Decision Support Systems 94 (2017) 65–76

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

Adapting sentiment lexicons to domain-specific social media texts

Shuyuan Deng, Atish P. Sinha, Huimin Zhao *

Sheldon B. Lubar School of Business, University of Wisconsin-Milwaukee, Milwaukee, WI 53201, United States

ARTICLE INFO

Article history:
Received 17 February 2016
Received in revised form 7 September 2016
Accepted 1 November 2016
Available online 3 November 2016

Keywords:
Sentiment analysis
Opinion mining
Sentiment lexicon
Lexicon expansion
Social media

ABSTRACT

Social media has become the largest data source of public opinion. The application of sentiment analysis to social media texts has great potential, but faces great challenges because of domain heterogeneity. 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-specific sentiment classification using an unannotated corpus and a dictionary. We evaluate our method using two large developing corpora, containing 743,069 tweets related to the stock market and one million tweets related to political topics, respectively, and five existing sentiment lexicons as seeds and baselines. The results demonstrate the usefulness of our method, showing significant improvement in sentiment classification performance.

1. Introduction

Social media has experienced exponential growth in the past few years. Posting messages on social media websites has become one of the most popular activities on the Internet today. The vast amount of user-generated content has made social media the largest data source of public opinion [1,2]. Such a data source is invaluable for business intelligence and analytics since opinion is a key predictor of human behavior [3,4]. Despite the tremendous effort in influencing customers through marketing campaigns on social media [5], extracting public opinion from social media is still in its infancy [1,6]. 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 [7]. Moreover, user-generated text data usually contain opinions and are more likely to influence other users than traditional media [8]. As a result, text data in social media have great potential to keep businesses informed in real-time, if appropriate techniques are employed [9].

Sentiment analysis, as a class of techniques to extract and assess opinions in texts, has been used to analyze text data in social media [10–12]. Sentiment analysis has been used to solve a variety of business problems. Aggarwal et al. [13] found that the sentiment of blogs significantly influences the financial performance of ventures. Mishne and Glance [14] analyzed the correlation between movie sales and the sentiment in blog postings related to the movies. They found that the sentiment is correlated more with the sales than with the volume of the relevant blogs. Bollen et al. [15] measured tweet sentiment in different dimensions and found a certain dimension of sentiment, namely, "calm", predicts the daily return of the Dow Jones Industrial Average index. Oh and Sheng [16] 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 [17]. In the past decade, many studies have applied sentiment analysis to online reviews and have obtained satisfactory results [11,18–20]. However, textual social media data raise new challenges for sentiment analysis.

Sentiment analysis commonly employs techniques from text mining and natural language processing [10]. Most research and applications implement the task as classifying the directional states of sentiment, e.g., positive, negative, or neutral [3,18]. 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 [21]. This approach has been shown to be more accurate than the lexicon approach when

* Corresponding author.
E-mail addresses: dengs@uwm.edu (S. Deng), sinha@uwm.edu (A.P. Sinha), hzhao@uwm.edu (H. Zhao).

http://dx.doi.org/10.1016/j.dss.2016.11.001
0167-9236/

the training and testing data are from the same domain [22]. 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 has raised a bigger challenge for analyzing social media texts because social media texts are generally short and heterogeneous in expression. They require a much larger training set than traditional media texts for building an effective supervised 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 [3]. The overall sentiment of the document is determined by the dominant polarity (i.e., positive or negative) among the indicators [23]. This approach relies on an accurate and extensive sentiment lexicon. The effectiveness of the existing lexicons, however, is quite limited when applied to new problems [24].

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 [3,25,26]. Consider 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 a general context, *red* would be interpreted 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 severe misunderstanding of information [27].

In this study, we refer to a context of discourse as a *content domain*. There has been some research to derive the contextual polarity of words using annotated data for sentiment classification [28]. However, given the unavailability of sufficient training data for social media, learning context-specific sentiment continues to be a major challenge [29]. 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 study 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., "cooooool") [30]. Without incorporating the latest language elements of social media, the power of sentiment analysis in this language domain would be quite limited.

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. The proposed approach contributes to sentiment analysis research and applications by mitigating the threat posed by the domain-specific challenges mentioned above. Applications such as analyzing customer opinion for specific products [18], assessing public sentiment toward political candidates [31], and using investor opinion to predict stock returns [32] will directly benefit from a domain-specific sentiment lexicon. Our work will also prompt researchers and practitioners to explore new opportunities for domain-specific sentiment analysis.

In the following section, we review the lexicon scoring approach to sentiment classification, as well as the lexicon generation methods. Next, we describe our approach for generating a domain-specific lexicon and report on our evaluation of our method with respect to existing lexicons. In the final section, we summarize our contributions and discuss future research directions.

2. Literature review

2.1. Sentiment classification using supervised learning

The supervised learning approach for sentiment classification uses machine learning classifiers to learn 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 [33,34]. Binary feature representation, indicating the presence or absence of words, has also been used for short texts. Assigning higher weights 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 [19]. More complex features have also been proposed, but they do not consistently outperform plain word features [35]. Bag-of-words features usually lead to a large feature set and are likely to cause overfitting [11]. Various feature selection methods have been proposed to alleviate the overfitting problem [11,36].

Supervised learning methods for sentiment analysis require a large training corpus so that the classifier can 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 a large volume of messages in just a few seconds. This suggests that the training data sets used in previous studies did not utilize big data and were not large enough to capture the characteristics of the language used on social media platforms, such as Twitter. Under such circumstances, the lexicon scoring approach provides an attractive alternative for sentiment analysis.

2.2. Sentiment classification using lexicons

A sentiment lexicon, also called opinion lexicon [3], is a collection of words or phrases that are commonly used to express feelings [25]. Some of the most widely used sentiment lexicons are General Inquirer (GI) [37], Multi-Perspective Question Answering Subjectivity Lexicon (MPQA) [38], SentiWordNet (SWN) [39], and Opinion Lexicon (OL) [19]. 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, most related studies did not use a continuous sentiment score since it adds another layer of complexity.

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 [19,23]. Some studies impose more weights on certain words. Other studies produce two sentiment scores for each document, namely positive and negative, by counting the number of words in each of the two categories [40,41]. For example, Das and Chen [42] 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 possible POS tags [43]. 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 [44]. It is also suitable for real-time sentiment classification, given its relatively lower computation requirement [22].

2.3. Sentiment lexicon generation methods

General sentiment lexicons are robust across different domains and applications since they capture the most commonly used sentiment words [3]. However, their performance usually suffers from two aspects, insufficiency and inaccuracy [30]. Insufficiency refers to the issue that general sentiment lexicons lack sentiment words that are specific to a domain. Inaccuracy means that the sentiment orientation of a word might change as the domain changes. Thus, domain-specific sentiment classification is likely to be improved if a domain-specific sentiment lexicon is used. In this section, we review methods that have been proposed to generate a sentiment lexicon.

The sentiment orientation of a word in a sentiment lexicon is usually assigned manually or using an automatic method. The two most popular sentiment lexicons, GI [37] and MPQA [43], 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 tend to get overlooked, thereby imposing a size limit on manually developed lexicons. In fact, the largest manually developed lexicon, MPQA, has only 7630 entries.

Automatic methods usually utilize one of two types of language resources, dictionary or corpus [3]. 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 those relationships. For instance, Liu's Opinion Lexicon [19] 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 the relationships 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 each other. Hatzivassiloglou and McKeown [45] 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 [45], Hatzivassiloglou and McKeown 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, for example, *respect* (unmarked) versus *disrespect* (marked).

Other corpus-based approaches utilize seed words to identify new sentiment words. A seed word is simply a word with known sentiment polarity. A seed lexicon is a collection of seed words. Turney [46] 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 [47] 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 [46] are "excellent" and "poor", which are believed to be the most frequently used in online reviews.

Wiebe & Riloff [48] 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. They then used several patterns to extract subjective and objective words. However, this method does not identify the polarity of subjective words.

Although prior studies have proposed various lexicon generation methods, most of them did not attempt to incorporate domain information in the new lexicon. As most sentiment analysis pertains to specific domains, it is important to close this gap. In recent years, domain-specific lexicon generation methods have been proposed. For instance, Oliveira et al. [49] demonstrated a novel method that creates a sentiment lexicon for microblog messages related to the stock market. Using a corpus of labeled Stocktwits messages, they effectively identified sentiment words based on the strength of association with the sentiment classes (bullish and bearish). Despite the usefulness of this method with labeled (i.e., annotated) data, it cannot be readily applied to unlabeled data. In this study, we propose a lexicon generation method that addresses the aforementioned domain challenges using unlabeled data. Compared to labeled data, unlabeled data can be inexpensively acquired. A method that works for unlabeled data has fewer restrictions on where it can be applied.

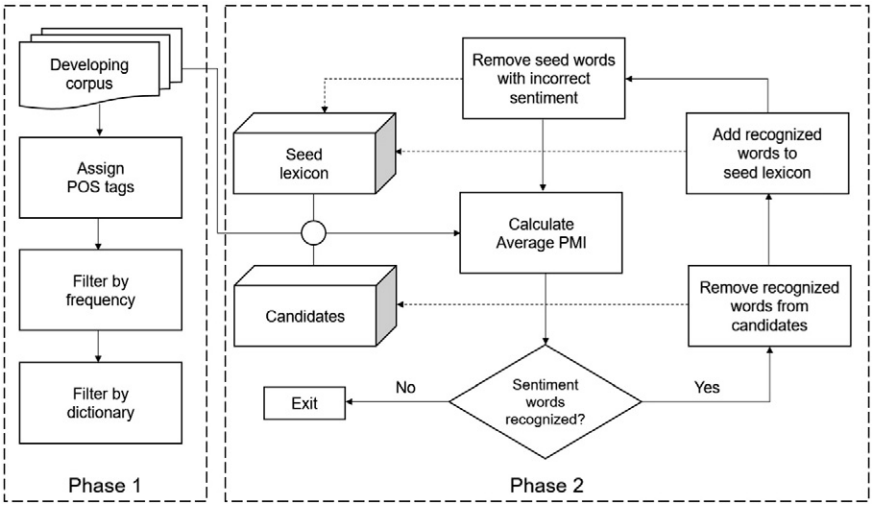

Fig. 1. The proposed lexicon generation method.

3. Proposed method

As discussed earlier, the manual approach for developing a 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 identity a limited number of sentiment words within a reasonable time frame (i.e., low recall). The dictionary-based approach is able to discover many more sentiment indicators if the dictionary contains sufficient synonyms and antonyms. However, the words identified are generally less accurate, without human supervision, than the manual approach. 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 precision with 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.

3.1. 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. As mentioned earlier, content domain refers to the context of discourse and language domain refers to the lexical and syntactical choices of language. Our method focuses on retaining such information in the new sentiment lexicon.

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. Wiktionary [50] is an appropriate choice because, compared to other dictionaries, Wiktionary is actively updated and contains more terms used in social media.

Fig. 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 the seed lexicon.
3) Determine the sentiment orientation of the candidates, add the recognized sentiment words to the seed lexicon, and remove the words with incorrect sentiment from the seed lexicon.
4) Repeat steps 2–3 until no more new words are added.

3.2. Extracting candidates

In the first phase of our method, we extract candidates for sentiment words from the developing corpus. Not all words in the developing

Let $C = \{c_1, c_2, ..., c_n\}$ denote the set of candidates. Let $n = \text{count}(C)$.
Let $PW = \{pw_1, pw_2, ..., pw_A\}$ and $NW = \{nw_1, nw_2, ..., nw_B\}$ denote the positive and negative seed word sets, respectively.
For each c_i in C,

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

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

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

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

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

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

Let $\text{SS}(c_i)$ denote the sentiment score of c_i, i.e.,

$$\text{SS}(c_i) = \text{APMI}(c_i, PW) - \text{APMI}(c_i, NW)$$

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

If $\text{SS}(c_i) > H$,

$\text{Polarity}(c_i) = positive$ and move c_i from C to PW;

If $c_i \in NW$, remove c_i from NW.

Else if $\text{SS}(c_i) < -H$,

$\text{Polarity}(c_i) = negative$ and move c_i from C to NW.

If $c_i \in PW$, remove c_i from PW.

If $\text{count}(C) < \text{n}$, repeat the procedure.

Fig. 2. Procedure for recognizing sentiment words from candidates.

corpus are equally likely to be sentiment words. The linguistic nature of words must be differentiated [51]. 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 [52]. Most previous work favors adjectives and adverbs, which are by nature more likely to carry sentiment [53]. Several extant studies on sentiment lexicon creation consider only adjectives and adverbs as sentiment clues [19,45,46,54]. Nouns and verbs are less frequently used since many of them do not carry any sentiment. Using these words will raise additional challenges in creating a sentiment lexicon. But to maximize the recall in identifying sentiment words, we consider nouns and verbs in our method. Note that both SWN and MPQA also include a fair number of nouns and verbs [39,43].

Intuitively, automatically learning nouns and verbs would 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, thereby reducing the precision of the lexicon. That is the reason why most of the previous studies have not used them. However, it would also add many domain-specific 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 four sets of candidates: the first set includes all four types of words (ARVN, representing adjective, adverb, verb, and noun, respectively), the second candidate set includes adjectives, adverbs, and nouns (ARN), the third candidate set includes adjectives, adverbs, and verbs (ARV), and the fourth candidate set includes adjective and adverbs (AR). In each candidate set containing nouns, proper nouns are excluded since they rarely bear any sentiment. In order to understand the impact of each individual POS category, in our experiments, we also evaluate our method using candidate sets with individual POS tags. Thus, the results for additional four candidate sets, A, R, V, and N, will also be presented.

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 program.

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 that 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.

3.3. Identifying polarity

The second phase in our method is to identify the sentiment polarity of the candidate words using a seed sentiment lexicon. This requires measuring the similarity between words. For this purpose, prior research has used the similarity between the fixed-window context vectors of two words in areas such as machine translation [55]. However, such measures tend to limit the similar words identified to words with similar POS properties. This is desirable in lexical-level translation since a verb should be translated to a verb and a noun should be translated to a noun, but such measures are not ideal for identifying sentiment words. Words with similar sentiment but different POS tags can occur in the same sentence, e.g., *soaring profit* and *tragic loss*. Therefore, we do not use context similarity in this study.

We use PMI [47] to measure the association between the candidate words and the seed words. The PMI between two words w_1 and w_2 is $\mathrm{PMI}(w_1, w_2) = log_2 \frac{\mathrm{p}(w_1, w_2)}{\mathrm{p}(w_1)\mathrm{p}(w_2)}$, where $\mathrm{p}(w_i)$ is the probability that word w_i occurs in a document and $\mathrm{p}(w_1, w_2)$ is the probability that w_1 and w_2 co-occur in a document. The sentiment polarity of each candidate c is determined by $\mathrm{Polarity}(c) = \frac{1}{N_p}\sum_{i=1}^{N_p} \mathrm{PMI}(c, \mathrm{p}w_i) - \frac{1}{N_n}\sum_{i=1}^{N_n} \mathrm{PMI}(c, nw_i)$, where $\{pw_i, i = 1,2, ..., N_p\}and\{nw_i, i = 1,2, ..., N_n\}$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 that with 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 candidate with a polarity score between $-H$ and H is not added to the seed lexicon. 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 a seed word is recalculated using the above procedure. By doing so, the algorithm corrects the polarity of the seed word using domain knowledge.

Fig. 2 presents our algorithm. The only parameter that needs to be tuned is the sentiment benchmark threshold, H. A high value of H would recognize too few sentiment words. A low value of H would add too much noise.

4. Evaluation procedure

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 [56]. As of early 2013, Twitter had over 200 million active users creating over 400 million tweets each day [57]. 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 [1,15,16,58], but the treatment of domain characteristics (both content and language domains simultaneously) for sentiment analysis in the extant literature has been minimal.

Table 1
Agreements and conflicts between existing lexicons.

Lexicon	Size	Agreements				Conflicts			
		MPQA	SWN	OL	GI	MPQA	SWN	OL	GI
MPQA	6457								
SWN	29,355	5402				1260			
OL	6789	5418	4782			50	1245		
GI	3642	2542	2563	2149		60	611	26	
LM	2703	951	1232	1043	508	26	292	9	15

Table 2
Positive and negative entries in existing lexicons.

Lexicon	Positive	Negative	Total
MPQA	2304	4153	6457
GI	1637	2005	3642
OL	2006	4783	6789
LM	354	2349	2703
Combined4	3649	7231	10,880
SWN	13,854	15,501	29,355

4.1. Seed lexicons and baselines

We used four general sentiment lexicons and one domain-specific lexicon as seed lexicons. We also used them as baselines in sentiment classification experiments. The effectiveness of the expanded lexicons can be shown if they outperform the seed sentiment lexicons in sentiment classification. The general lexicons are General Inquirer (GI), MPQA Subjectivity Lexicon (MPQA), Opinion Lexicon (OL), and SentiWordNet (SWN). The domain-specific lexicon we used is published by Loughran and McDonald [27]; it is coded as LM in this paper. LM contains positive and negative words picked by the authors after reviewing a large number of financial reports.

The GI lexicon [37] refers to the positive and negative word lists among its various dictionaries for content analysis. It contains 1637 (unique) positive words and 2005 (unique) negative words. These words were manually picked from psychological dictionaries by experts. It is frequently used by business researchers [42,59], as well as by computational linguistic researchers [28,60].

The MPQA sentiment lexicon was also compiled manually from multiple resources, including GI [38]. Each entry contains both the polarity

Table 3
Experiment results.

Lexicon	Positive words	Negative words	Precision (%)	Recall (%)	F-measure (%)
Combined4_ARVN	5935	16,543	80.27	80.28	80.31
Combined4_ARN	5482	13,340	76.96	76.97	76.88
Combined4_N	5163	11,926	76.39	76.39	76.28
Combined4_V	4130	8570	73.34	73.09	73.12
Combined4_ARV	4446	9453	74.10	73.87	73.89
Combined4	3649	7231	74.48	74.14	74.06
Combined4_AR	3965	7912	72.94	72.65	72.69
Combined4_A	3929	7791	72.85	72.56	72.60
Combined4_R	3684	7340	72.59	72.30	72.35
GI_ARVN	8119	8509	66.48	60.09	55.05
GI_ARN	6658	6914	65.79	59.12	53.51
GI_N	5656	5934	66.26	59.06	53.17
GI_ARV	4380	4932	64.74	57.73	51.23
GI	1637	2005	65.81	57.28	50.62
GI_V	3302	3918	64.67	57.41	50.60
GI_A	2709	3009	64.94	57.11	49.82
GI_AR	2890	3170	64.84	57.03	49.69
GI_R	2082	2442	65.00	56.88	49.29
OL_ARVN	5124	16,863	73.30	73.20	73.02
OL_ARN	4382	13,638	71.61	71.19	70.81
OL_N	3960	11,887	71.29	70.65	70.15
OL_ARV	3172	8557	71.26	70.08	69.30
OL	2006	4783	71.24	69.18	68.41
OL_V	2729	7069	70.69	69.20	68.23
OL_AR	2436	5788	70.28	68.48	67.33
OL_A	2372	5574	68.40	67.24	67.81
OL_R	2062	4957	70.06	68.16	66.94
LM_ARN	2874	12,000	64.30	64.12	63.81
LM_ARVN	3623	14,847	63.99	63.89	63.67
LM_N	2354	9848	64.05	63.69	63.21
LM_ARV	1662	7725	61.66	60.60	59.30
LM	354	2349	64.21	61.33	58.90
LM_V	1128	5422	61.75	60.27	58.54
LM_AR	857	4195	62.48	60.23	57.84
LM_A	792	3795	62.44	60.16	59.54
LM_R	416	2633	62.81	60.13	57.35
MPQA_ARVN	5497	16,081	74.86	74.85	74.74
MPQA_ARN	4759	11,774	73.09	72.78	72.47
MPQA_N	4312	9957	72.15	71.64	71.22
MPQA_ARV	3496	6762	69.60	68.25	67.30
MPQA	2304	4153	69.78	67.98	67.24
MPQA_V	3041	5687	68.93	67.35	66.24
MPQA_AR	2738	4869	68.75	67.18	66.06
MPQA_A	2687	4746	68.54	66.93	65.78
MPQA_R	2345	4248	68.47	66.85	65.69
SWN	13,854	15,501	60.17	58.99	57.76
SWN_N	15,018	16,348	59.07	58.02	56.35
SWN_ARN	15,155	16,504	59.07	58.02	56.35
SWN_ARVN	15,454	16,767	59.07	58.02	56.35
SWN_ARV	14,267	15,890	58.96	57.93	56.27
SWN_AR	13,990	15,641	58.96	57.93	56.27
SWN_V	14,120	15,742	58.96	57.93	56.27
SWN_A	13,979	15,624	58.96	57.93	56.27
SWN_R	13,865	15,517	58.96	57.93	56.27

Table 4
Sample sentiment words.

Positive	Negative
Reputations, nodding, legalization, $hower, pennants, drums, ceremony, quips, unblocked, fastest-rising, freshness, wellllll, risers, pumpjack, niceeee, ass-kicking, upsize, prevailing, healthfully, earner, top-pick, dopest	Subpar, godawful, ceded, soulless, piss, dreading, vacillating, slapped, sorrows, unfollows, cursing, devaluation, starving, farewell, trafficking, victimized, plonge, darkest, deaths, muting, dogshit, twilight, surmised, terrorists

Table 5
Class-level results.

Lexicon	Positive			Negative		
	Precision (%)	Recall (%)	*F*-measure (%)	Precision (%)	Recall (%)	*F*-measure (%)
Combined4_ARVN	79.64	79.36	79.50	80.92	81.19	81.05
Combined4_ARN	74.33	79.36	76.76	79.58	74.59	77.00
Combined4_N	73.55	79.18	76.26	79.22	73.60	76.30
Combined4_ARV	70.11	79.72	74.60	78.45	68.48	73.13
Combined4_V	69.25	79.36	73.96	77.86	67.33	72.21
Combined4	69.05	79.02	73.70	77.88	66.69	71.85
Combined4_AR	68.66	79.54	73.70	77.76	66.34	71.59
Combined4_A	68.56	79.54	73.64	77.71	66.17	71.48
Combined4_R	68.36	79.18	73.37	77.37	66.01	71.24
GI_ARVN	54.32	91.64	68.21	78.64	28.55	41.89
GI_ARN	53.55	92.70	67.88	77.94	26.24	39.26
GI_N	53.63	91.99	67.76	78.97	25.41	38.45
GI_ARV	52.68	92.53	67.14	76.80	22.94	35.32
GI_V	52.46	92.88	67.05	76.88	21.95	34.15
GI_A	52.23	93.59	67.05	77.64	20.63	32.59
GI_AR	52.18	93.59	67.01	77.50	20.46	32.38
GI_R	52.07	93.95	67.01	77.92	19.80	31.58
GI	49.32	91.54	64.10	76.64	18.76	30.14
OL_ARVN	69.70	77.76	73.51	76.89	68.65	72.54
OL_ARN	66.62	79.18	72.36	76.60	63.20	69.26
OL_N	65.65	80.25	72.22	76.92	61.06	68.08
OL_ARV	64.32	82.74	72.37	78.20	57.43	66.22
OL_V	63.21	83.45	71.93	78.17	54.95	64.53
OL_AR	62.32	84.06	71.61	78.24	52.81	63.05
OL_A	62.24	84.16	71.56	78.19	52.64	62.92
OL_R	61.96	84.34	71.44	78.16	51.98	62.44
OL	61.33	84.01	70.90	77.98	50.99	61.66
LM_ARN	60.70	70.64	65.30	67.90	57.59	62.32
LM_ARVN	60.82	69.04	64.67	67.17	58.75	62.68
LM_N	59.85	72.42	65.54	68.24	54.95	60.88
LM	56.47	72.02	63.30	67.23	47.65	55.77
LM_ARV	56.24	76.16	64.70	67.08	45.05	53.90
LM_V	55.68	78.47	65.14	67.82	42.08	51.93
LM_AR	55.29	81.85	66.00	69.64	38.61	49.68
LM_A	55.21	82.03	66.00	69.67	38.28	49.41
LM_R	55.05	83.45	66.34	70.57	36.80	48.37
MPQA_ARVN	71.99	77.76	74.76	77.72	71.95	74.72
MPQA_ARN	68.40	79.72	73.62	77.78	65.84	71.31
MPQA_N	66.81	80.25	72.92	77.48	63.04	69.52
MPQA_ARV	62.52	82.21	71.02	76.69	54.29	63.57
MPQA_V	61.54	82.56	70.52	76.33	52.15	61.96
MPQA_AR	61.41	82.38	70.36	76.09	51.98	61.76
MPQA_A	61.16	82.38	70.20	75.91	51.49	61.36
MPQA_R	61.08	82.38	70.15	75.85	51.32	61.22
MPQA	60.86	81.01	69.50	73.54	50.09	59.59
SWN_N	54.08	75.44	63.00	64.06	40.59	49.70
SWN_ARN	54.08	75.44	63.00	64.06	40.59	49.70
SWN_ARVN	54.08	75.44	63.00	64.06	40.59	49.70
SWN_ARV	54.05	75.27	62.90	63.90	40.59	49.65
SWN_AR	54.05	75.27	62.90	63.90	40.59	49.65
SWN_V	54.05	75.27	62.90	63.90	40.59	49.65
SWN_A	54.05	75.27	62.90	63.90	40.59	49.65
SWN_R	54.05	75.27	62.90	63.90	40.59	49.65
SWN	53.97	74.27	62.51	62.03	40.12	48.73

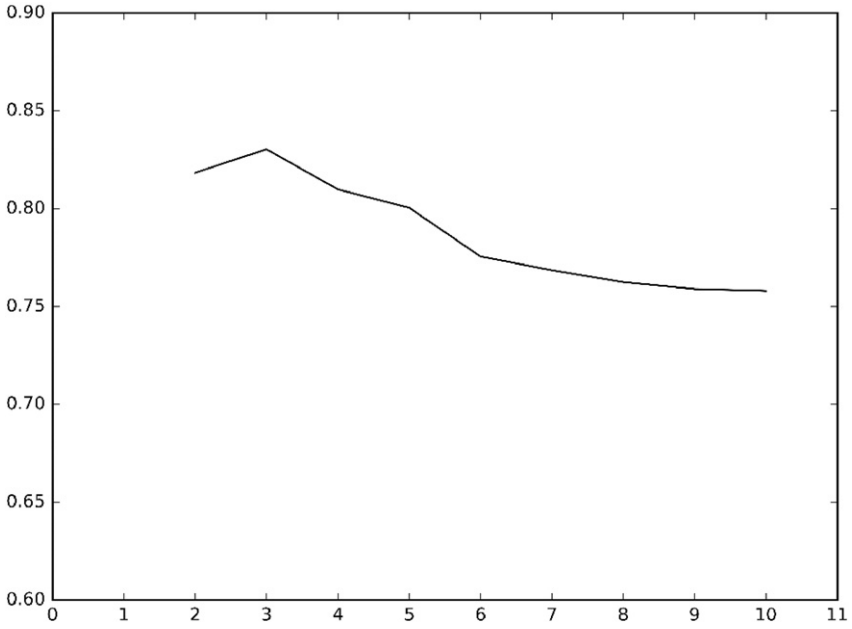

Fig. 3. Sensitivity of *F*-measure to word frequency threshold.

(i.e., positive, negative, or neutral) and the strength of subjectivity (i.e., strong or weak). A word that is neutral in polarity, for example, "absolute," can be strongly subjective. 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, 4153 against 2304. It has also been used in OpinionFinder (OF), a software package that is used by business researchers to analyze sentiment in social media texts [15] and news [61].

SWN contains sentiment-bearing synsets automatically identified from WordNet, a dictionary widely used for word sense disambiguation (WSD) [39]. 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 − (positive score + 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 [62] and Fu et al. [23] did such a conversion by averaging the sentiment score of a word across synsets. Zhang et al. [63] and Balasubramanyan et al. [31] converted the continuous scale to discrete values. Thet et al. [64] adopted the highest sentiment score of each word. Loughran and McDonald [65] used both the first sense of each word and the average

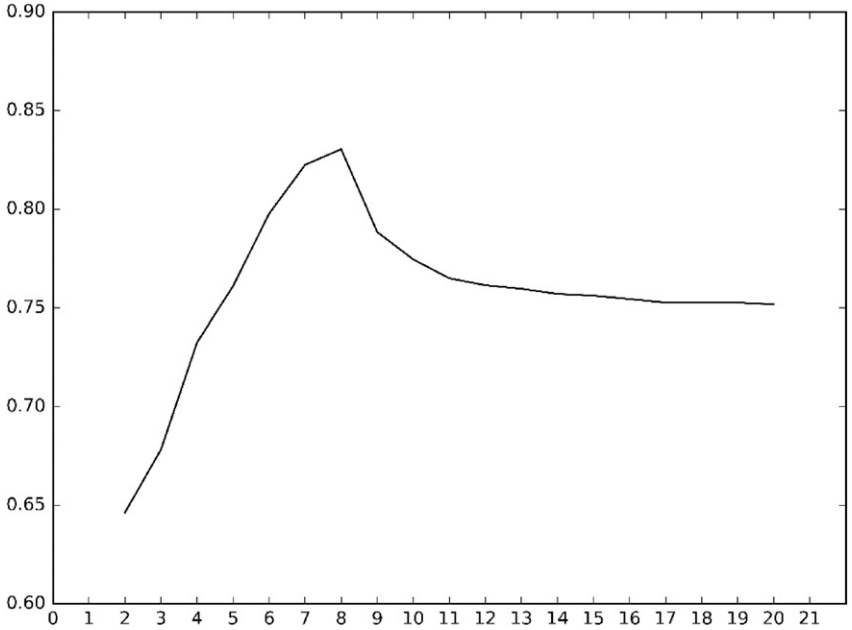

Fig. 4. Sensitivity of *F*-measure to *H*.

score and found that the average-sense method generally performs better than the first sense method. We used a conversion similar to that of [23]. That is, for each word in SWN, we used 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 [19]. 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. To be consistent with and comparable to the baseline lexicons, the expanded lexicons did not retain POS tags either.

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 high agreements between most lexicon pairs indicate that the general sentiment lexicons consist largely of commonly used sentiment words that are applicable across domains. It points to the need for adding domain-specific sentiment words to these lexicons. 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 the general sentiment lexicons severely lack domain-specific entries. This is precisely the area that our method aims to improve upon.

We also created another baseline lexicon (coded as Combined4) by combining all of the lexicons mentioned above, except SWN. Conflicts across lexicons were removed. This will not significantly reduce the lexicon size since the proportion of conflicts is small. SWN was excluded from the combined sentiment lexicon because it contains much more entries than any other lexicon. Were it included, it would have dominated the combined lexicon. Table 2 lists the numbers of positive and negative entries in the baseline lexicons.

4.2. 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 were 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 were written in a non-English language and only 11 tweets were not related to the stock market. This strongly indicates the validity of the developing corpus used in this study.

To extract candidates, we first POS-tagged the developing corpus. In this step, we used a state-of-the-art POS tagger for tweets [66]. After the POS tagging, 95,513 unique tokens were obtained. The eight candidate sets, ARVN, ARV, ARN, AR, A, R, V, and N were subsequently created. Hashtags, user mentions, and URL's do not really belong to any of these categories and were therefore not included in the candidate sets. Even if some of these terms could potentially bear sentiment, they are likely to be applicable only in certain contexts and cannot be generalized in the same way as A, R, V, or N. After removing words that occurred less than three times in the developing corpus, words that are not in Wiktionary, stop words, numbers, and hash marks, the A, R, V, N candidate sets contained 4522, 1006, 8796, and 19,332 candidates, respectively. We compared ARVN with all the baseline lexicons and found that many potential sentiment words, for example, *earnings*, *profits*, *rally*, *swing*, *drops*, and *lol*, were not contained in those lexicons, lending strong support to the motivation for our study.

We used the lexicon scoring approach for sentiment classification: if a tweet contained more negative words than positive words, then it was classified as negative; otherwise, it was classified as positive. Since our domain-specific lexicon generation method focuses on distinguishing between positive and negative words, we used only positive and negative tweets for testing. In such a binary classification, there is no neutral class. Thus, when a tweet has equal number of positive and negative tweets, it must be classified into either positive or negative. In this study, we classify such a case as positive. This is because there are more negative words in all of the seed and expanded lexicons. To minimize the impact of the large number of negative words, a tweet is only considered as negative when it has more negative words than positive words. Through some preliminary experiments, we set the sentiment

Table 6
Results using political tweets.

Lexicon	Positive words	Negative words	Precision (%)	Recall (%)	*F*-measure (%)
Combined4_ARVN	4530	10,453	86.02	85.51	85.46
Combined4_ARN	4287	9337	83.86	83.61	83.59
Combined4_N	4200	8837	83.19	83.01	82.99
Combined4_ARV	4008	8349	79.62	79.62	79.62
Combined4_V	3905	7985	79.52	79.52	79.52
Combined4_AR	3756	7562	78.43	78.42	78.42
Combined4_A	3732	7490	78.43	78.42	78.42
Combined4_R	3673	7299	78.24	78.22	78.22
Combined4	3649	7231	78.24	78.22	78.22
GI_ARVN	4348	5321	65.86	64.25	63.31
GI_N	3205	3790	65.80	63.85	62.70
GI_ARN	3673	4333	65.57	63.65	62.49
GI	1637	2005	65.87	62.54	60.48
GI_V	2571	3230	63.42	61.66	60.34
GI_A	2234	2653	65.25	62.26	60.30
GI_R	2010	2394	65.60	62.26	60.11
GI_ARV	3009	3758	63.05	61.36	60.05
GI_AR	2332	2765	64.97	61.96	59.93
OL_ARVN	3167	10,489	73.99	73.62	73.52
OL_ARN	2849	8471	73.13	72.92	72.86
OL_ARV	2519	6718	72.48	72.33	72.28
OL_V	2352	6061	72.27	72.03	71.96
OL_N	2689	7645	71.29	71.23	71.21
OL_R	2036	4882	70.88	70.34	70.14
OL	2006	4783	70.88	70.33	70.14
OL_AR	2173	5281	70.53	70.04	69.85
OL_A	2139	5163	70.53	70.04	69.85
LM_ARV	1261	9296	70.28	70.23	70.21
LM_N	1029	6061	70.11	70.03	70.00
LM_ARN	1261	7365	69.79	69.52	69.43
LM_AR	575	3336	71.13	69.84	69.37
LM_ARVN	1595	9296	70.14	69.42	69.15
LM_A	540	3123	70.97	69.54	69.01
LM_V	700	4072	69.12	68.64	68.63
LM_R	387	2515	71.73	69.15	68.19
LM	354	2349	71.41	68.63	67.59
MPQA_ARVN	3517	8631	75.32	75.32	75.32
MPQA_ARN	3165	6761	75.90	75.43	75.31
MPQA_N	3016	6146	75.94	75.33	75.18
MPQA_ARV	2828	5513	72.90	71.84	71.50
MPQA_V	2664	5037	72.69	71.54	71.17
MPQA_AR	2462	4525	72.06	70.44	69.88
MPQA_A	2427	4448	72.06	70.44	69.88
MPQA_R	2331	4219	72.11	70.44	69.87
MPQA	2304	4153	72.12	70.43	69.86
SWN_ARN	14,228	15,778	59.47	58.46	57.31
SWN_N	14,182	15,734	59.47	58.46	57.31
SWN_ARV	14,050	15,679	59.47	58.46	57.31
SWN_AR	13,906	15,546	59.47	58.46	57.31
SWN_V	13,999	15,633	59.47	58.46	57.31
SWN_A	13,893	15,540	59.47	58.46	57.31
SWN_R	13,867	15,507	59.47	58.46	57.31
SWN	13,854	15,501	59.47	58.44	57.30
SWN_ARVN	14,369	15,908	59.37	58.36	57.19

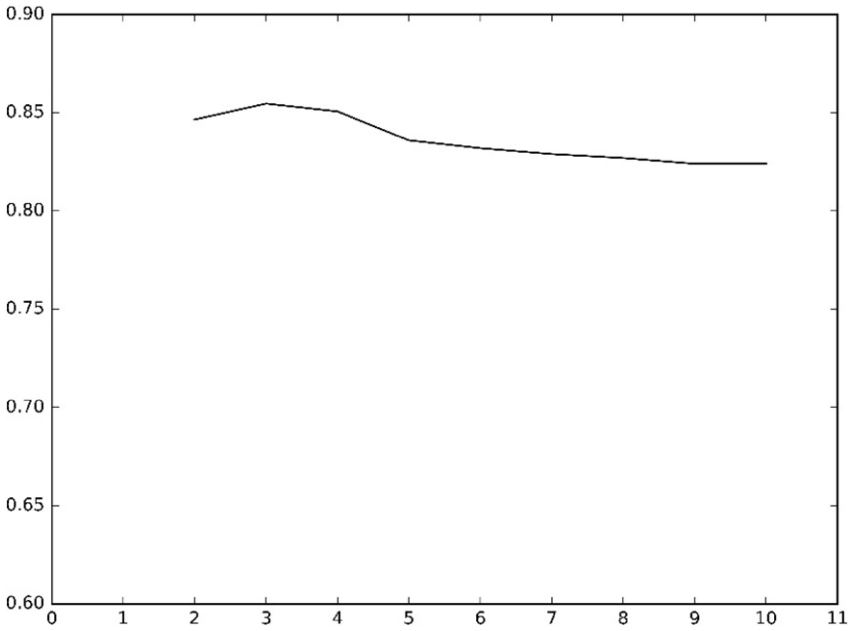

Fig. 5. Sensitivity of *F*-measure to word frequency threshold (political domain).

benchmark parameter *H* to 8. This *H* value prevents our method from identifying too many or too few sentiment words.

We focused on evaluating the performance of our expanded lexicons in tweet-level sentiment classification. We manually created a testing data set using real-world tweets. 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. Following Aggarwal, et al. [67], each tweet was labeled by two raters. When there was a conflict, a third student's decision was used. The labeled data set contains more negative tweets than positive tweets. The Cohen's Kappa is 0.79, indicating substantial agreement [68]. In order to make the classification metrics easier to interpret, we randomly sampled negative tweets to match the number of positive tweets. The final testing set contains 584 positive tweets and 584 negative tweets. The evaluation of our proposed method was based on its performance (precision, recall, and *F*-measure) in classifying these tweets.

5. Results and discussion

5.1. Main results

Table 3 presents the main results of our experiments. The expanded lexicons are denoted as <baseline lexicon>_<candidate set>. The results show that the expanded lexicons using candidate sets with multiple tags generally outperformed the baseline lexicons, except the SWN group. The performance of many of our expanded lexicons, especially

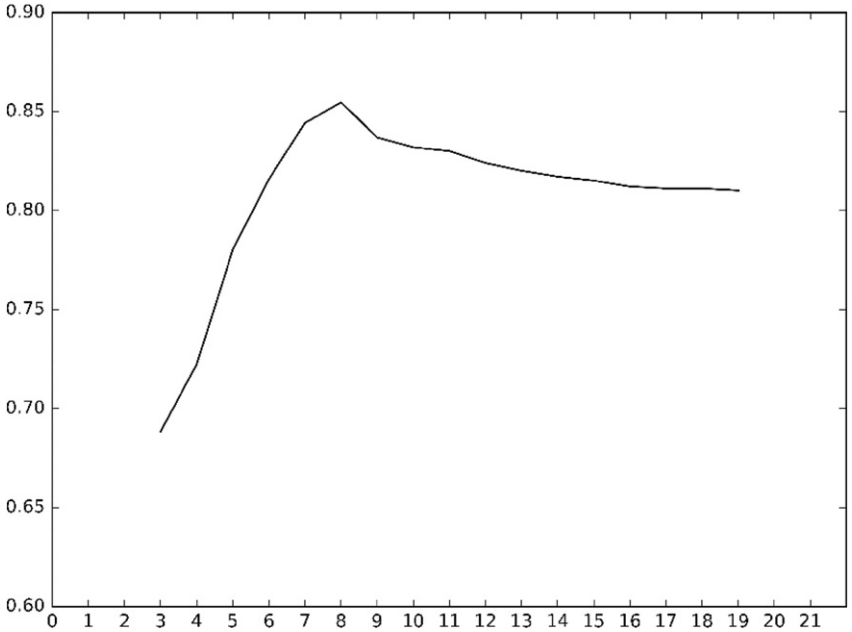

Fig. 6. Sensitivity of *F*-measure to *H* (political domain).

Combined4_ARVN, has been deemed acceptable in prior studies [13,42]. The Combined4 lexicon outperformed all of other seed lexicons, showing the usefulness of using a large and high quality set of sentiment words. Expanding the Combined4 lexicon with the ARVN candidates further significantly improved the performance by almost 10% in *F*-measure, 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. More sample words are presented in Table 4. We believe that the significantly better performance of the Combined4 lexicon is because of the large coverage of sentiment words and most of them were hand-picked.

In the MPQA group, the best result was achieved by the ARVN candidate set. The improvement in *F*-measure is above 7%. In the OL group, the best result was also achieved by the ARVN candidate set, with more than 4% increase in *F*-measure. These two lexicons were developed more recently than GI. Although the sentiment words in GI were carefully selected, the lexicon was created more than half a century ago. Thus, it is not likely to capture the latest characteristics in the social media language. Through our lexicon expansion process, the *F*-measure was improved by over 4% using the ARVN candidate set in the GI group. Although LM was created using a collection of financial reports and is specialized in the finance domain, it contains much more negative words than positive words. The original LM contains 2349 negative words and only 354 positive words. The overwhelming number of negative words dominated the lexicon generation procedure as the expanded lexicon contains 4350 negative words and only 365 positive words. Such an unbalanced lexicon caused a strong bias toward the negative class. Moreover, LM was developed based on a corpus of financial reports. In terms of language domain, financial reports differ greatly from microblogs. Thus, a lexicon developed using financial reports is not likely to effectively capture sentiment indicators in microblogs. This also supports our motivation to address the challenge of language domain in sentiment analysis. As a result, classification performance of LM on social media language is low. The performance of the expanded lexicons in this group is also limited. Nonetheless, the ARV candidate set was able to improve the *F*-measure by almost 5%. Among the original lexicons, SWN has the worst performance. Expanding SWN did not gain improvement, either. This is likely due to the fact that the creation process of SWN is largely automatic, resulting in the inclusion of a large number of words whose sentiment depends on context. These results indicate that the seed lexicon is crucial in lexicon expansion. With a high quality seed lexicon, more domain-specific sentiment words can be learned. This finding is consistent with Maks and Vossen [69].

We observed a pattern among the results using individual-tag candidate sets (namely, A, R, V, and N). Using only A or R did not improve the classification performance as we expected. Although adjectives and adverbs are most likely to bear sentiment, we cannot ignore the fact that most seed lexicons already contain a large amount of A and R. Among the four POS categories, N and V are the largest and second largest sets, providing more potential for identifying new sentiment words. However, combining all of the four POS categories tend to yield superior performance. We believe that the interaction between words of different POS tags played a key role in the lexicon expansion process. In particular, tweets are short. Thus, there are likely to be limited co-occurrence between words of the same POS tags. Considering co-occurrence between words with different tags allows more potential in capturing the sentiment indicators. The results support our motivation for combining POS categories when creating candidate sets.

We conducted paired *t*-test to examine if the *F*-measures of the expanded lexicons are significantly higher than those of the seed lexicons. We compared a list of the highest *F*-measures achieved by the expanded lexicons in each experiment group and a list of the *F*-measures achieved by the seed lexicons. The difference is significant at the 5% level ($p = 0.017$), suggesting that the performance results generated by the expanded lexicons are significantly better than those generated by the seed lexicons.

5.2. Class-level results

The class-level results are reported in Table 5. Across all groups, the *F*-measure in the negative class showed more improvement than that in the positive class through the lexicon expansion process. For instance, in the Combined4 group, the *F*-measure in the negative class improved from 71.85% to 81.05% while the *F*-measure in the positive class improved from 73.70% to 79.50%. The largest class-level improvement, 15.13%, is achieved in the MPQA group by the ARVN candidate set. We believe that the improvement in the negative class is higher than that in the positive class because of the fact that most seed lexicons we used contain much more negative words than positive words. Such a large amount of knowledge of negative words in the seed lexicons results in the better identification of negative words.

5.3. Sensitivity analysis

To obtain our main results, we set the threshold for word frequency to be 3 and the threshold for parameter *H* to be 8. To show the sensitivity of the results to these parameters, we further report the results using the Combined4 seed lexicon and the ARVN candidate set with different threshold values. We select this combination since it has achieved the best classification performance among the results reported in previous sections. First, we vary the threshold for word frequency, while fixing *H* at 8. Fig. 3 shows the change in *F*-measure by varying the word frequency threshold from 2 to 10. The vertical axis show *F*-measure while the horizontal axis shows word frequency threshold. The best *F*-measure was achieved when the threshold is set to 3. When the threshold further increases, the set of candidate words reduces, the expanded lexicons approach the seed lexicon, and thus the performance improvement diminishes. Although the number of candidate words is largest when the threshold is set to 2, the *F*-measure in this setting is not as high as when the threshold is 3. This indicates 3 is an appropriate threshold for avoiding using too many or too few candidate words for our tweet corpus related to stocks.

We further present the results using the Combined4 seed lexicon and the ARVN candidate set by fixing the word frequency threshold to 3 and varying *H* from 2 to 20. Fig. 4 shows the change in *F*-measure for this setting. The *F*-measure peaked when *H* was 8 and decreased when *H* moved away from the peak. When *H* is too small, too much noise is introduced, hurting the sentiment classification performance. When *H* is too large, too few candidate words are identified as sentiment words, the expanded lexicons approach the seed lexicon, and thus the performance improvement diminishes.

5.4. Testing in another content domain

Our proposed method is general and can be readily used to learn sentiment words in another domain. To show such generalizability, we conducted another round of experiments using a different corpus. This corpus consists of one million tweets related to political topics, such as United States presidential election. Analyzing these tweets has important practical implications since Twitter is an important source for monitoring public opinion during political campaigns [12,70,71]. We also created a testing set containing 500 positive political tweets and 500 negative political tweets in the same manner as we created the testing set in the stock market domain.

We used similar settings as in our main results. That is, we set the word frequency threshold to 3 and *H* to 8. We used all of the aforementioned seed lexicons and candidate sets. Table 6 shows the numbers of positive and negative words, precision, recall, and *F*-measure, grouped by seed lexicons. The highest *F*-measure, 85.46%, was obtained using

Combined4 as seed lexicon and ARVN as candidate set. This reflects more than 7% improvement over the seed lexicon baseline, 78.22%. Most other candidate sets also showed improvement. This is consistent with our main results. These results support the domain awareness capability of the proposed lexicon expansion method. Figs. 5 and 6 show the sensitivity analysis of the word frequency threshold and *H*, respectively. The classification performance peaked when word frequency threshold was set to 3 and when *H* was set to 8. These results are also consistent with our previous evaluation in the stock market domain.

6. Conclusion

In this study, we identified the challenges posed by the content domain and language domain in sentiment classification. To address these challenges, we proposed a lexicon expansion method to improve sentiment classification by learning domain knowledge. We tested our method with two large unannotated developing corpora and five existing sentiment lexicons as seeds and baselines. The evaluation results show that the expanded lexicons improved the sentiment classification performance significantly compared to the seed lexicons.

Our research makes several contributions to both research and practice. First, the proposed algorithm can automatically generate a domain-specific sentiment lexicon based on a seed lexicon and an unannotated corpus, both of which can be easily obtained. It can be used by sentiment analysis practitioners and researchers to improve lexicon-based sentiment analysis without incurring expensive human labor costs. It also creates opportunities for new research and applications in domains where sentiment analysis was previously not effective. Second, our experiments show that the proposed algorithm is indeed powerful in constructing a better sentiment lexicon in the microblogging and finance domains. This will directly benefit analytics research using microblogging messages to predict financial outcomes, such as [1,15, 72–74]. Third, through a series of experiments, we have shown that the effectiveness of the generated sentiment lexicon depends on the seed lexicons and word classes of choice. Based on the results, we speculate that a conservative candidate set (i.e., AR or ARV) should be used if a small seed lexicon is used. When a large and high-quality seed lexicon is used, the expanded lexicon will benefit more from a large and complete candidate set (i.e., ARVN).

Our proposed method also has methodological implications. It has shown that co-occurrence based measures are useful in finding important associations in unstructured data. This applies not only to lexicon expansion, but to a wide range of text analytics applications as well. By finding semantic similarity between words, our method has also shown that resolving word ambiguity can indeed improve lexicon-based sentiment classification.

This study has a few limitations, which could be addressed in future research. First, the convergence of the lexicon expansion process depends on the threshold used to establish a relationship between candidate words. The lower the threshold, the higher is the number of words added; increasing the threshold reduces that number. When more words are added to the expanded lexicon, they are likely to contain more non-sentiment words (i.e., higher false-positive rate). When fewer words are added, more true sentiment words are likely to be missed (i.e., higher false-negative rate). How to choose an optimal threshold remains an open question and is worth exploring in future research. Second, the proposed method is based on the heuristic that frequently co-occurring words are likely to have similar sentiment orientation. The effectiveness of the proposed method is also limited by the limitation of that heuristic. Third, many hashtags (e.g., #FeelsGoodMan) are strong sentiment expressions. However, our method does not consider hashtags since only A, R, V, N words are used as candidate words. How to identify sentiment-bearing hashtags is an interesting future research direction.

This study can also be extended in other directions. 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. In our lexicon expansion process, positive and negative words are identified. In future research, we are interested in further identifying neutral words to increase the effectiveness of the process.

References

[1] Y. Yu, W. Duan, Q. Cao, The impact of social and conventional media on firm equity value: a sentiment analysis approach, Decis. Support. Syst. 55 (2013) 919–926.
[2] A. Bifet, E. Frank, Sentiment knowledge discovery in twitter streaming data, Discovery Science 2010, pp. 1–15.
[3] B. Liu, Sentiment analysis and opinion mining, Synthesis Lectures on Human Language Technologies, 5, 2012, pp. 1–167.
[4] X. Luo, J. Zhang, How do consumer buzz and traffic in social media marketing predict the value of the firm? J. Manag. Inf. Syst. 30 (2013) 213–238.
[5] R. Divol, D. Edelman, H. Sarrazin, Demystifying social media, McKinsey Q. 2 (2012) 66–77.
[6] H. Chen, R.H. Chiang, V.C. Storey, Business intelligence and analytics: from big data to big impact, MIS Q. 36 (2012) 1165–1188.
[7] M. Chau, J. Xu, Business intelligence in blogs: understanding consumer interactions and communities, MIS Q. 36 (2012) 1189–1216.
[8] B. Bickart, R.M. Schindler, Internet forums as influential sources of consumer information, J. Interact. Mark. 15 (2001) 31–40.
[9] T. Sakaki, M. Okazaki, Y. Matsuo, Earthquake shakes Twitter users: real-time event detection by social sensors, Proceedings of the 19th International Conference on World Wide Web 2010, pp. 851–860.
[10] H. Chen, D. Zimbra, AI and opinion mining, IEEE Intell. Syst. 25 (2010) 74–80.
[11] A. Abbasi, S. France, Z. Zhang, H. Chen, Selecting attributes for sentiment classification using feature relation networks, IEEE Trans. Knowl. Data Eng. 23 (2011) 447–462.
[12] S. Stieglitz, L. Dang-Xuan, Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior, J. Manag. Inf. Syst. 29 (2013) 217–248.
[13] R. Aggarwal, R. Gopal, A. Gupta, H. Singh, Putting money where the mouths are: the relation between venture financing and electronic word-of-mouth, Inf. Syst. Res. 23 (2012) 976–992.
[14] G. Mishne, N. Glance, Predicting movie sales from blogger sentiment, Proceedings of AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW 2006) 2006, pp. 301–304.
[15] J. Bollen, H. Mao, X. Zeng, Twitter mood predicts the stock market, J. Comput. Sci. 2 (2011) 1–8.
[16] C. Oh, O.R.L. Sheng, Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement, ICIS 2011 Proceedings 2011, pp. 57–58.
[17] E. Boiy, M.F. Moens, A machine learning approach to sentiment analysis in multilingual web texts, Inf. Retr. 12 (2009) 526–558.
[18] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up?: sentiment classification using machine learning techniques, Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, 10, 2002, pp. 79–86.
[19] M. Hu, B. Liu, Mining and summarizing customer reviews, Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2004, pp. 168–177.
[20] T.L. Ngo-Ye, A.P. Sinha, Analyzing Online Review Helpfulness Using a Regressional Relief-enhanced Text Mining Method, *ACM Transactions on Management Information Systems*, 3, 2012, 1–20.
[21] E. Riloff, J. Wiebe, Learning extraction patterns for subjective expressions, Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing 2003, pp. 105–112.
[22] P. Chaovalit, L. Zhou, Movie review mining: a comparison between supervised and unsupervised classification approaches, Proceedings of the 38th Annual Hawaii International Conference on System Sciences 2005, p. 112c.
[23] T. Fu, A. Abbasi, D. Zeng, H. Chen, Sentimental Spidering: leveraging opinion information in focused crawlers, ACM Trans. Inf. Syst. 30 (2012) 24.
[24] D.E. O'Leary, Blog mining-review and extensions: "From each according to his opinion", Decis. Support. Syst. 51 (11//2011) 821–830.
[25] B. Pang, L. Lee, Opinion Mining and Sentiment Analysis: Now Pub, 2008.
[26] X. Ding, B. Liu, P.S. Yu, A holistic lexicon-based approach to opinion mining, Proceedings of the International Conference on Web Search and Web Data Mining 2008, pp. 231–240.
[27] T. Loughran, B. McDonald, When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks, J. Financ. 66 (2011) 35–65.
[28] T. Wilson, J. Wiebe, P. Hoffmann, Recognizing contextual polarity: an exploration of features for phrase-level sentiment analysis, Comput. Linguist. 35 (2009) 399–433.

[29] R. Raina, A. Battle, H. Lee, B. Packer, A.Y. Ng, Self-taught learning: transfer learning from unlabeled data, Proceedings of the 24th International Conference on Machine Learning 2007, pp. 759–766.
[30] S. Brody, N. Diakopoulos, Cooooooooooooooollllllllllll!!!!!!!e!!!!!!!: using word lengthening to detect sentiment in microblogs, Proceedings of the Conference on Empirical Methods in Natural Language Processing 2011, pp. 562–570 (Edinburgh, United Kingdom).
[31] R. Balasubramanyan, W.W. Cohen, D. Pierce, D.P. Redlawsk, What pushes their buttons?: predicting comment polarity from the content of political blog posts, Proceedings of the Workshop on Languages in Social Media 2011, pp. 12–19.
[32] R.P. Schumaker, H. Chen, Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZF in Text System, *ACM Transactions on Information Systems*, 27, 2009, 12.
[33] A. Go, R. Bhayani, L. Huang, Twitter sentiment classification using distant supervision, CS224N Project Report 2009, pp. 1–12 (Stanford).
[34] J. Smailović, M. Grčar, N. Lavrač, M. Žnidaršič, Predictive sentiment analysis of tweets: a stock market application, Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data, Springer 2013, pp. 77–88.
[35] R. Abbott, M. Walker, P. Anand, J.E.F. Tree, R. Bowmani, J. King, How can you say such things?!?: recognizing disagreement in informal political argument, Proceedings of the Workshop on Languages in Social Media 2011, pp. 2–11 (Portland, Oregon).
[36] C.-H. Chou, A.P. Sinha, H. Zhao, A hybrid attribute selection approach for text classification, J. Assoc. Inf. Syst. 11 (2010) 491–518.
[37] P.J. Stone, D.C. Dunphy, M.S. Smith, The General Inquirer: A Computer Approach to Content Analysis, M.I.T. Press, Oxford, England, 1966.
[38] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, et al., OpinionFinder: a system for subjectivity analysis, Proceedings of HLT/EMNLP on Interactive Demonstrations 2005, pp. 34–35.
[39] A. Esuli, F. Sebastiani, Sentiwordnet: a publicly available lexical resource for opinion mining, Proceedings of 5th Conference on Language Resources and Evaluation 2006, pp. 417–422.
[40] B. O'Connor, R. Balasubramanyan, B. Routledge, N. Smith, From tweets to polls: linking text sentiment to public opinion time series, Proceedings of the International AAAI Conference on Weblogs and Social Media, 2010.
[41] D. Garcia, Sentiment during recessions, J. Financ. 68 (2013) 1267–1300.
[42] S.R. Das, M.Y. Chen, Yahoo! for Amazon: sentiment extraction from small talk on the web, Manag. Sci. 53 (2007) 1375–1388.
[43] T. Wilson, J. Wiebe, P. Hoffmann, Recognizing contextual polarity in phrase-level sentiment analysis, Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing 2005, pp. 347–354.
[44] A. Bermingham, A.F. Smeaton, Classifying sentiment in microblogs: is brevity an advantage, Proceedings of the 19th ACM International Conference on Information and Knowledge Management 2010, pp. 1833–1836.
[45] V. Hatzivassiloglou, K.R. McKeown, Predicting the semantic orientation of adjectives, Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics 1997, pp. 174–181.
[46] P.D. Turney, Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews, Proceedings of the 40th Annual Meeting on Association for Computational Linguistics 2002, pp. 417–424.
[47] K.W. Church, P. Hanks, Word association norms, mutual information, and lexicography, Comput. Linguist. 16 (1990) 22–29.
[48] J. Wiebe, E. Riloff, Creating subjective and objective sentence classifiers from unannotated texts, Computational Linguistics and Intelligent Text Processing, Springer 2005, pp. 486–497.
[49] N. Oliveira, P. Cortez, N. Areal, Stock market sentiment lexicon acquisition using microblogging data and statistical measures, Decis. Support. Syst. 85 (5//2016) 62–73.
[50] Wiktionary, Wiktionary. Available http://www.wiktionary.org/2013.
[51] L. Zhou, J.K. Burgoon, D.P. Twitchell, T. Qin, J.F. Nunamaker Jr., A comparison of classification methods for predicting deception in computer-mediated communication, J. Manag. Inf. Syst. 20 (2004) 139–166.
[52] L. Velikovich, S. Blair-Goldensohn, K. Hannan, R. McDonald, The viability of web-derived polarity lexicons, Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics 2010, pp. 777–785.
[53] B. Liu, M. Hu, J. Cheng, Opinion observer: analyzing and comparing opinions on the Web, Proceedings of the 14th International Conference on World Wide Web 2005, pp. 342–351.
[54] G. Qiu, B. Liu, J. Bu, C. Chen, Expanding domain sentiment lexicon through double propagation, Proceedings of the 21st International Joint Conference on Artifical Intelligence 2009, pp. 1199–1204.
[55] P. Koehn, K. Knight, Learning a translation lexicon from monolingual corpora, Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition, 9, 2002, pp. 9–16.
[56] Alexa, Available http://www.alexa.com/2013.
[57] Twitter, Celebrating #Twitter7. Available https://blog.twitter.com/2013/celebrating-twitter72013.
[58] W. Xu, T. Li, B. Jiang, C. Cheng, Web mining for financial market prediction based on online sentiments, Proceedings of PACIS 2012, 2012.
[59] P.C. Tetlock, Giving content to investor sentiment: the role of media in the stock market, J. Financ. 62 (2007) 1139–1168.
[60] J. Jiang, N. Yu, C.Y. Lin, FoCUS: learning to crawl web forums, Proceedings of the 21st International Conference Companion on World Wide Web 2012, pp. 33–42.
[61] R.P. Schumaker, Y. Zhang, C.-N. Huang, H. Chen, Evaluating sentiment in financial news articles, Decis. Support. Syst. 53 (6//2012) 458–464.
[62] A. Fahrni, M. Klenner, Old wine or warm beer: target-specific sentiment analysis of adjectives, Proc. of the Symposium on Affective Language in Human and Machine, AISB 2008, pp. 60–63.
[63] Z. Zhang, X. Li, Y. Chen, Deciphering Word-of-Mouth in Social Media: Text-based Metrics of Consumer Reviews, *ACM Transactions on Management Information Systems*, 3, 2012, 5.
[64] T.T. Thet, J.-C. Na, C.S. Khoo, Aspect-based sentiment analysis of movie reviews on discussion boards, J. Inf. Sci. 36 (2010) 823–848.
[65] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, M. Stede, Lexicon-based methods for sentiment analysis, Comput. Linguist. 37 (2011) 267–307.
[66] O. Owoputi, B. O'Connor, C. Dyer, K. Gimpel, N. Schneider, N.A. Smith, Improved part-of-speech tagging for online conversational text with word clusters, Proceedings of NAACL-HLT 2013, pp. 380–390.
[67] R. Aggarwal, R. Gopal, R. Sankaranarayanan, P.V. Singh, Blog, blogger, and the firm: can negative employee posts lead to positive outcomes? Inf. Syst. Res. 23 (2012) 306–322.
[68] J.R. Landis, G.G. Koch, The measurement of observer agreement for categorical data, Biometrics (1977) 159–174.
[69] I. Maks, P. Vossen, Different approaches to automatic polarity annotation at synset level, Proceedings of the First International Workshop on Lexical Resources, 2011 (WoLeR).
[70] A. Tumasjan, T.O. Sprenger, P.G. Sandner, I.M. Welpe, Election forecasts with twitter: how 140 characters reflect the political landscape, Soc. Sci. Comput. Rev. 29 (November 1 2011) 402–418.
[71] D. Gayo-Avello, "I Wanted to Predict Elections With Twitter and All I Got Was This Lousy Paper" – A Balanced Survey on Election Prediction Using Twitter Data, 2012.
[72] S. Nann, J. Krauss, D. Schoder, Predictive analytics on public data—the case of stock markets, ECIS, 2013.
[73] N. Oliveira, P. Cortez, N. Areal, On the predictability of stock market behavior using StockTwits sentiment and posting volume, Progress in Artificial Intelligence, Springer 2013, pp. 355–365.
[74] T.O. Sprenger, A. Tumasjan, P.G. Sandner, I.M. Welpe, Tweets and trades: the information content of stock microblogs, European Financial Management, 20, 2014, pp. 926–957.

Shuyuan (Lance) Deng is a visiting Assistant Professor of Information Technology Management at the Lubar School of Business, University of Wisconsin-Milwaukee. He has published several refereed conference papers in the areas of text mining, sentiment analysis, and the social and economic impact of IS. He is a member of the Association for Information Systems. He received his Ph.D. degree in Information Technology Management from the University of Wisconsin-Milwaukee, an M.B.A. degree from the University of Illinois at Chicago, and an M.A. degree from South China University of Technology.

Atish P. Sinha is a Professor of Information Technology Management and Director of the Center for Technology Innovation at the Lubar School of Business, University of Wisconsin-Milwaukee. He earned his Ph.D. in business, with a concentration in Artificial Intelligence, from the University of Pittsburgh. His current research interests are in the areas of business intelligence and analytics, data mining, text mining, service-oriented computing, and healthcare informatics. His research has been published in several journals, including ACM Transactions on MIS; Communications of the ACM; Decision Support Systems; IEEE Transactions on Engineering Management; IEEE Transactions on Software Engineering; IEEE Transactions on Systems, Man, and Cybernetics; Information Systems Research; International Journal of Human-Computer Studies; Journal of the Association for Information Systems; and Journal of Management Information Systems. Professor Sinha is a member of ACM, AIS, and INFORMS. He served as the program co-chair of the 6th Design Science Research in Information Systems and Technology (DESRIST) conference, and as the co-chair of the 16th Workshop on Information Technologies and Systems (WITS).

Huimin Zhao is a Professor of Information Technology Management at the Lubar School of Business, University of Wisconsin-Milwaukee. He received the B.E. and M.E. degrees in Automation from Tsinghua University, China and the Ph.D. degree in Management Information Systems from the University of Arizona, USA. His current research interests include data mining and healthcare informatics. He has published in such journals as MIS Quarterly, Communications of the ACM, ACM Transactions on MIS, IEEE Transactions on Knowledge and Data Engineering, IEEE Transactions on Systems, Man, and Cybernetics, Information Systems, Journal of Management Information Systems, Journal of the AIS, and Decision Support Systems. He has served as a senior editor for Decision Support Systems and an associate editor for MIS Quarterly. He served as a co-chair of the 19th Workshop on Information Technologies and Systems (WITS), the 5th INFORMS Workshop on Data Mining and Health Informatics, and the 9th China Summer Workshop on Information Management (CSWIM).