Modeling public mood and emotion: Blog and news sentiment and socio-economic phenomena

Mu-Yen Chen, Ting-Hsuan Chen

PII:	S0167-739X(17)32375-0
DOI:	https://doi.org/10.1016/j.future.2017.10.028
Reference:	FUTURE 3767

To appear in: *Future Generation Computer Systems*

Received date :	30 December 2016
Revised date :	16 October 2017
Accepted date :	18 October 2017

Please cite this article as: M. Chen, T. Chen, Modeling public mood and emotion: Blog and news sentiment and socio-economic phenomena, *Future Generation Computer Systems* (2017), https://doi.org/10.1016/j.future.2017.10.028

Modeling Public Mood and Emotion: Blog and News Sentiment and Socio-Economic Phenomena

Mu-Yen Chen[1], Ting-Hsuan Chen[2,]*

mychen.academy@gmail.com, thchen@nutc.edu.tw

[1]Department of Information Management, National Taichung University of Science and Technology, Taichung 404, Taiwan, R.O.C

[2] Department of Finance, National Taichung University of Science and Technology, Taichung 404, Taiwan, R.O.C

Correspondence: Ting-Hsuan Chen, Associate Professor

Department of Finance,

National Taichung University of Science and Technology

Postal address: 129 Sec. 3, San-Min Road, Taichung 404, Taiwan, R.O.C.

E-mail address: thchen@nutc.edu.tw

Tel: +886-4-22196863

Fax: +886-4-22196181

Modeling Public Mood and Emotion: Blog and News Sentiment and Socio-Economic Phenomena

Abstract

The development of online virtual communities has raised the importance in analyzing massive volume of text from websites and social networks. This research analyzed financial blogs and online news articles to develop a public mood dynamic prediction model for stock markets, referencing the perspectives of behavioral finance and the characteristics of online financial communities. This research applies big data and opinion mining approaches to the investors' sentiment analysis in Taiwan. The proposed model was verified using experimental datasets from ChinaTimes.com, cnYES.com, Yahoo stock market news, and Google stock market news over an 18 month period. Empirical results indicate the big data analysis techniques to assess emotional content of commentary on current stock or financial issues can effectively forecast stock price movement.

Keywords: Sentiment mining analysis, text mining, opinion mining, stock price, public mood and emotion

1. Introduction

The rapid development of online communities and the mobile Internet have driven a rapid expansion in online news forums and discussions which potentially include data useful for investment decision making. Various approaches have been developed for analyzing "Big Data", referred to as text, web and sentiment mining. Sentiment Mining is often referred to as opinion mining, sentiment analysis or subjectivity analysis. It is a form of textual analysis which automatically extracts words and sentences which appear with higher frequency and are potentially meaningful. “Sentiment” refers to contextualized attitudes, comments, and feelings, thus sentiment mining is designed to detect, extract, and analyze hidden sentiment or semantic orientation.

Sentiment mining has been mostly applied to textual analysis social network content. In 2015, eMarketer found that 89% of US companies use social media as a marketing tool [1]. In 2013, RBC Capital found that the return on investment for advertising on Facebook is only slightly lower than that for top-ranked Google, and is ahead of Twitter, LinkedIn, Yahoo, AOL and other platforms [2], reflecting the significant marketing impact of social media. In addition, after U.S. Securities and Exchange Commission (SEC) officially allowed listed companies to disclose their earnings on social networks in 2013, the world’s leading data providers including Townsend Reuters and Bloomberg Data began to provide data analysis services for social network services. In 2014, the worlds’ largest social data provider GNIP noted that sentiment analysis social networks first began in 2010 [3]. The initial purpose of such activities was to allow companies to assess customer reaction to and satisfaction with their products and services. However, sentiment analysis of social networks has significant potential in other domains, such as the prediction of stock price movements.

Bollen et al. (2011) analyzed a massive volume of Twitter content to determine the use of mass emotion in predicting future stock market trends [4]. Applying the OpinionFinder and Google-Profile of Mood States API tools, they scraped nearly ten million tweets, and subjected the results to Granger causality analysis and Self-Organizing Fuzzy Neural Network (SOFNN) to assess correlations between Twitter-based financially related emotional content and the Dow Jones Industrial Average Index (DJIA). Result showed that a change in “calm,” as determined by Google-Profile of Mood States (GPOMS), can predict the movement of the DJIA over the following 3 to 4 days with an accurate rate as high as 86.7%. In 2012, Datasift used 95,019 Tweets from 58,665 Twitter users to assess the emotional moment for correspondence to the stock price on the day of THE Facebook IPO. Findings showed that sentimental tendencies can be used as an effective predictor of stock price movements [5]. Moreover, in May 2012 Derwent Capital Markets launched the world's first hedge fund based on Twitter public sentiment, promising annual rates of return between 15 and 20% [6].

Vindhini et al. (2012) pointed out that sentiment mining is primarily a natural language processing technique used to analyze emotions and opinions in text [7]. It thus covers a broad range of research fields and application domains, such as article collection in social network, computational linguistics to analyze the grammatical structure of articles, and determination of whether the emotional polarity of vocabulary is positive or negative. In addition, analysis based on statistical techniques and artificial intelligence have been increasingly integrated into sentiment mining to improve outcomes. However, sentiment mining applications are still immature and are subject to certain challenges including: (1) Inconsistent article structures and lengths [8]. (2) Data processing and transmission bottlenecks for real-time online analysis, though the increasing popularity of cloud computing potentially provides a solution [9, 10]. (3) Lack of a consistent and universal development framework reduces the efficiency of software development and analysis, resulting in poor flexibility in algorithm development and maintenance difficulty [11]. This research proposes and a sentiment mining approach designed to address massive quantities of short text articles based on the Academia Sinica Bilingual Ontological Wordnet (BOW-WordNet) [12], the National Taiwan University Sentiment Dictionary (NTUSD) [13], and Python-Jieba to detect abbreviations and slang, delete stop words, and detect positive and negative connotations in words. A sentiment mining framework is then developed for use in subsequent research.

This research uses content from finance-related blogs and related news articles to integrate features from behavioral finance and social networks for sentiment mining, seeking to empirically determine whether public sentiment in online Taiwanese communities can be used to predict future stock market trends in Taiwan. The remainder of this paper is organized as follows. Section 2 surveys the literature related to sentiment mining, sentiment classification, feature selection, and feature weighting methods. Section 3 describes the research methodology. Section 4 discusses the experimental design, dataset collection, experimental results, and model performance. Finally, the research contribution and future works are presented in Section 5.

2. Literature Review

2.1. Sentiment Mining Analysis

The Internet has emerged as a global medium of communication, particularly through social network applications. Social networks offer private individuals an easy and convenient way to widely disseminate their opinions, and an array of techniques for social behavior analysis and prediction have been developed and applied to analyze such content, which are important sources for sentiment mining analysis. Current technologies allow for the effective filtering and analysis of massive amounts of text. Sentiment analysis includes the identification of opinion holders, feature words, and opinion words. These three main methods are introduced below.

Kim et al. (2004) used the BNN2 named entity tagger-IdentiFinder to identify potential opinion holders [14], but was restricted to individual people and institutions. When more than one holder occurs in a sentence, the one with the closest to the targeted opinion is selected. Ku et al. (2007) noted that a proper noun or a synonym appearing in front of the verb which expresses an opinion usually identifies the opinion holder [15].

In terms of feature word identification, Hu et al. (2004) proposed an Association Rule mining method to identify the most common nouns or noun phrases in given sentences [16]. They found that customer comments typically include considerable description which is not directly related to the product itself. However, product-related comments tend to use the same vocabulary, thus nouns or noun phrases which appear with higher frequency in the comments are thus more likely to be related to the product. Su et al. (2008) proposed a method using the association approach to effectively identify implicit product features [17]. Clustering is conducted based on the intrarelationship of feature vocabulary and opinion words. A sentiment association set is then established to describe the association between the feature and opinion sets. Using the previously established association set, feature vocabularies not explicitly appearing in the comments can be identified to provide more accurate view of comments.

In terms of opinion word identification, Ku et al. (2007) proposed using the frequency of a Chinese character appearing in a dictionary to determine whether it is an opinion word or not [15]. The meaning of a Chinese opinion word can be seen as a function of character combinations because, when encountering an unknown character, readers will attempt to interpret it according to its ideographic content. They then calculate the probability of each character in a certain word appearing in a dictionary file. Finally, whether or not the sum of the scores for the word exceeds the threshold value is used to determine if the word is an opinion. Hu et al. (2004) proposed a simple yet effective way of using synonym and antonym sets of adjectives in WordNet to predict the semantic direction of adjectives [16]. Su et al. (2008) proposed an opinion mining approach to evaluate the consumers’ comments of buying the new car [17]. In general, adjectives and their synonyms have similar meanings while antonyms have the opposite meaning. Hence, when the synonyms or antonyms of an adjective is known, we can use above concept to predict the meaning of said adjective. Qiu et al. (2009) proposed to use double propagation to expand domain sentiment lexicon in certain field [18]. The main concept is that opinion words in a comment are almost always accompanied by feature vocabulary. Thus we can use known opinion words to identify feature vocabulary items. New opinion words and feature vocabulary items can then again be used to identify new feature vocabulary items and opinion words [19]. Such double propagation of words is repeated until no new opinion words or feature vocabulary items are identified.

2.2. Sentiment Classification

Sentiment classification focuses on four aspects: subjectivity classification, word sentiment classification, document sentiment classification, and opinion extraction [20]. This paper mainly focuses on the sentiment classification of document content to identify positive or negative opinions.

In recent years, many studies have adopted feature selection and machine learning methods for sentiment classification (see Tables 1 and 2). For example: Pang et al. applied simple document frequency (DF) as the threshold of filtering features to movie reviews to assess the effectiveness of machine learning methods such as SVM, Naïve Bayes (NB) and Maximum entropy in processing sentiment classification [21]. They found that SVM has the best classification performance. Na et al. (2005) introduced negation phrases into feature selection to further improve the effectiveness of sentiment classification [22]. Abbasi (2007) combined Entropy and Genetic Algorithm (GA) to propose the EWGA (Entropy weighted genetic algorithm) feature selection method for sentiment classification of articles on internet forums and movie reviews [23]. Wang et al. (2011) proposed the FLDA (Fisher linear discriminant analysis) method based on DF to combine with SVM classifier for sentiment classification of online product reviews [24]. Fersini et al. (2014) proposed a novel ensemble learning methodology to effectively solve the polarity classification issue [32]. Abdel Fattah (2015) [33] and Wu et al. (2016) [34] also proposed novel approaches to handle and classify complex sentiment knowledge from unstructured internet review comments or microblog articles. Recently, Gui et al. (2017) built a heterogeneous network and polarity lexicon to connect users, products, words in related product reviews from IMDB and Yelp, thus significantly improving classification performance [35].

Table 1. Sentiment Classification Research

Type of Research	Methodology comparison	R1
	New method	R2
Machining Learning Method	Support vector machine (SVM)	ML1
	Naive Bayes (NB)	ML2
	Others (N-gram, neural network, deep learning)	ML3
Feature Selection Method	Document Frequency (DF)	FS1
	Information gain	FS2
	POS-Labeling algorithm	FS3
	Others (CHI, CPD…etc)	FS4
Feature Weighting Method	TF	FW1
	TF-IDF	FW2
	TP	FW3
Experimental Dataset	Internet reviews (product, movie, lodging…etc)	D1
	Internet forums or blogs	D2
	Other (news, open database…etc)	D3

Table 2. Literatures Review for Sentiment Classification

Scholars	Research Type		Machining Learning			Feature Selection				Feature Weighting			Experiment Dataset		
	R1	R2	ML1	ML2	ML3	FS1	FS2	FS3	FS4	FW1	FW2	FW3	D1	D2	D3
Pang et al. (2002) [21]	✓		✓	✓	✓	✓		✓		✓		✓	✓		
Na et al. (2005) [22]		✓	✓			✓		✓		✓	✓	✓	✓		
Whitelaw et al. (2005) [25]		✓	✓					✓		✓			✓		
Abbasi et al. (2007) [23]		✓	✓				✓					✓	✓	✓	
Li et al. (2007) [26]		✓	✓			✓		✓				✓	✓		
Tan and Zhang (2008) [27]	✓		✓	✓	✓	✓	✓		✓		✓		✓		
Zhang et al. (2008) [28]		✓	✓	✓	✓			✓			✓		✓		
Chen and Chiu (2009) [29]		✓			✓			✓			✓		✓		
O'keefe and Koprinska (2009) [30]		✓	✓	✓		✓			✓	✓	✓	✓	✓		
Bollen et al. (2011) [31]		✓			✓	✓				✓	✓	✓		✓	
Fersini et al. (2014) [32]		✓	✓	✓	✓			✓		✓	✓	✓	✓		
Abdel Fattah (2015) [33]		✓	✓		✓	✓	✓		✓	✓	✓	✓	✓		
Wu et al. (2016) [34]	✓		✓	✓	✓	✓				✓				✓	
Gui et al. (2017) [35]		✓	✓		✓	✓				✓			✓		

2.3. Feature selection

Feature selection selects the most important features that can present the meaning of the document [36], reducing the dimension of feature space, required computing time and costs, and noise while improving classification performance [26, 37]. Subjective online text such discussion threads and blogs are often unstructured or semi-structured, and feature selection is a key issue for the classification of such data [24].

Categorical Proportional Difference (CPD) is a feature selection method proposed by Simeon and Hilderman (2008) and is often applied to multi-category document classification [38]. O'Keefe and Koprinska (2009) used this method in two categories of semantic classification studies [30]. This method is mainly used to calculate the difference between a feature's positive and negative document frequency, allowing for the selection of features that can effectively distinguish between the two categories. The CPD method is calculated as follows:

$d_{P,i}(i = 1, 2,\ldots, m)$ and $d_{N,j}(j = 1, 2,\ldots, n)$ respectively represent the i-th positive category document and the j-th negative category document. *m* and *n* respectively represent the number of positive and negative category files. *t* represents the number of terms in the document. Random variables $d_{P,i}(t)$ and $d_{N,j}(t)$ are defined as follows:

$$d_{P,i}(t) = \begin{cases} 1 & if\ t \text{ occurs in } d_{P,i} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

$$d_{N,j}(t) = \begin{cases} 1 & if\ t \text{ occurs in } d_{N,j} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$$CPD = \frac{\left|\sum_{i=1}^{m} d_{P,i}(t) - \sum_{j=1}^{n} d_{N,j}(t)\right|}{\sum_{i=1}^{m} d_{P,i}(t) + \sum_{j=1}^{n} d_{N,j}(t)} = \frac{|m_1 - m_2|}{m_1 + m_2} \tag{3}$$

In Eq. 3, m_1 and m_2 respectively represent the positive and negative document frequency of a feature in all documents. CPD calculation produces a CPD value (importance) of a feature between 0 and 1 where a higher value indicates greater importance. Two types of documents can be effectively distinguished, thus a CPD feature value of 1 indicates that this feature only appears in a single category.

2.4. Feature Weight Method

When classifying document data, each document will be represented as a feature vector [28], that is, the construction of the term-document matrix (TDM). In the TDM, each feature vector represents the weight of a term in a document, with the last column indicating the category of each document. In the information retrieval field, feature weights are used to represent the usefulness of a feature in the retrieval process [39].

There are many ways of calculating feature weight in the document classification field. For example:

- Term frequency (TF)

  TF uses the frequency with which a term appears in a document to represent its weight. It is primarily used to measure a term’s importance in a document, and can be called a "Local Term Weight" [40].

- Inverse document frequency (IDF)

  IDF calculates weight according to a term’s document frequency (DF). It is primarily used to measure a term’s importance to all documents, and can be called a "Global Term Weight" [40]. The IDF weight of a term t can be expressed as follows:

$$IDF = log\frac{N}{m_t} \tag{4}$$

In Eq. 2.9, N represents the total number of all documents and m_t represents the total number of documents containing feature t.

- Term frequency-inverse document frequency (TF-IDF)

  TF-IDF combines TF and IDF weights, and is a popular weighting method in the field of information retrieval [39, 40, 41]. TF-IDF may be calculated as follows:

$$TF - IDF = TF \times IDF \tag{5}$$

- Term presence (TP)

  Pang et al. (2002) first used TP for the semantic classification of two categories [21]. TP judges weight by determining whether a term appears in a document, where 1 represents yes, and 0 represents no.

In studies related to semantic classification, TF weight, TP weight and TF-IDF weight are frequently used feature weighting methods. Studies have also confirmed that TP weights and TF-IDF weight generally have better classification performance, therefore TF-IDF weighting is used to construct term matrix in the present study.

3. Research Methodology

The research framework is described as below and illustrated in Fig. 1.

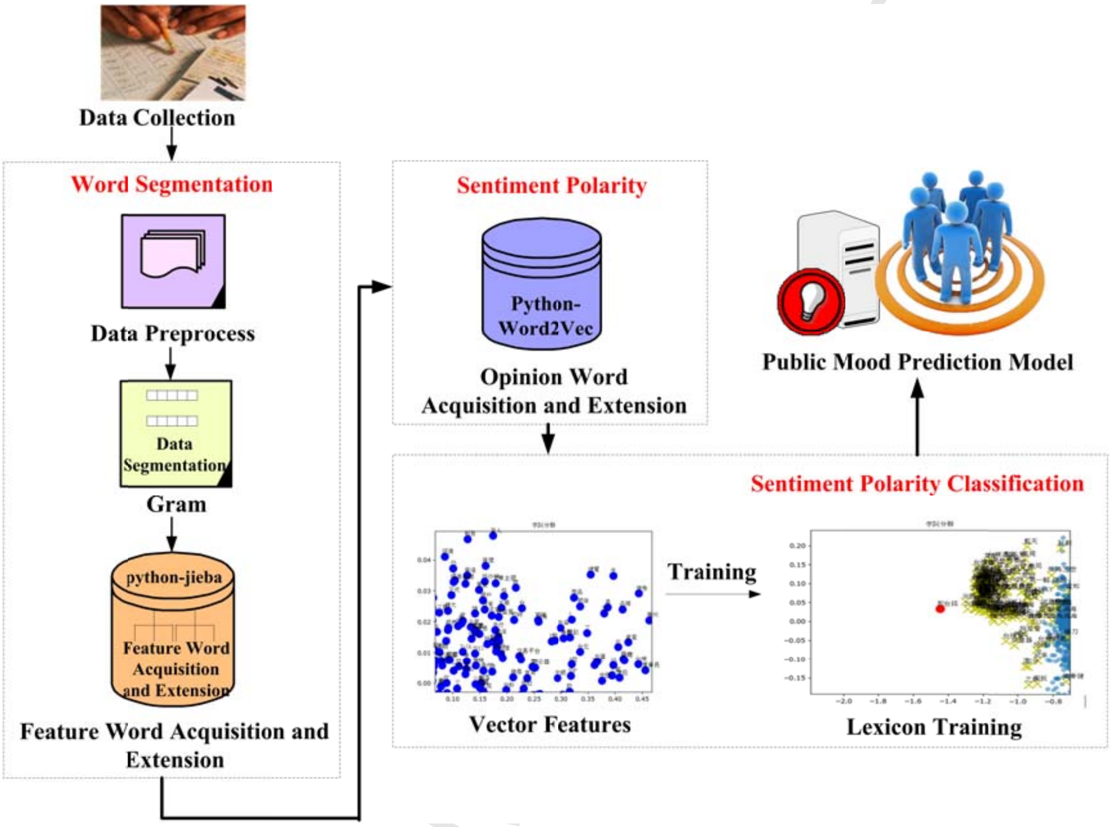
Data Collection
Word Segmentation
Data Preprocess
Data Segmentation
Gram
python-jieba
Feature Word Acquisition and Extension
Feature Word Acquisition and Extension
Sentiment Polarity
Python-Word2Vec
Opinion Word Acquisition and Extension
Public Mood Prediction Model
Sentiment Polarity Classification
Training
Vector Features
Lexicon Training

Fig.1. Research Framework

(1) Data Sources and Collection

This research processed financial data collected from Taiwan's most popular news websites including ChinaTimes.com, cnYES.com, Yahoo, and Google from January 1, 2016 to July 31, 2017. Stock prices of the individual stock and the TWSE capitalization weighted stock index (TAIEX) for the period were also recorded.

(2) Data Preprocessing

Raw data was collected from mainstream financial news articles, and the data preprocessing tasks are including incomplete data correction, duplication data remove, or missing values deletion. HTML tags and incomplete data were removed, along with unnecessary information to improve training efficiency.

(3) Gram

This step prepares the raw data for later experimental analysis. The “Python-Jieba” Chinese text segmentation toolkit is used to break the text into words as a feature unit, and remove low frequency vocabulary. The Python-Jieba was used to handle the articles, it is convenient for feature word expansion and processing speed, and is more widely used. This step also used a custom dictionary to assist Python-Jieba in word segmentation, after which useful key words are extracted with their corresponding polarities.

(4) Feature Word Acquisition and Extension

This step uses the feature word identification to analyze the emotional moment for online financial news articles. The target feature words are manually acquired and expanded using Python-Jieba. We used Python-Jieba to extend the collected feature words. This can be accomplished through two approaches. The first segments words through the full article without respect to part of speech or polarity. The other captures nouns because it is easier to distinguish between sentences based on polarity.

(5) Opinion Word Acquisition and Extension

Opinion words are acquired and extended using Python-Word2Vec, a popular sentiment dictionary that can be used to calculate the relationship between words in terms of similarity and weight, allowing for words to be transformed in advance to multidimensional vectors which are then used to represent the word, as shown in Fig.2.

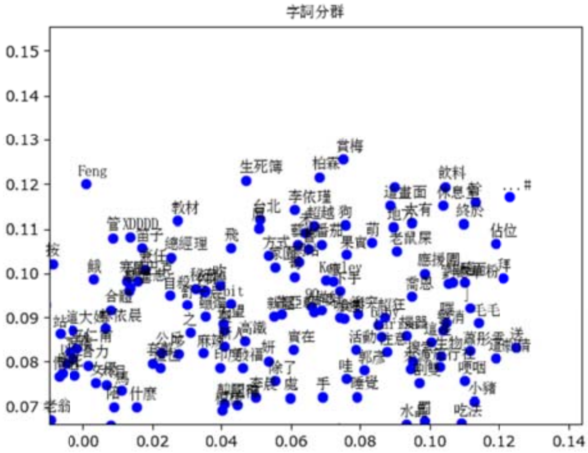

Fig.2. Two-dimensional spatial representation of vector features

(6) Lexicon Training

First, we use PCA for dimensionality reduction of word characteristics. We then transform the word into a vector and throw it into the spatial distribution, capturing 200 similar words around the specific keyword. Finally, these 200 words are used for training to determine the relevant words for upward and downward trends for stock prices. In addition, these 200 words were inserted into the lexicon according to the number of occurrences and files calculated, as shown in Fig.3.

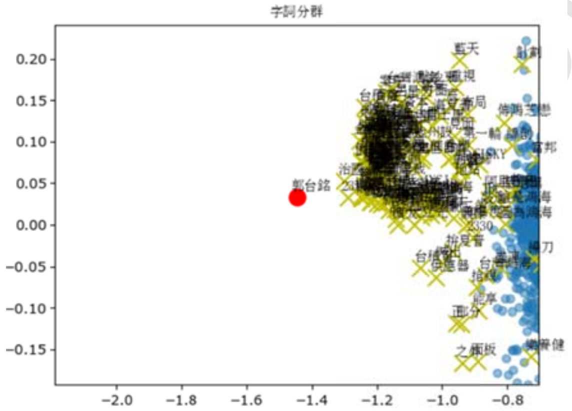

Fig.3. Similar words around the keyword

(7) Sentiment Polarity Modeling and Experimental Results

This research applies sentiment mining approaches to real-time analysis of the emotional moment of online financial news articles. In this step, sentiment polarity modeling followed raw data processing, feature word acquisition, and opinion word acquisition.

4. Experimental Results

4.1. Experimental Design and Data Collection

The proposed model was verified using experimental datasets from ChinaTimes.com, cnYES.com, Yahoo stock market news, and Google stock market news from January 1, 2016 to July 31, 2017.

Based on the lexicon training results, this research was calculated the positive or negative emotion of each article from the time period *t-5* to *t-1* to determine the correlation of stock price movement on time *t*, Experiments were designed to address the following research questions. (1) Does the amount of data used for prediction affect forecast accuracy? (2) Do larger lexicons improve accuracy? (3) Does the training result affect forecast accuracy?

4.2. Single Stock Price Prediction

To assess prediction performance, this experiment used only two keywords: "Hon Hai Precision Industry Company" and "Terry Gou" (Hon Hai's board chairman), for comparison against Hon Hai's stock performance over the same time period.

As shown in Table 3, lexicon polarity does not cover the amount of forecasting data, predicting too many unknown words in the article, leading to reduce the precision. In addition, increased lexicon polarity does not necessarily produce better results: excessive duplication in the training data can result in of the wrong direction of the positive and negative emotion, then resulting in inconsistent predictive accuracy instability.

Table 3. Results comparison for Single Stock Price Prediction

Scenario	Training Dataset	Training Period	Testing Dataset	Accuracy
1	2016/01/01~2016/12/31 10 articles per day	1 Year	2017/01/01~2017/07/31 10 articles per day	59.71%
2	2016/01/01~2016/12/31 10 articles per day	1 Year	2017/01/01~2017/07/31 50 articles per day	52.51%
3	2016/01/01~2017/07/31 50 articles per day	1.5 Year	2017/01/01~2017/07/31 10 articles per day	58.27%
4	2016/01/01~2017/07/31 50 articles per day	1.5 Year	2017/01/01~2017/07/31 50 articles per day	53.95%

To improve the accuracy rate, we then used the TF-IDF weighting method to investigate actual performance. We selected Scenario 1 to combine the TF-IDF weighting method because it has the highest accuracy results among the four scenarios.

Table 4. Results comparison with/without TF-IDF weighting method

Scenario	Training Dataset	Testing Dataset	Accuracy	Average Lexicon Coverage
1	2016/01/01~2016/12/31 10 articles per day	2017/01/01~2017/07/31 10 articles per day	59.71%	43.17%
5	2016/01/01~2016/12/31 10 articles per day	2017/01/01~2017/07/31 10 articles per day	53.43%	21.89%

Table 4 shows that Scenario 5 is less accurate than Scenario 1. We then investigated the TF-IDF weighting method effects on lexicon in the same article content and found that lexicon terms used in Scenario 1 are not realistic and inflate the importance of unimportant words. The lexicon used in Scenario 5 is closer to the character meaning and the words used are all important for the calculation prior to forecasting.

Using the same article content, we can see the average lexicon coverage rate in Scenario 1 is 43.17% and thus higher than in Scenario 5. However, the lexicon used in Scenario 1 is not measured by the TF-IDF weighting method, and contains many unimportant words. In contrast, Scenario 5 has a low average lexicon coverage; despite the use of the TF-IDF weighting method, the accuracy is not effectively improved.

4.3. Taiwan 50 and TAIEX Price Prediction

As shown in Table 4, in order to improve the lexicon average coverage rate of less than 30%, resulting in the accuracy cannot effectively enhance the problem, we collected financial articles on Taiwan's largest 50 publicly traded companies ("Taiwan 50") to establish lexicon training materials for enhancing lexicon coverage and forecasting movement of the TWSE Capitalization Weighted Stock Index (TAIEX). The experiment sought to determine whether forecast accuracy varied with lexicon and dataset size. Scenario 1 and 2 are the previous experiment in Section 4.2, to predict price movements of individual stocks, whereas Scenario 3 uses TF-IDF to adjust the lexicon, and Scenario 4 uses the Taiwan 50 to establish the lexicon and forecast fluctuations in the TAIEX.

Table 5. Results for individual stock and TAIEX price prediction

Scenario	Training Dataset	Training Period	Testing Dataset	Prediction	TF-IDF	Accuracy
1	2016/01/01~2016/12/31 10 articles per day	1 Year	2017/01/01~2017/07/31 10 articles per day	Single Stock	No	59.71%
2	2016/01/01~2017/07/31 50 articles per day	1.5 Year	2017/01/01~2017/07/31 10 articles per day	Single Stock	No	58.27%
3	2016/01/01~2016/12/31 10 articles per day	1 Year	2017/01/01~2017/07/31 10 articles per day	Single Stock	Yes	53.43%
4	2016/01/01~2016/12/31 10 articles per day	1 Year	2017/01/01~2017/07/31 10 articles per day	TAIEX	Yes	57.62%

To determine the impact of lexicon size on accuracy, we compared Scenarios 3 and 4, with results summarized in Table 5. In the testing dataset, the results showed accuracy increased with training lexicon size, but the new experimental result of Scenario 4 is 57.62% and not higher than the previous experimental result of Scenario 1 with 59.71% accuracy, but the meaning of the sentence is closer to the real meaning with higher average lexicon coverage, as shown in Table 6.

Table 6. Average lexicon coverage for individual stock and TAIEX price prediction

Scenario	Training Dataset	Testing Dataset	Average Forecasting Article Words	Average Lexicon Coverage
3	2016/01/01~2016/12/31 10 articles per day	2017/01/01~2017/07/31 10 articles per day	258,017	21.89%
4	2016/01/01~2016/12/31 10 articles per day	2017/01/01~2017/07/31 10 articles per day	258,017	59.01%

Using the same texts, Scenario 4 provides significantly greater lexicon coverage than Scenario 3. Therefore, if increasing the lexicon data can improve the low rate of lexicon coverage, TF-IDF can be used to boost prediction accuracy, as shown in Table 6.

In Table 7, the only difference between Scenario 4 and Scenario 5 is that the training period, but accuracy is significantly increased in Scenario 5. This is because the *t-1* to *t-5* articles are predicted at the predicted time t; when t+1 is predicted, t has already become part of the training lexicon, thus the t+1 accuracy will generally increase with coverage. The results suggest that using TF-IDF weighting to adjust the training lexicon can improve forecasting accuracy and higher amount of lexicon with TF-IDF weighting method can also achieve high accuracy.

Table 7. Results comparison for different training periods

Scenario	Training Dataset	Training Period	Testing Dataset	Prediction	TF-IDF	Accuracy
4	2016/01/01~2016/12/31 10 articles per day	1 Year	2017/01/01~2017/07/31 10 articles per day	TAIEX	Yes	57.62%
5	2016/01/01~2017/07/31 50 articles per day	1.5 Year	2017/01/01~2017/07/31 10 articles per day	TAIEX	Yes	65.81%

5. Conclusion

This research integrated content from financial blogs and news articles to develop a public mood dynamic prediction model for stock prices, referencing behavioral finance and online financial community characteristics. A public mood time series prediction model is also presented, integrating features from social networks and behavioral finance, and uses big data analysis to assess emotional content of commentary on current stock or financial issues to forecast changes for Taiwan stock index. The proposed model was verified using experimental datasets from the ChinaTimes.com, cnYES.com, Yahoo stock market news and Google stock market news from January 1, 2016 to July 31, 2017.

This research is subject to several limitations. First, experiments were conducted only using stock prices from the TAIEX, and further validation could be provided by duplicating the experiments with other stock markets, such as the DJIA, S&P 500, or IBOVESPA. Second, this research discusses the financial behavior in social networks, and future research could adopt other popular social network platforms, such as Twitter, Instagram, or Snapchat. Finally, the proposed model can also consider other factors as input variables, including personality traits, trust, and other psychological indicators.

ACKNOWLEDGMENT

The authors wish to thank the Ministry of Science and Technology (National Scientific Council) of the Republic of China for financially supporting this research under Contract Grants No. MOST106-2634-F-025-001, and MOST105-2410-H-025-015-MY2.

References

[1] eMarketer (2013), "The Year of Social? Nearly nine in 10 marketers will use social media marketing next year," http://www.emarketer.com/Article/Year-of-Social/1010386

[2] RBC Capital Markets and Advertising Age (2013). "Facebook, Inc." September 12, 2013

[3] GNIP Whitepaper (2014), "Social Media in Financial Markets: The Coming of Age," GNIP. https://gnip.com/pages/social-media-and-the-markets-the-coming-of-age-whitepaper/

[4] J. Bollen, H. Mao, A. Pepe, Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena. Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (2011) 450-453.

[5] Datasift (2012). Facebook IPO - stock market and social media data. http://s3.amazonaws.com/ DataSiftReports/2012-05-18_Facebook_IPO-Market_and_social_media_data/index.html

[6] Dylan Tweney (2012), "Twitter-fueled hedge fund bit the dust, but it actually worked," http://venturebeat.com/2012/05/28/twitter-fueled-hedge-fund-bit-the-dust-but-it-actually-worked/

[7] G. Vinodhini, R.M. Chandrasekaran, Sentiment Analysis and Opinion Mining: A Survey. International Journal of Advanced Research in Computer Science and Software Engineering 2(6) (2012) 282-292.

[8] K. Guo, L. Shi, W. Ye, X. Li, A survey of Internet public opinion mining. 2014 IEEE International Conference on Progress in Informatics and Computing (PIC), 2014.

[9] F. Zheng, Y. Xu, Y. Li, Research on Internet Hot Topic Detection Based on MapReduce Architecture. 4th International Conference on Intelligent Human-Machine Systems and Cybernetics, 2012.

[10] L. Jiang, B. Ge, W. Xiao, M. Gao, BBS Opinion Leader Mining Based o A Improved PageRank Algorithm Using MapReduce. IEEE Conference on Chinese Automation Congress, 2013.

[11] M.Z. Khan, M.N.A. Khan, Enhancing Software Reusability through Value Based Software Repository. International Journal of Software Engineering and Its Applications 8(11) (2014) 75-88.

[12] BOW-WordNet Website. http://bow.ling.sinica.edu.tw/wn/

[13] National Taiwan University Sentiment Dictionary (NTUSD) Website. http://nlg18.csie.ntu.edu.tw:8080/opinion/index.html

[14] S.M. Kim, E. Hovy, Determining the Sentiment of Opinions. Proceedings of the 20th international conference on Computational Linguistics, Geneva, Switzerland, 2004.

[15] L.W. Ku, Y.S. Lo, H.H. Chen, Using Polarity Scores of Words for Sentence-level Opinion Extraction. Proceedings of NTCIR-6 Workshop Meeting, May 15-18, Tokyo, Japan, 2007.

[16] M. Hu, B. Liu, Mining and Summarizing Customer Reviews. KDD'04, August 22–25, Seattle, Washington, USA, 2004.

[17] Q. Su, X. Xu, H. Guo, Z. Guo, X. Wu, X. Zhang, B. Swen, Z. Su, Hidden Sentiment Association in Chinese Web Opinion Mining. WWW 2008, April 21–25, 2008, Beijing, China. ACM 978-1-60558-085-2/08/04, 2008.

[18] G. Qiu, B. Liu, J. Bu, C. Chen, Expanding Domain Sentiment Lexicon through

Double Propagation. Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence (IJCAI-09), 2009.

[19] X. Ding, B. Liu, P.S. Yu, A Holistic Lexicon-Based Approach to Opinion Mining. WSDM'08, February 11-12, Palo Alto, California, USA, 2008.

[20] H. Tang, S. Tan, X. Cheng, A survey on sentiment detection of reviews. Expert Systems with Applications 36 (2009) 10760-10773.

[21] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up?: Sentiment Classification Using Machine Learning Techniques. Annual Meeting of the ACL Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, 10 (2002) 79-86.

[22] J.C. Na, C. Khoo, P.H.J. Wu, Use of negation phrases in automatic sentiment classification of product reviews. Library Collections, Acquisitions & Technical Services 29 (2005) 180-191.

[23] A. Abbasi, H. Chen, A. Salem, Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums. ACM Transactions on Information Systems, 26(3) (2007) 1-34.

[24] S. Wang, D. Li, X. Song, Y. Wei, H. Li, A Feature Selection Method Based on Improved Fisher's Discriminant Ratio for Text Sentiment Classification. Expert Systems with Applications, 38(7), (2011) 8696-8702.

[25] C. Whitelaw, N. Garg, S. Argamon, Using appraisal groups for sentiment analysis. Proc. of the 14th ACM international conf. on Information and knowledge management (2005) 625-631.

[26] S. Li, C. Zong, X. Wang, Sentiment Classification through Combining Classifiers with Multiple Feature Sets. Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering (2007) 135-140.

[27] S. Tan, J. Zhang, An Empirical Study of Sentiment Analysis for Chinese Documents. Expert Systems with Applications 34(4) (2008) 2622-2629.

[28] C. Zhang, W. Zuo, T. Peng, F. He, Sentiment Classification for Chinese Reviews Using Machine Learning Methods Based on String Kernel. Proceedings of the Third International Conference on Convergence and Hybrid Information Technology, 2 (2008) 909-914.

[29] L.S. Chen, H.J. Chiu, Developing a Neural Network based Index for Sentiment Classification. Proceedings of the International MultiConference of Engineers and Computer Scientists (2009) 744-749.

[30] T. O'Keefe, I. Koprinska, Feature Selection and Weighting Methods in Sentiment Analysis. Proceedings of the 14th Australasian Document Computing Symposium, 2009.

[31] J. Bollen, H. Mao, X. Zeng, Twitter mood predicts the stock market. Journal of Computational Science, 2, 1-8.

[32] E. Fersini, E. Messina, F.A. Pozzi, Sentiment analysis: Bayesian Ensemble Learning. Decision Support Systems, 68 (2014) 26-38.

[33] M. Abdel Fattah, New term weighting schemes with combination of multiple classifiers. Neurocomputing,167 (2015) 434-442.

[34] F. Wu, Y. Song, Y. Huang, Microblog sentiment classification with heterogeneous sentiment knowledge. Information Sciences 373 (2016) 149-164.

[35] L. Gui, Y. Zhou, R. Xu, Y. He, Q. Lu, Learning representations from heterogeneous network for sentiment classification of product reviews. Knowledge-Based Systems, 124 (2017) 34-45.

[36] T. Wang, H. Huang, S. Tian, J. Xu, Feature Selection for SVM via Optimization of Kernel Polarization with Gaussian ARD Kernels. Expert Systems with Applications, 37(9) (2010) 6663-6668.

[37] K. Polat, S. Gunes, A New Feature Selection Method on Classification of Medical Datasets: Kernel F-score Feature Selection. Expert Systems with Applications, 36(7) (2002), 10367-10373.

[38] M. Simeon, R. Hilderman, Categorical Proportional Difference: A Feature Selection Method for Text Categorization. Proceedings of the 17th Australasian Data Mining Conference, (2008) 201-208.

[39] A. Aizawa, An Information-theoretic Perspective of TF-IDF Measures. Information Processing and Management, 39(1) (2003) 45-65.

[40] X. Tian, W. Tong, An Improvement to TF: Term Distribution Based Term Weight Algorithm. Proceedings of the second International Conference on Networks Security Wireless Communications and Trusted Computing (NSWCTC) (2010) 252-255.

[41] A. Singhal, Modern Information Retrieval: A Brief Overview. IEEE Data Engineering Bulletin, 24(4) (2001) 35-43.

Mu-Yen Chen, Ph.D

Dr. Chen is a Professor of Department of Information Management at National Taichung University of Science and Technology, Taiwan. His current research interests include artificial intelligent, soft computing, bio-inspired computing, financial engineering, and data mining. Dr. Chen's research is published or is forthcoming in Information Sciences, Applied Soft Computing, Neurocomputing, Neural Computing and Applications, Journal of Educational Technology & Society, Journal of Information Science, The Electronic Library, Computers and Mathematics with Applications, Quantitative Finance, Expert Systems with Applications, Soft Computing, and a number of national and international conference proceedings.

Ting-Hsuan Chen, Ph.D

Dr. Chen is an Associate Professor of Department of Finance at National Taichung University of Science and Technology, Taiwan. Her current research interests include financial engineering, corporate social responsibility, spatial analysis and text mining. Dr. Chen's research is published or is forthcoming in Economic Modelling, Emerging Markets Finance and Trade, Review of Quantitative Finance and Accounting, International Review of Economics & Finance, Quarterly Review of Economics and Finance and a number of national and international conference proceedings.

- It develops an opinion mining framework based on TF-IDF in public blogs and news
- It builds a public mood dynamic prediction model in Taiwan stock market
- It uses big data technique to conduct sentiment analysis of emotions and reactions