is negligible if the overall variance of stock return is not largely explained by the model.

Table 13 summarizes the data sets used and the findings of these studies. The findings contradict each other. In our opinion, many of these findings are not conclusive due to the small sample size and short testing period. More interestingly, none of them considered the predictive power of stock return on social media sentiment. It is intuitive to ask if social media users discuss what has already happened in the stock market on a daily basis. If this is true, the tiny predictive power of social media sentiment on stock return may only attribute to the autocorrelation of return itself. This study differs from all related work by theoretically postulating and empirically testing a set of hypotheses on the effects of social media sentiment on stock return. It also differs from previous

research by using a much larger data set, spanning a much longer time period. This grants us much higher statistical power and enhances the validity of our conclusions.



### 3. Theoretical Background and Hypotheses

#### **H1: Investor sentiment vs. stock return**

It is widely believed that stock prices react to new information (Luo et al. 2013; Malkiel and Fama 1970). Online word-of-mouth (WOM) information, reflected in blogs and customer reviews, provides useful cues to a firm's future performance (Luo and Zhang 2013). Many studies have found that the online WOM of a product influences its sales positively (Dewan and Ramaprasad 2014; Rui et al. 2013). Social media sentiment is also a reflection of consumer satisfaction and creates a positive image for investors (Luo et al. 2013). Given that customer satisfaction has been found to increase firm value (Anderson et al. 2004), it seems reasonable to believe that social media sentiment predicts firm value in the form of stock return. However, previous studies have found contradictory results on the predicting power of social media sentiment. Out of curiosity, we test the following hypothesis again in our paper using a different data set than in previous studies:

*H1a: Social media sentiment does not predict daily stock return.*

Lo et al. (2005) found that traders have strong emotional response to stock return and other market events. Human emotion response has at least two components, arousal and valence (Deng and Poole 2010). The former refers to the non-specific and non-directional component while the latter refers to the directional component, which ranges from positive to negative. Valence is a result of the cognitive appraisal to an external stimulus. Investors, whose economic outcomes are closely associated with the financial market, are likely to be consistently influenced by stock return. In fact, by studying consumer sentiment survey and stock prices, Otoo (1999) found that an increase in stock price signals a good economic trend, which in turn boosts public sentiment. Emotions are not

just feelings. They also create impulse to act (Frijda 1986; Gross and Thompson 2007). One action commonly associated with emotion is to post messages on social media (Aggarwal et al. 2012a). From another perspective, the positive feedback trading theory posits that noise investors' beliefs about future stock return are heavily influenced by the return in the previous period (De Long et al. 1990b). Such investor behavior is referred to as extrapolative expectation. As a result, the latest stock market movement constantly updates the investors' forward-looking opinions.

We therefore hypothesize that:

*H1b: Daily stock return predicts social media sentiment.*

## **H2: Positive vs. negative**

Previous research in psychology and organizational studies found that people respond to positive and negative stimuli differently (Stieglitz and Dang-Xuan 2013). In particular, the emotional response to negative stimuli is stronger than that to positive stimuli. This “negativity bias” applies to a variety of domains. This phenomenon has also been studied extensively in social media. For instance, negative online WOM receives more attention than positive ones (Luo 2007). Aggarwal et al. (2012b) found that negative blogs attract more readers.

Research in behavioral finance has found that investors react to good news and bad news differently, especially for popular stocks (Barberis et al. 1998; Conrad et al. 2002). In particular, the stock market responds to bad news in a much stronger fashion than to good news. Due to extrapolative expectation (De Long et al. 1990b), investors, in general, infer

a stock's future based on its past performance (Conrad et al. 2002). Popular stocks typically have superior historical performance. This forms investors' prior beliefs about their return. As a result, when good news comes in, investors take it for granted and show little reaction (Conrad et al. 2002). However, if bad news of popular stocks is released, which strongly contradicts investors' prior beliefs, investors will exhibit greater reaction. This theory has been supported in many empirical studies. For instance, Tetlock (2007) observed that negative stock return leads to more pessimistic sentiment in the Wall Street Journal the next day. Garcia (2013) found that investors are more sensitive to economic downturns. In a controlled experiment, (Ko and Huang 2012) also observed that investors do not pay attention to news that conform to their prior beliefs. On social media platforms, discussions on popular stocks are dominant. When investors on social media observe negative stock return, they are likely to report more negative sentiment than when they observe positive stock return. Hence, I posit the following hypothesis:

*H2: Stock return has a stronger influence on negative sentiment on social media than on positive sentiment.*

#### **4. Data and Measurement**

We use a data set containing all messages on Stocktwits.com between September 2009 and September 2014. Stocktwits is a microblog website for investors and financial professionals to share information and ideas about financial markets (StockTwits 2014).

Similar to tweets, each posting is limited to 140 characters. The data set contains 18,226,067 messages. Over 9,000 financial assets were mentioned in these messages over the period. Such a large data set has never been used in related work. This grants us more statistical power (Garcia 2013).

To measure investor sentiment, we use SentiStrength (Thelwall et al. 2010) to classify each message into positive, negative, or neutral. While some previous research manually classified the sentiments (Aggarwal et al. 2012b), it is more practical to use a computational method for this task, given big data we are using from social media (Luo et al. 2013). Previous work that benchmarked 20 popular Twitter sentiment analysis tools have shown that SentiStrength yields the best classification accuracy among others (Abbasi et al. 2014). Given the similarity between Stocktwits and Twitter, we believe SentiStrength is currently our best choice for sentiment classification. We aggregated the number of positive and negative messages at the daily level. The overall positive sentiment on a specific day is measured as the ratio of the number of positive messages to the total number of messages on that day. The overall negative sentiment is calculated similarly. Using these ratios greatly reduces the difference in scale between sentiment measures and return. The overall sentiment is calculated as the difference between the overall positive sentiment and the overall negative sentiment. This is referenced as pessimism factor (Garcia 2013) or bull-bear-spread (Bormann 2013; Brown and Cliff 2004) in previous work. We also combine the messages on Saturdays and Sundays with that on Fridays. There is neither new information nor trading activity during the weekend (Garcia 2013). The data set also shows that there are much fewer messages during weekends than on weekdays.

We collected the historical prices of DJIA for the same period from Yahoo! Finance. To calculate return, we use the percentage change between the daily closing prices. We also subtract the daily Treasury bill rate from the return to get the risk premium. Use of the risk premium allows us to only consider meaningful variations of DJIA prices.

**Table 14. Summary Statistics of Variables (N = 1259)**

	Return	Positive	Negative	Sentiment Spread
Mean	-0.0002	0.1126	0.0851	0.0757
Standard Deviation	0.0092	0.0219	0.0111	0.0327
Min	-0.0556	0.0568	0.0449	-0.0581
Max	0.0423	0.1853	0.1183	0.3510

## 5. Model Specification

We use the vector autoregression (VAR) to model the dynamic interaction between return and social media sentiment. Both return and sentiment are treated as endogenous. Each variable is predicted by the lagged values of itself and the other variable. By doing so, the model captures the possible feedback loops among the variables. That is, return in the current period may predict the sentiment in the next period, which, in turn, predicts the return after the next period. Although we do not believe that sentiment has predictive power on return, we still model this effect for the sake of rigor and completeness.

$$\begin{bmatrix} Return_t \\ Sentiment_t \end{bmatrix} = \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} + \sum_{p=1}^p \begin{bmatrix} \pi_{11}^{t-p} & \pi_{12}^{t-p} \\ \pi_{21}^{t-p} & \pi_{22}^{t-p} \end{bmatrix} \begin{bmatrix} Return_{t-p} \\ Sentiment_{t-p} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix} \quad (1)$$

where  $C_i (i = 1, 2)$  are constants,  $\pi_{i,j} (i, j = 1, 2)$  are coefficients,  $p$  is lag length, and  $\varepsilon_i (i = 1, 2)$  are residuals. To test Hypothesis 2, we specify another model using positive sentiment and negative sentiment as separate variables.

$$\begin{bmatrix} Return_t \\ Positive_t \\ Negative_t \end{bmatrix} = \begin{bmatrix} C_1 \\ C_2 \\ C_3 \end{bmatrix} + \sum_{p=1}^p \begin{bmatrix} \pi_{11}^{t-p} & \pi_{12}^{t-p} & \pi_{13}^{t-p} \\ \pi_{21}^{t-p} & \pi_{22}^{t-p} & \pi_{23}^{t-p} \\ \pi_{31}^{t-p} & \pi_{32}^{t-p} & \pi_{33}^{t-p} \end{bmatrix} \begin{bmatrix} Return_{t-p} \\ Positive_{t-p} \\ Negative_{t-p} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \varepsilon_{3,t} \end{bmatrix} \quad (2)$$

where  $C_i (i = 1, 2, 3)$  are constants,  $\pi_{i,j} (i, j = 1, 2, 3)$  are coefficients,  $p$  is lag length, and  $\varepsilon_i (i = 1, 2, 3)$  are residuals. The lag lengths of both models are selected using the Bayesian information criterion (BIC). In this procedure, models with different lag lengths are estimated and the model having the lowest BIC is chosen.

Once the VAR model is estimated, we use impulse response functions (IRFs) to analyze how each variable responds to one unit of unexpected shock from another variable,

controlling for other changes. This also allows us to understand the dynamics between each pair of variables. We also use forecast error variance decomposition (FEVD) to examine each predictor's ability to explain the variance of each endogenous variable.

## 6. Results

To ensure the stability of the estimated parameters, we first use the test for unit root to check the stationarity of each variable. This is the precondition for finding time-invariant associations. We conduct the augmented Dickey-Fuller (ADF) test on each variable. The p-values of all the ADF tests are smaller than 0.001. This indicates that the estimates of the VAR model are stable. Next, we select the optimal lag length based on BIC. The optimal lag is 5 for both models. This corresponds to 5 trading days, or one week.

We report the regression coefficients of Model 1 in Table 15 and the results of the Granger causality tests in Table 16 (\*\*\*) indicates that results are significant at  $p < 0.01$ , \*\* indicates that results are significant at  $p < 0.05$ ). The first column from the left shows the predictors and the lag orders. The other two columns show the coefficients of the predictors in each of the two equations. In the return equations, only the Lag 5 of return is significant. None of the sentiment lags are significant in predicting return. The Granger causality tests show that sentiment does not “Granger cause” return. However, return significantly causes sentiment. Table 17 shows the results from the Forecasting Error Variance Decomposition (FEVD). Sentiment spread only explains 0.2% of the return variation. This value is not significant as the lower bound of the 95% confidence interval falls below zero. The  $R^2$  for the return equation and the sentiment spread equation are

1.7% and 47.9%, respectively. This indicates that return is almost not explained in the model but sentiment is largely explained. Return explains 14.5% of the sentiment variation. This is a fairly large effect. These results support Hypothesis 1b but not Hypothesis 1a.

We plot the impulse response functions of model 1 in Figure 4. The response of return to one unit shock on sentiment is not significantly different from zero. However, sentiment has a sharp increase in response to one unit shock on return on the next day. This effect decreases on Day 2 and eventually fades away after that.

**Table 15. Coefficients of Model 1**

	Dependent Variable	
	Return	Sentiment Spread
Return		
L1.	-0.058	0.377***
L2.	0.038	0.086
L3.	-0.050	0.018
L4.	-0.002	-0.042
L5.	-0.082***	-0.226***
Sentiment Spread		
L1.	-0.002	0.275***
L2.	-0.002	0.143***
L3.	0.002	0.133***
L4.	-0.008	0.108***
L5.	0.006	0.183***
*** significant at 0.01 ** significant at 0.05		

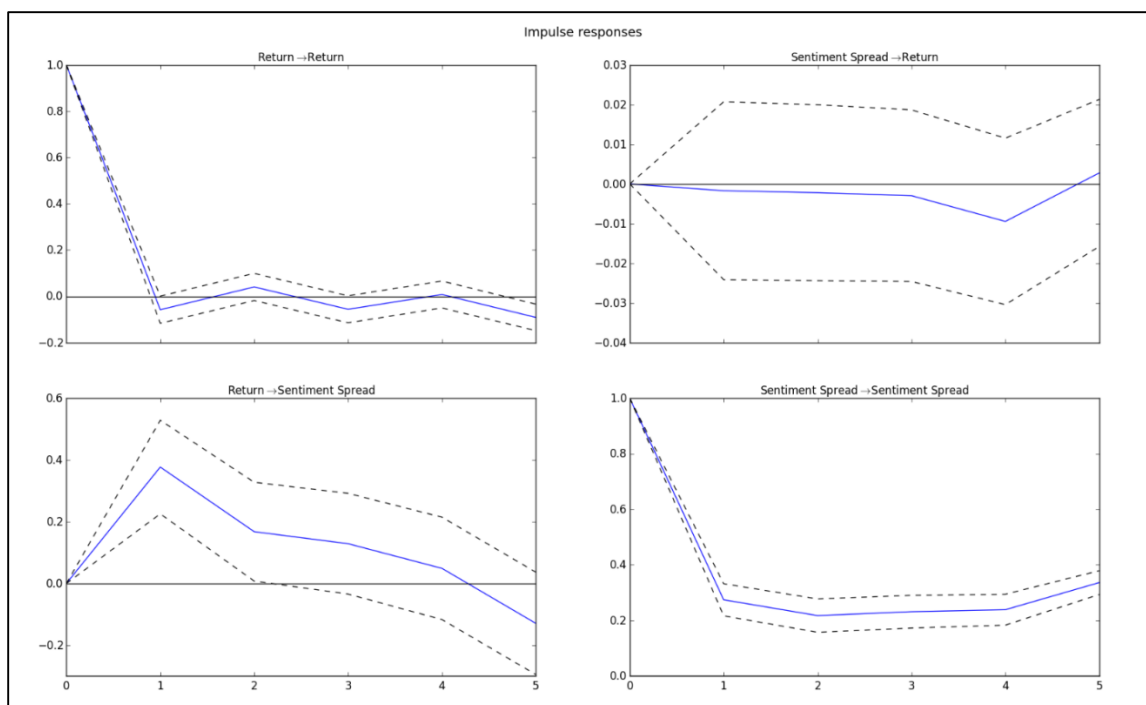
**Table 16. Granger Causality Test for Model 1**

**Table 17. FEVD for Model 1 (in %)**

	Dependent Variable	
	Return	Sentiment Spread
Return	99.98	14.56
Sentiment Spread	0.02	85.44

	Dependent Variable	
	Return	Sentiment Spread
Return	-	6.615***
Sentiment Spread	0.246	-
*** significant at 0.01 ** significant at 0.05		





**Figure 4. IRF Plot of Model 1**

For the equation of return, only Lag 5 of return, Lag 4 of positive sentiment, and Lag 5 of negative sentiment are significant. However, the Granger causality test shows that neither positive sentiment nor negative sentiment Granger causes return. For the equation of positive sentiment, all self-lags are significant and the Lag 5 of negative sentiment is significant. The Granger causality test shows that negative sentiment significantly predicts positive sentiment with ( $p = 0.045$ ). For the equation of negative sentiment, return is significant at Lags 1 and 5. All self-lags are significant. Positive sentiment is also significant at Lag 4. The results from Granger causality tests supports that return significantly Granger causes negative sentiment but not positive sentiment.

Next, we examine how each predictor explains the forecasting error variance in each equation. The  $R^2$  for the return, positive, and negative equations are 2.8%, 52.0%, and

46.5%, respectively. This indicates that the variance of return is only explained in a tiny fraction in the VAR system. However, the variances of positive and negative sentiments are largely explained. In the equation for return, positive sentiment and negative sentiment together only account for 0.6% of the variation. Neither of them is statistically significant. In the equation for positive sentiment, return accounts for 2.3% of the variation. In the equation for negative sentiment, return explains 25.0% of the variation. The positive and negative sentiment variances explained by return are significantly different. This supports Hypothesis 2.

Next, we examine the impulse response functions between the two polarities of sentiment and return. The responses of return to the shock of positive sentiment and the shock of negative sentiment are not significantly different from zero. Positive sentiment also exhibits no significant response to the shock from return. However, negative sentiment shows an immediate decrease in response to the shock of return. This effect wears in during the following period. The wear-in time refers to the number of periods it takes for the response to reach its peak (Luo et al. 2013).

**Table 18. Coefficients of Model 2**

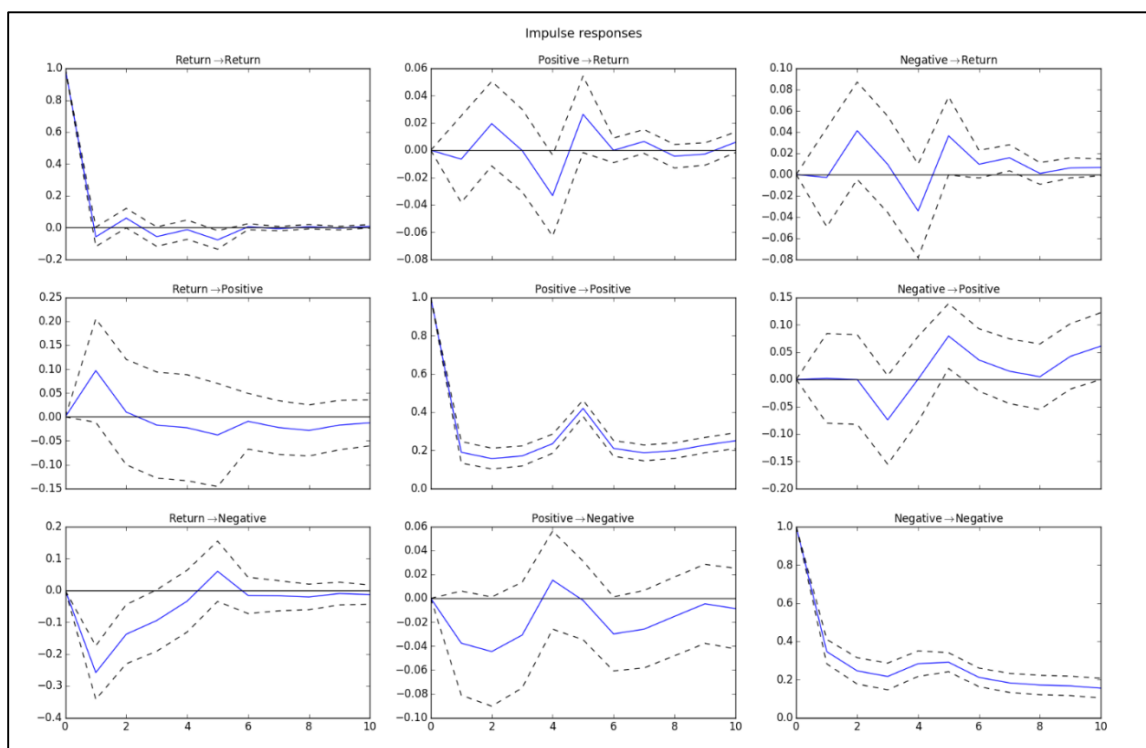
	Dependent Variable		
	Return	Positive	Negative
Return			
L1.	-0.058	0.097	-0.258***
L2.	0.056	-0.002	-0.06
L3.	-0.042	-0.037	0
L4.	-0.018	-0.049	0.03
L5.	-0.078***	-0.048	0.124***
Positive			
L1.	-0.007	0.19***	-0.038
L2.	0.02	0.122***	-0.026
L3.	0	0.116***	0.005
L4.	-0.035**	0.159***	0.046**
L5.	0.029	0.308***	0.002
Negative			
L1.	-0.003	0.002	0.346***
L2.	0.042	-0.001	0.125***
L3.	-0.002	-0.079	0.098***
L4.	-0.046	0.041	0.146***
L5.	0.045**	0.098**	0.081**
*** significant at 0.01 ** significant at 0.05			

**Table 19. Granger Causality Test for Model 2**

	Dependent Variable		
	Return	Positive	Negative
Return	-	0.999	9.068***
Positive	1.612	-	1.557
Negative	2.043	2.267**	-
*** significant at 0.01 ** significant at 0.05			

**Table 20. FEVD for Model 2 (in %)**

	Dependent Variable		
	Return	Positive	Negative
R-square	2.8	52.0	46.5
Return	99.4	2.3	25.0
Positive	0.2	97.4	8.5
Negative	0.4	0.3	66.5



**Figure 5. IRF Plot for Model 2**

### Robustness Tests

Sentiment classification typically does not achieve perfect accuracy. One may reasonably doubt if the VAR results are affected by this fact. To examine how the bias of using SentiStrength affects the results, we used Senti140, another effective sentiment analysis tool, to classify the sentiment again (Abbasi et al. 2014). For lag selection, the optimal lag for both Model 1 and 2 is 5. This is consistent with previous results. For

Model 1, Lag 5 of sentiment significantly predicts return with a p value of 0.015. However, the Granger causality test shows that the causal effect is not significant. In contrast, return significantly Granger causes sentiment. The IRF plot also shows that the change of sentiment in response to one unit shock of return peaks at the next day and decreases in the following days. The FEVD shows similar results. For Model 2, the Granger causality test, IRF plot, and FEVD show similar results as in the previous section.

## 7. Discussion

In this essay, we proposed an alternative explanation on the relationship between social media sentiment and stock return. We tested our hypotheses using a large social media data set that spans over 5 years. The results suggest that social media sentiment does not predict stock return. Instead, stock return explains the variation of sentiment, especially the negative sentiment, to a large extent.

Re-examining the theoretical foundation, we found that it has never been discussed by any previous research why social media sentiment can (or cannot) predict daily stock return. While it seems reasonable that online WOM can predict a firm's long-run financial performance, it is unclear how it can explain the daily variation of stock prices. A public firm's product and service quality is a relatively stable characteristic of the firm and is not expected to vary on a daily basis. So is the perception of this characteristic by consumers. For a fixed group of consumers, their opinions toward a company are not likely to vary on a daily basis. We believe that this is the reason why social media sentiment does not have predictive power on daily stock return.

From another perspective, there are different user groups on social media. It is naïve to assume that the sentiment of different groups are associated with stock return in the same manner. While many social media users are consumers, there are also many investors. We believe there is a large difference between consumer media sentiment and investor sentiment. As discussed in related work, consumer sentiment reflects their satisfaction with products and services. Investor sentiment reflects investor's direct emotional response to stocks. There are also different types of investors. De Long et al. (1990a)

distinguish noise traders from rational investors. Oh and Sheng (2011) believe that noise traders are prevalent on investor-focused social media sites. Park et al. (2013) also find that noise traders are consistently attracted by online communities. Noise traders do not make their investment decisions based on a firm's economic prospects and are constantly seeking information from unofficial channels and sharing them. Noise traders tend to overestimate return. They hold beliefs on future performance and risks that are not based on facts (Baker and Wurgler 2007). Lo et al. (2005) found that emotional traders consistently fare badly in predicting stock return. As a result, the sentiment of noise traders on social media is not likely to consistently predict stock return. We believe this is the theoretical explanation on why Hypothesis 1a was not supported.

These results has a number of theoretical contributions. One important contribution is the revelation that stock return influences social media sentiment. Although the relationship between the two variables has been examined in a few related studies, this particular effect has been largely overlooked. Examining this effect helps closing a major gap in this stream of research. Most related research has treated social media sentiment as exogenous. Our results show that its endogeneity cannot be ignored. This finding also indicates that investor psychology, such as extrapolative expectation, characterized in the behavioral finance research, is also exhibited in the social media domain. Future research in behavioral finance can directly measure investor sentiment from social media messages.

This study also has theoretical contributions to research related to behavioral finance. Our results indicate that the discussion of the stock market on social media merely represents noise, rather than new information. There is no information on social media that has not

already been incorporated in the stock price. This is a distinctive difference between social media and traditional media. From a practical perspective, this finding cautions investors who intend to develop trading strategies based on social media information. For instance, Derwent Capital Markets, which established the first hedge fund to trade based on social media sentiment, was shut down in just three months (Bloomberg 2013). Our findings could serve as a caution to companies in regard to executing improper investment strategies.

Our results also show that investors exhibit biased opinions on social media, as demonstrated by the stronger response to negative stock return than to positive return. This cautions public firm managers to treat good news and bad news of their firms differently. While good news might not significantly enhance the image of a large public firm, bad news will significantly hurt it.

The IRF analysis in our results have shown that the influence of return on social media peaks on the next day. This has important implications to public company managers in understanding the timing for managing their public image on social media. Social media is an effective platform for information diffusion (Luo et al. 2013). In the stock market, where new information is critical, investors actively seek the latest market information. When the stock market fluctuates, noise traders are compelled to look for others' opinions.

That stock return predicts social media sentiment may not sound as exciting as the other way around. However, our findings still have significant managerial implications. Based on our findings, social media analytics firms should focus more on monitoring values rather than predicting values. For instance, after a marketing campaign, consumer



response to the campaign can be gauged using the change of social media sentiment. For a public company, the continuous decrease of its stock price could also lead to doubts of its financial strength. The managers can assess the degree of such doubts using social media sentiment. Our findings also indicate that managers should immediately access social media sentiment after an event since it takes a very short time for the response to peak on social media.

## **8. Conclusion**

As social media analytics are being widely applied today, how social media sentiment is associated with stock return is an important question in both research and practice.

Empirically analyzing this relationship using a large data set has been challenging, given the difficulty of obtaining and processing large-scale social media data. To our knowledge, this paper is the first to examine the bidirectional effects of this relationship.

This study has some limitations. First, the hypotheses in this study were tested on the market index level. Previous research in finance did find that investors react differently to the price fluctuation of popular stocks and non-popular stocks. While most stocks being discussed on social media are popular stocks, we do not know if our hypotheses will apply to non-popular stocks. As observed in the finance literature, noise traders may be able to influence small firm stock return under herding. It could be possible to recover the predictability of social media sentiment on certain small firm stocks.

Second, this study used a data set from a single type of social media that is used mostly by investors. Users on other social media platforms may exhibit different behaviors.

Future research could examine how the relationship between social media sentiment and stock return varies across different social media platforms.

## REFERENCES

- Abbasi, A., France, S., Zhang, Z., and Chen, H. 2011. "Selecting Attributes for Sentiment Classification Using Feature Relation Networks," *IEEE Transactions on Knowledge and Data Engineering* (23:3), pp. 447-462.
- Abbasi, A., Hassan, A., and Dhar, M. 2014. "Benchmarking Twitter Sentiment Analysis Tools," *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, Reykjavik, Iceland.
- Abbott, R., Walker, M., Anand, P., Tree, J. E. F., Bowmani, R., and King, J. 2011. "How Can You Say Such Things?!?: Recognizing Disagreement in Informal Political Argument," *Proceedings of the Workshop on Languages in Social Media*, Portland, Oregon: Association for Computational Linguistics, pp. 2-11.
- Aggarwal, R., Gopal, R., Gupta, A., and Singh, H. 2012a. "Putting Money Where the Mouths Are: The Relation between Venture Financing and Electronic Word-of-Mouth," *Information Systems Research* (23:3-Part-2), pp. 976-992.
- Aggarwal, R., Gopal, R., Sankaranarayanan, R., and Singh, P. V. 2012b. "Blog, Blogger, and the Firm: Can Negative Employee Posts Lead to Positive Outcomes?," *Information Systems Research* (23:2), pp. 306-322.
- Alexa. 2013. from <http://www.alexa.com/>
- Anderson, E. W., Fornell, C., and Mazvancheryl, S. K. 2004. "Customer Satisfaction and Shareholder Value," *Journal of Marketing* (68:4), pp. 172-185.
- Ang, A., and Bekaert, G. 2007. "Stock Return Predictability: Is It There?," *Review of Financial Studies* (20:3), pp. 651-707.
- Baker, M., and Wurgler, J. 2007. "Investor Sentiment in the Stock Market." National Bureau of Economic Research Cambridge, Mass., USA.
- Balasubramanyan, R., Cohen, W. W., Pierce, D., and Redlawsk, D. P. 2011. "What Pushes Their Buttons?: Predicting Comment Polarity from the Content of Political Blog Posts," *Proceedings of the Workshop on Languages in Social Media*: Association for Computational Linguistics, pp. 12-19.
- Barberis, N., Shleifer, A., and Vishny, R. 1998. "A Model of Investor Sentiment," *Journal of financial economics* (49:3), pp. 307-343.



- Chen, H., Chiang, R. H., and Storey, V. C. 2012. "Business Intelligence and Analytics: From Big Data to Big Impact," *MIS Quarterly* (36:4), pp. 1165-1188.
- Chen, H., and Zimbra, D. 2010. "Ai and Opinion Mining," *Intelligent Systems, IEEE* (25:3), pp. 74-80.
- Chomsky, N. 1965. *Aspects of the Theory of Syntax*. MIT press.
- Chomsky, N. 2002. *Syntactic Structures*. Walter de Gruyter.
- Chou, C.-H., Sinha, A. P., and Zhao, H. 2010. "A Hybrid Attribute Selection Approach for Text Classification," *Journal of the Association for Information Systems* (11:9), pp. 491-518.
- Church, K. W., and Hanks, P. 1990. "Word Association Norms, Mutual Information, and Lexicography," *Computational Linguistics* (16:1), pp. 22-29.
- Conrad, J., Cornell, B., and Landsman, W. R. 2002. "When Is Bad News Really Bad News?," *The Journal of Finance* (57:6), pp. 2507-2532.
- Covington, M. A. 2001. "A Fundamental Algorithm for Dependency Parsing," *Proceedings of the 39th annual ACM southeast conference*: Citeseer, pp. 95-102.
- Das, S. R., and Chen, M. Y. 2007. "Yahoo! For Amazon: Sentiment Extraction from Small Talk on the Web," *Management Science* (53:9), pp. 1375-1388.
- De Long, J. B., Shleifer, A., Summers, L. H., and Waldmann, R. J. 1990a. "Noise Trader Risk in Financial Markets," *Journal of Political Economy* (98:4), pp. 703-738.
- De Long, J. B., Shleifer, A., Summers, L. H., and Waldmann, R. J. 1990b. "Positive Feedback Investment Strategies and Destabilizing Rational Speculation," *The Journal of Finance* (45:2), pp. 379-395.
- Deng, L., and Poole, M. S. 2010. "Affect in Web Interfaces: A Study of the Impacts of Web Page Visual Complexity and Order," *MIS Quarterly* (34:4), pp. 711-730.
- Dewan, S., and Ramaprasad, J. 2014. "Social Media, Traditional Media, and Music Sales," *MIS Quarterly* (38:1), pp. 101-121.
- Ding, X., Liu, B., and Yu, P. S. 2008. "A Holistic Lexicon-Based Approach to Opinion Mining," *Proceedings of the International Conference on Web Search and Web Data Mining*: ACM, pp. 231-240.

- Divol, R., Edelman, D., and Sarrazin, H. 2012. "Demystifying Social Media," *McKinsey Quarterly* (2), pp. 66-77.
- Esuli, A., and Sebastiani, F. 2006. "Sentiwordnet: A Publicly Available Lexical Resource for Opinion Mining," *Proceedings of 5th Conference on Language Resources and Evaluation*, pp. pp. 417-422.
- Fan, W., and Gordon, M. D. 2013. "Unveiling the Power of Social Media Analytics," *forthcoming in Communications of the ACM*.
- Frijda, N. H. 1986. *The Emotions*. Cambridge University Press.
- Fu, T., Abbasi, A., Zeng, D., and Chen, H. 2012. "Sentimental Spidering: Leveraging Opinion Information in Focused Crawlers," *ACM Transactions on Information Systems* (30:4), p. 24.
- Garcia, D. 2013. "Sentiment During Recessions," *The Journal of Finance*.
- Gross, J. J., and Thompson, R. A. 2007. "Emotion Regulation: Conceptual Foundations," *Handbook of emotion regulation* (3, p. 24.
- Hatzivassiloglou, V., and McKeown, K. R. 1997. "Predicting the Semantic Orientation of Adjectives," *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics: Association for Computational Linguistics*, pp. 174-181.
- Hu, M., and Liu, B. 2004. "Mining and Summarizing Customer Reviews," *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: ACM*, pp. 168-177.
- Jacobs, A. 2009. "The Pathologies of Big Data," *Communications of the ACM* (52:8), pp. 36-44.
- Ko, K. J., and Huang, Z. 2012. "Persistence of Beliefs in an Investment Experiment," *Quarterly Journal of Finance* (02:01), p. 1250005.
- Liu, B. 2012. "Sentiment Analysis and Opinion Mining," *Synthesis Lectures on Human Language Technologies* (5:1), pp. 1-167.
- Liu, B., Hu, M., and Cheng, J. 2005. "Opinion Observer: Analyzing and Comparing Opinions on the Web," *Proceedings of the 14th international conference on World Wide Web: ACM*, pp. 342-351.

- Lo, A. W., Repin, D. V., and Steenbarger, B. N. 2005. "Fear and Greed in Financial Markets: A Clinical Study of Day-Traders," National Bureau of Economic Research.
- Loughran, T., and McDonald, B. 2011. "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10 - Ks," *The Journal of Finance* (66:1), pp. 35-65.
- Luo, X. 2007. "Consumer Negative Voice and Firm-Idiosyncratic Stock Returns," *Journal of Marketing* (71:3), pp. 75-88.
- Luo, X., and Zhang, J. 2013. "How Do Consumer Buzz and Traffic in Social Media Marketing Predict the Value of the Firm?," *Journal of Management Information Systems* (30:2), pp. 213-238.
- Luo, X., Zhang, J., and Duan, W. 2013. "Social Media and Firm Equity Value," *Information Systems Research* (24:1), pp. 146-163.
- Malkiel, B. G., and Fama, E. F. 1970. "Efficient Capital Markets: A Review of Theory and Empirical Work\*," *The journal of Finance* (25:2), pp. 383-417.
- Mishne, G., and Glance, N. 2006. "Predicting Movie Sales from Blogger Sentiment," *Proceedings of AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW 2006)*, pp. 301-304.
- Ngo-Ye, T. L., and Sinha, A. P. 2012. "Analyzing Online Review Helpfulness Using a Regression Relief-Enhanced Text Mining Method," *ACM Transactions on Management Information Systems* (3:2), pp. 1-20.
- O'Leary, D. E. 2011. "Blog Mining-Review and Extensions: "From Each According to His Opinion"," *Decision Support Systems* (51:4), pp. 821-830.
- Oh, C., and Sheng, O. R. L. 2011. "Investigating Predictive Power of Stock Micro Blog Sentiment in Forecasting Future Stock Price Directional Movement," *ICIS 2011 Proceedings*, pp. 57-58.
- Oliveira, N., Cortez, P., and Areal, N. 2013. "On the Predictability of Stock Market Behavior Using Stocktwits Sentiment and Posting Volume," in *Progress in Artificial Intelligence*. Springer, pp. 355-365.
- Otoo, M. W. 1999. "Consumer Sentiment and the Stock Market."
- Pang, B., and Lee, L. 2008. *Opinion Mining and Sentiment Analysis*. Now Pub.

- Pang, B., Lee, L., and Vaithyanathan, S. 2002. "Thumbs Up?: Sentiment Classification Using Machine Learning Techniques," *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*: Association for Computational Linguistics, pp. 79-86.
- Park, J., Konana, P., Gu, B., Kumar, A., and Raghunathan, R. 2013. "Information Valuation and Confirmation Bias in Virtual Communities: Evidence from Stock Message Boards," *Information Systems Research* (24:4), pp. 1050-1067.
- Qiu, G., Liu, B., Bu, J., and Chen, C. 2009. "Expanding Domain Sentiment Lexicon through Double Propagation," *Proceedings of the 21st international joint conference on Artificial intelligence*: Morgan Kaufmann Publishers Inc., pp. 1199-1204.
- Raina, R., Battle, A., Lee, H., Packer, B., and Ng, A. Y. 2007. "Self-Taught Learning: Transfer Learning from Unlabeled Data," *Proceedings of the 24th international conference on Machine learning*: ACM, pp. 759-766.
- Riloff, E., and Wiebe, J. 2003. "Learning Extraction Patterns for Subjective Expressions," *Proceedings of the 2003 conference on Empirical methods in natural language processing*: Association for Computational Linguistics, pp. 105-112.
- Rui, H., Liu, Y., and Whinston, A. 2013. "Whose and What Chatter Matters? The Effect of Tweets on Movie Sales," *Decision Support Systems* (55:4), pp. 863-870.
- Sakaki, T., Okazaki, M., and Matsuo, Y. 2010. "Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors," *Proceedings of the 19th international conference on World wide web*: ACM, pp. 851-860.
- Schumaker, R. P., Zhang, Y., Huang, C.-N., and Chen, H. 2012. "Evaluating Sentiment in Financial News Articles," *Decision Support Systems* (53:3), pp. 458-464.
- Stefan Nann, J. K., Detlef Schoder. 2013. "Predictive Analytics on Public Data - the Case of Stock Markets," *ECIS*.
- Stieglitz, S., and Dang-Xuan, L. 2013. "Emotions and Information Diffusion in Social Media—Sentiment of Microblogs and Sharing Behavior," *Journal of Management Information Systems* (29:4), pp. 217-248.
- StockTwits. 2014. "A Communications Platform for the Investing Community." Retrieved June 15, 2014, from <http://stocktwits.com/about>



- Stone, P. J., Dunphy, D. C., and Smith, M. S. 1966. "The General Inquirer: A Computer Approach to Content Analysis."
- Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. 2011. "Lexicon-Based Methods for Sentiment Analysis," *Computational Linguistics* (37:2), pp. 267-307.
- Tetlock, P. C. 2007. "Giving Content to Investor Sentiment: The Role of Media in the Stock Market," *The Journal of Finance* (62:3), pp. 1139-1168.
- Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., and Kappas, A. 2010. "Sentiment Strength Detection in Short Informal Text," *Journal of the American Society for Information Science and Technology* (61:12), pp. 2544-2558.
- Thet, T. T., Na, J.-C., and Khoo, C. S. 2010. "Aspect-Based Sentiment Analysis of Movie Reviews on Discussion Boards," *Journal of Information Science* (36:6), pp. 823-848.
- Tsytsarau, M., and Palpanas, T. 2012. "Survey on Mining Subjective Data on the Web," *Data Mining and Knowledge Discovery* (24:3), pp. 478-514.
- Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. 2011. "Election Forecasts with Twitter: How 140 Characters Reflect the Political Landscape," *Social Science Computer Review* (29:4), pp. 402-418.
- Turney, P. D. 2002. "Thumbs up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews," *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics: Association for Computational Linguistics*, pp. 417-424.
- Twitter. 2013. "Celebrating #Twitter7." from <https://blog.twitter.com/2013/celebrating-twitter7>
- Velikovich, L., Blair-Goldensohn, S., Hannan, K., and McDonald, R. 2010. "The Viability of Web-Derived Polarity Lexicons," *Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Association for Computational Linguistics*, pp. 777-785.
- Wiebe, J., and Riloff, E. 2005. "Creating Subjective and Objective Sentence Classifiers from Unannotated Texts," in *Computational Linguistics and Intelligent Text Processing*. Springer, pp. 486-497.
- Wikipedia. 2013. "Social Media." from [http://en.wikipedia.org/wiki/Social\\_media](http://en.wikipedia.org/wiki/Social_media)

- Wiktionary. 2013. "Wiktionary." from <http://www.wiktionary.org/>
- Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., and Patwardhan, S. 2005a. "Opinionfinder: A System for Subjectivity Analysis," *Proceedings of HLT/EMNLP on Interactive Demonstrations*: Association for Computational Linguistics, pp. 34-35.
- Wilson, T., Wiebe, J., and Hoffmann, P. 2005b. "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis," *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*: Association for Computational Linguistics, pp. 347-354.
- Wilson, T., Wiebe, J., and Hoffmann, P. 2009. "Recognizing Contextual Polarity: An Exploration of Features for Phrase-Level Sentiment Analysis," *Computational Linguistics* (35:3), pp. 399-433.
- Xu, W., Li, T., Jiang, B., and Cheng, C. 2012. "Web Mining for Financial Market Prediction Based on Online Sentiments," *Proceedings of PACIS 2012*
- Yu, Y., Duan, W., and Cao, Q. 2013. "The Impact of Social and Conventional Media on Firm Equity Value: A Sentiment Analysis Approach," *Decision Support Systems* (55:4), pp. 919-926.
- Zhang, Z., Li, X., and Chen, Y. 2012. "Deciphering Word-of-Mouth in Social Media: Text-Based Metrics of Consumer Reviews," *ACM Transactions on Management Information Systems (TMIS)* (3:1), p. 5.

## CURRICULUM VITAE

### RESEARCH INTERESTS

---

- Business Intelligence & Analytics
- Social Media Analytics
- Data/Text/Web Mining
- Health Informatics
- Big Data
- Online Communities
- Natural Language Processing

### TEACHING INTERESTS

---

- Business Intelligence & Analytics
- Data/Text/Web Mining
- Data Management/Database Design
- Big data
- Programming
- Enterprise Resource Planning
- Business Statistics

### ACADEMIC APPOINTMENTS

---

**Assistant Professor of Information Systems (Tenure-Track), 2014 – present**  
Dakota State University (College of Business and Information Systems)

### EDUCATION

---

**Ph.D. Candidate, Information Technology Management, 2009 – 2014**

University of Wisconsin –Milwaukee (Sheldon B. Lubar School of Business)

- **Dissertation Title:** Three Essays on Opinion Mining of Social Media Texts
- **Dissertation Committee Chairs:** Atish Sinha, Huimin Zhao
- **Minor Area:** Computer Science

**Master of Business Administration, 2007 – 2008**

University of Illinois at Chicago

**Bachelor of Arts, English 2003 – 2007**

South China University of Technology

### REFEREED CONFERENCE PROCEEDINGS

---

Deng, S., Sinha, A., Zhao, H., and Wu, J. (2013). “Learning Domain-Specific Lexicon for Sentiment Analysis of Social Media Texts,” in *Proceedings of the 23rd Workshop on Information Technologies and Systems (WITS 2013)*, Milan, Italy. (**Best Paper Award Winner**)

Wu, J. and Deng, S. (2013). “Why do users prefer an experienced IT Provider? A Study of New Online Services Adoption,” in *Proceedings of the 43<sup>rd</sup> Decision Science Institute Annual Meeting (DSI 2013)*, Baltimore, MD.

Kwak, D.-H., Deng, S., Neely, D., and Zhao, H. (2012). “What Should be Considered to Attract Charity Website Visitors?” in *Proceedings of the 11<sup>th</sup> Workshop on e-Business (WEB 2012)*, Orlando, FL.

## INVITED TALKS AND PRESENTATIONS

---

Shuyuan Deng, Atish P. Sinha, Huimin Zhao. “Expanding Sentiment Lexicons for Sentiment Analysis of Domain-Specific Social Media Text,” Hefei University of Technology, China, 2014.

Shuyuan Deng, Atish P. Sinha, Huimin Zhao. “Adapting sentiment lexicons to domain-specific social media texts,” University of Cincinnati, 2014.

Shuyuan Deng, Atish P. Sinha, Huimin Zhao. “Adapting sentiment lexicons to domain-specific social media texts,” Oregon State University, 2014.

Shuyuan Deng, Atish P. Sinha, Huimin Zhao. “Adapting sentiment lexicons to domain-specific social media texts,” University of Utah, 2013.

## WORKING PAPERS

---

Deng, S., Sinha, A., and Zhao, H. “Predicting the stock market using real-time social media sentiment”

Deng, S., Sinha, A., and Zhao, H. “Learning syntactic cues for unsupervised sentiment classification”

Deng, S., Sinha, A., and Zhao, H. “Target-dependent opinion mining in social media”

Kwak, D.-H., Deng, S., Kim, K., Kumar, S., and Zhao, H. “Connection Does Matter: The Effect of Charity Website Network on Performance”

Deng, S., Jain, H., and Ross, A. “Monitoring Disruption Risks in Supply Chains: The Contribution of Web Text Mining”

Deng, S., Abbasi, A., and Zhao, H. “Detecting Management Fraud using Text Mining – Does Syntax Matter?”

## TEACHING EXPERIENCE

---

### Lecturer:

Fall 2014, Dakota State University

INFS 732: Emerging Technologies (online and classroom, Master elective)

INFS 830: Decision Support Systems (online and classroom, Doctoral elective)

Summer 2014, University of Wisconsin–Milwaukee

BusMgmt 735 (online): Advanced Spreadsheet Tools (MBA elective)

(Mean Instructor Rating: 4.8/5)

Spring 2014, University of Wisconsin–Milwaukee

BusMgmt 735 (online): Advanced Spreadsheet Tools (MBA elective)

(Mean Instructor Rating: 4.2/5)

Fall 2013, University of Wisconsin–Milwaukee

BusMgmt 735: Advanced Spreadsheet Tools (MBA elective)

(Mean Instructor Rating: 4.5/5)

Bus Adm 395: Special Topics in Business: Advanced Spreadsheet Tools

(Mean Instructor Rating: 4.5/5)

**Instructional Assistant:**

Spring 2013, University of Wisconsin –Milwaukee

Bus Adm 211: Business Scholars: Introduction to Management Statistics

**Teaching Assistant:**

Fall 2012, University of Wisconsin –Milwaukee

Bus Adm 231: Business Scholars: Introduction to Information Technology

Management

Fall 2009 – Spring 2011 & Fall 2012 – present, University of Wisconsin –Milwaukee

Bus Adm 230: Introduction to Information Technology Management

**Project Assistant**

Fall 2011 – Spring 2012, University of Wisconsin –Milwaukee

**INDUSTRY EXPERIENCE**

---

**Data Analyst Intern at Morningstar**

March 2009 – May 2009, Morning Star, Inc.

**IS TECHNICAL SKILLS**

---

**Programming Languages:** Java, Python, VB.NET, C#, JavaScript

**Operating Systems:** Microsoft Windows, Linux/Unix

**Database Systems:** Microsoft SQL Server, Oracle Database, MySQL, Microsoft Access

**Data Analytics Tools:** R, SAS, Weka, RapidMiner, Excel & VBA, Tableau, Minitab, Stata

**Big Data Technologies:** Hadoop, MongoDB, Cassandra, Hive, Spark, Neo4j

**SAP Applications:** Enterprise Portal, Business Warehouse, Lumira, HANA

**PROFESSIONAL ACTIVITIES**

---

Member of the Association for Information Systems (AIS)

Student Volunteer, Design Science Research in Information Systems and Technology, 2011

Reviewer for the following conferences:

- DESRIST (2011)
- ICIS (2013)
- Pre-ICIS SIGDSS (2013)
- PACIS (2014)
- AMCIS (2014)

Reviewer for the following journals:

- MIS Quarterly
- DATA BASE

## **PROFESSIONAL CERTIFICATIONS**

---

SAS Certified Advanced Programmer for SAS 9

Oracle Certified SQL Expert

EMC Certified Data Science Associate

Cloudera Certified Professional: Data Scientist (Exam passed)