A Method for Constructing Large-scale Domain-specific Lexicon Based on Deep Learning

Cai Chongchao[1, 2, *], Xu Huahu[1], Wan Jie[2], Wu Jiaqi[1]

[1]School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China

[2]College of Logistic and Information Engineering, Huzhou Vocational & Technical College, Huzhou, Zhejiang 313000, China

Email:caichongchao@163.com huahuxu@163.com

Keywords: Sentiment Lexicon, Deep Learning, Word2Vec, Sentiment analysis

Abstract: Sentiment analysis mainly refers to the mining and extraction of useful knowledge from subjective texts. Most of existing analysis methods involve either text classification through machine learning or polarity classification through sentiment lexicon. The limited scale of Chinese sentiment lexicon impedes the performance of sentiment analysis for social network. In this context, a domain-specific method is proposed to extract large-scale sentiment lexicon. To achieve this, social network data are extracted and cleaned. Next, the deep representation model Word2Vec is used to learn the extracted data in order to obtain the Chinese lexicon vector and determine the seed words of the stock domain. Experimental results show that in terms of large social network sentiment analysis, the proposed method for acquisition of domain-specific sentimental lexicon is superior to other methods for popular sentiment lexicons and small-scale domain-specific lexicons.

1. Introduction

The broad variety of Internet-based open social networks (blogs, forums and microblogs) has reshaped the transmission of information. In a microblog, users discuss current events or hot topics in the society. Efficiently extracting opinions from the social media to inform enterprise decision-making is called social network-based sentiment analysis [1]. The government can monitor and guide the public opinion on the Internet. The e-commerce shops can maximize their profitability by adapting their production and service to user feedback.

It is a complex task to judge the emotional polarity of text [2,3]. In terms of sentiment analysis, the main techniques used are divided into two categories: One is a machine learning method [1] Tagging training corpus and testing corpus, Using Long Short Term Memory (LSTM) deep learning model [4], Bayes [5,6], SVM [1,7,8]. Wang and Manning [6] found that naive Bayes method is more effective for short texts, The SVM method performs better in classifying long text or longer comments.

Another method is using sentiment lexicon [9] to count the number of positive and negative words in the text to be analyzed. Using their difference to analysis the emotional polarity of text, The construction of an sentiment lexicon can be divided into three types, Manual construction [9,10], Construction by seed words [11,12], Using Wordnet to construct sentiment lexicon [13].

Our focus is the construction of domain-specific sentimental lexicon for the social network. The proposed method provides an approach to guarantee polarity of sentiment words in different domains, which is impossible with the traditional sentiment lexicon. The algorithm steps are as follows. First, a web crawler is used to extract data from different social networks, and the collected data is then cleaned. Second, the Chinese word vector is obtained by learning the extracted data through Word2Vec model. Third, domain seed words are selected, continuously expanding the seed lexicon. Finally, the effectiveness of the sentiment lexicon is tested on a stock forum.

The contribution of this paper is two-fold: a method for constructing a domain-specific sentiment lexicon, and a stock-domain sentiment lexicon.

The remainder of this paper is organized as follows. Section 2 reviews related work on sentiment

analysis and lexicon construction. Section 3 describes the proposed method for the construction of a sentiment lexicon. Section 4 evaluates the effectiveness of the domain-specific lexicon through experimentation and analyzes the experimental results. Section 5 presents the conclusions drawn from the experiment and the direction of future work.

2. Related Work

Sentiment analysis consists of sentiment information mining and summarization. It can be performed at various levels of a sentence, an article or a short text. The sentiment dictionary technique is almost applicable to all levels of sentiment analysis. Sentiment lexicons can be constructed manually, based on existing lexicons, corpus or graph models. The method based on existing lexicons construct sentiment lexicons by jointly selecting seed words and exploiting the relationship between synonym and antonym [14, 15, 16]. The methods based on corpus are available in [17, 18, 19] and the methods based on graph model are available in [20, 21].

2.1 Methods based on lexicon

Expert tagging is one of the most direct methods, for example WordNet[22], General Inquirer(GI) and hownet.

Hu and Liu [23] manually selected some adjectives as the seed words, the sentiment lexicon is expanded by using the synonyms and antonyms of WordNet to expand the seed words. Kamps [14] investigate a graph-theoretic model of WordNet most important relation-synonymy-and propose measures that determine the semantic orientation of adjectives for three factors of subjective meaning. Esuli build SENTIWORDNET [13,24], a lexical resource in which each WORDNET synsets is associated to three numerical scores Obj(s), Pos(s) and Neg(s), describing how objective, positive, and negative the terms contained in the synset are.

2.2 Methods based on corpus

Based on the corpus, it is assumed that the common emotional words in the corpus have the same emotional polarity, and the emotional polarity of words are calculated by using the concurrence information and context information.

Turney [12,25] proposed a method that the phrase has a positive semantic orientation when it has good associations (e.g., “subtle nuances“) and a negative semantic orientation when it has bad associations (e.g., “very cavalier“).the semantic orientation of a phrase is calculated as the mutual information between the given phrase and the word “excellent“, minus the mutual information between the given phrase and the word “poor“. A review is classified as recommended if the average semantic orientation of its phrases is positive.

Our focus is the construction of domain-specific sentimental lexicon for the social network, with an eye toward constructing large-scale Chinese lexicons for different domains. Three stock-domain datasets are constructed, i.e., microblogs short-text dataset, stock-blog short-text dataset, and a stock-forum long-text dataset.

3. Data and Algorithm

3.1 Financial data on sentimental lexicon

Our training data comes from the sina finance blog, Focus on the blog about stock discussion. A total of 410211 articles, with a total size of 2.08G. We use these articles to generate the word2vec sentiment lexicon.

3.2 Testing Data

In order to verify the validity of sentiment lexicon,Weibo (Chinese twitter) are crawled as test data sources respectively.

Table 1: Crawling Test Data

	Data Crawling Date	Data Size
Microblog data	2014.10.24-2017.10.24	177707 microblogs

3.3 Construction of the stock-domain lexicon

(1) Read the 2.08G financial blog data collected using the web crawler from the database, define the line feed sign "\n" as the break point, and sequentially store the data into the same txt document T.

(2) Open the document T, traverse it line by line, and store the contents into the variable L[n], where n denotes the line number, L denotes the text contents. If n=5, L[5] denotes the text content of the 5^{th} line.

(3) Segment the Chinese words in each L[n] using the Jieba approach and store the segmentation results into the temporary variable S. Remove useless words and Arabic numbers by putting the data of S through the "Harbin Institute of Technology lexicon". Store the final results into N[n]. The data N is the text corpus input for Word2Vec.

(4) Perform online batch incremental training using the Online Word2Vec15 approach in the Gensim library.

(5) Choose two seed words "rise" and "fall" and adopt the Word2Vec[26]. This function involves two parameters for class extraction, i.e., positive and negative. Hence, the parameter positive=[rise], and negative=[fall]. The parameter topn can be adjusted to control the number of similar words in the output pairs. It is set to 20 in this paper.

(6) Execute this function, output 20 words and extract those positively related to rise and negatively related to fall.

By repeating the steps above and deleting words that are obviously irrelevant to the stock domain, 365 sentiment words were finally obtained.

4. Experiment and Results

Data preprocessing includes cleaning corpus, simplified Chinese character transformation, segmentation and removal of stop words, and removing duplicated data. There are a lot of news headlines, emoticons and URL links in the original micro-blog, which will increase the noise of the micro-blog text, so the regular expression is used to remove the information in the phase of the corpus cleaning.

After data cleaning, three annotated persons are asked to manually annotate the weibo data. The annotate divides all the weibo into three categories of postive, negtive, and neutral according to the emotional polarity of weibo text. If the annotation results are not consistent, the sentiment of the weibo text is determined by voting method. The results of manual tagging of these data are shown in Table 2.

Table 2: Weibo Data

Postive	Negtive	Neutral	Total
9891	1882	14405	26178

When evaluating the performance of emotion classifiers, the accuracy, recall and F1 scores are used to evaluate the standard of each lexicon.

Table 3: Negtive Calculation formula

	Actual Postive	Actual Negtive
Prediction Postive	PTP	PFP
Prediction Negtive	PFN	PTN

Table 4: Negtive Calculation formula

	Actual Negtive	Actual Postive
Prediction Negtive	NTP	NFP
Prediction Postive	NFN	NTN

Calculation formula of accuracy

$$\text{Precision} = \frac{PTP\left(NTP\right)}{PTP(NTP) + PFP(NFP)} \tag{1}$$

Calculation formula of recall

$$\text{Recall} = \frac{PTP\left(NTP\right)}{PTP(NTP) + PFN(NFN)} \tag{2}$$

Calculation formula of F1 score

$$F1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision+Recall}} \tag{3}$$

Table 5: statistics of different sentiment lexicon

Sentiment Lexicon	Postive	Negtive	Total
Hownet	4568	4372	8940
DLUT	11229	10783	22012
NTUSD	2812	8278	11090
BosonNLP	83084	31680	114764
Tsinghua	5568	4470	10038
BosonNLP+Word2Vec	83214	31736	114950
DLUT+Word2Vec	11437	10887	22324
Word2Vec	240	125	365

The polarity of weibo text as shown in the following formula:

$$\text{SO}_c = \frac{\sum_{i=1}^{n} wt(w_i)}{n}\left(c \in (Postive, Negtive)\right) \tag{4}$$

$\text{wt}(w_i)$ represent the emotional intensity of words in the sentiment lexicon, n denotes the number words of weibo content in the lexicon.

$$\text{SO}= \begin{cases} 1 & \left|SO_{postive}\right| > \left|SO_{negtive}\right| \\ 0 & \left|SO_{postive}\right| = \left|SO_{negtive}\right| \\ -1 & \left|SO_{postive}\right| < \left|SO_{negtive}\right| \end{cases} \tag{5}$$

$so_{postive}$ denotes the emotional intensity of positive words, $so_{negtive}$ denotes the emotional intensity of negtive words.

Input: Weibo content;

Outpt: Accuracy, Recall,F1;

Step 1: Chinese Word Segmentation, The result of the Segmentation is an emotional feature word set, SW = { sw1,sw2,sw3…swn} ;

Step 2: Get the weight of emotion, Get weibo emotional feature vector, VWT = { vwt1, vwt2, vwt3,…,vwtn} .

Step 3 :Put the Emotional feature vecto into (4) and (5) ;

Step 4: Calculate accuracy, recall, F1 score;

Step 5: Output the average of F1 Scor

Table 6: Positive Results

	Postive Accuracy	Postive Recall	Postive F1 Score	Postive Content Num
Hownet	0%	0%	0%	0
DLUT	46.59%	51.72%	49.02%	11173
NTUSD	100%	100%	100%	4
Tsinghua	100%	100%	100%	15
BosonNLP	84.44%	95.86%	89.79%	25076
Word2Vec	93.22%	96.15%	94.66%	7464
DLUT+Word2Vec	88.94%	89.68%	89.31%	13684
BosonNLP+Word2Vec	85.26%	97.22%	90.85%	25044

The results shows:

a) The BosonNLP dictionary and DLUT lexicon have good performance as a general sentiment lexicon

b) Word2Vec performs best, whether in the Positive or Negtive.

In the process of our analysis, we used the methods of counting postive and negtive words separately. The following formula is used to consider the average F1 Score,The statistical results are shown in the Table.

$$Polarity(F1) = \frac{2 * Pos(F1) * Neg(F1)}{Pos(F1) + Neg(F1)} \tag{6}$$

Table 7: Average F1 Score

	Average F1 Score
Hownet	0
DLUT	35.82%
NTUSD	0
Tsinghua	0
BosonNLP	19.80%
Word2Vec	73.65%
DLUT+Word2Vec	51.64%
BosonNLP+Word2Vec	30.75%

The overall performance of Word2Vec is optimal. Then, we divide the weibo data set into 4 sub datasets. Each data set contains 5000 pieces of data and then tests the performance of different sentiment lexicon on the subsets respectively. The test results are shown in the diagram.

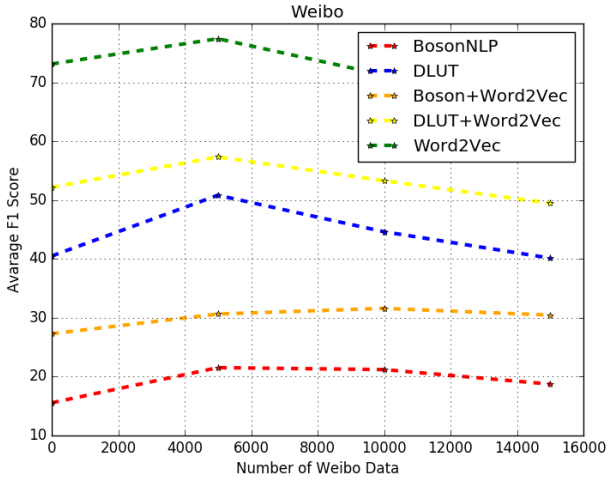

Fig 1. Word2Vec's performs best followed by DLUT combined with Word2Vec

5. Conclusion

Existing domain-specific Chinese sentiment lexicons are few and very limited. To address this problem, a method was proposed to construct large-scale domain-specific lexicons.In this method, the mass data of the social network was collected, denoised and processed using the deep-learning technique Word2Vec. The stock-domain sentiment lexicons were constructed and then applied to several social networks. Results demonstrate the vast superiority of the lexicons constructed using the proposed method.

6. Future Work

The proposed lexicon construction method is semi-automated. More effort will be exerted in making the construction fully automated. Also, its performance was limited when it comes to long texts. In the future, we want to enhance the proposed method and apply it to actual film and product reviews.

References

[1] Pang B, Lee L. Opinion mining and sentiment analysis[J]. Foundations and Trends® in Information Retrieval, 2008, 2(1–2): 1-135.

[2] Soo-Min Kim and Eduard Hovy. 2004. Determining the sentiment of opinions. In Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics, 1367

[3] Soo-Min Kim and Eduard Hovy. 2006. Extracting opinions, opinion holders, and topics expressed in online news media text. In Proceedings of the Workshop on Sentiment and Subjectivity in Text. Association for Computational Linguistics,1–8

[4] Wang J, Yu L C, Lai K R, et al. Dimensional sentiment analysis using a regional CNN-LSTM model[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2016, 2: 225-230.

[5] Zhang H. The optimality of naive Bayes. In: Proceedings of the Seventeenth Florida Artificial Intelligence Research Society Conference, 2004, pp.562–567. American Association for Artificial Intelligence Press.

[6] Wang S and Manning CD. Baselines and bigrams : simple, good sentiment and topic classification. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Jeju, Republic of Korea, 2012, pp.90–94. Association for Computational Linguistics.

[7] Cortes C and Vapnik V. Support vector networks. Machine Learning 1995; 20(3): 273-297.

[8] Vapnik VN and Vapnik V. Statistical Learning Theory. New York: John Wiley and Sons, 1998.

[9] Taboada M, Brooke J, Tofiloski M, et al. Lexicon-based methods for sentiment analysis[J]. Computational linguistics, 2011, 37(2): 267-307.

[10] Stone PJ, Dunphy DC, Smith MS, et al. The General Inquirer: A Computer Approach to Content Analysis. Cambridge, MA: MIT Press, 1966.

[11] Hatzivassiloglou V, Mckeown KR. Predicting the semantic orientation of adjectives. In: Proceedings of 35th Meeting of the Association for Computational Linguistics, 1997, pp. 174–181. Association for Computational Linguistics

[12] Turney P and Littman M. Measuring praise and criticism: inference of semantic orientation from association. ACM Transactions on Information Systems 2003, 21(4): 315–346.

[13] Esuli A and Sebastiani F. SentiWordNet: a publicly available lexical resource for opinion mining. In: Proceedings of 5th International Conference on Language Resources and Evaluation,

2006, pp.417–422

[14] Jaap Kamps, MJ Marx, Robert J Mokken, M de Rijke, and others. 2004. Using wordnet to measure semantic orientations of adjectives. (2004).

[15] Saif Mohammad, Cody Dunne, and Bonnie Dorr. 2009. Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2. Association for Computational Linguistics, 599–608.

[16] Wei Peng and Dae Hoon Park. 2004. Generate adjective sentiment dictionary for social media sentiment analysis using constrained nonnegative matrix factorization. Urbana 51 (2004), 61801.

[17] Yejin Choi and Claire Cardie. 2009. Adapting a polarity lexicon using integer linear programming for domain-specifc sentiment classifcation. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2. Association for Computational Linguistics, 590–598.

[18] Vasileios Hatzivassiloglou and Kathleen R McKeown. 1997. Predicting the semantic orientation of adjectives. In Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 174–181.

[19] Hiroshi Kanayama and Tetsuya Nasukawa. 2006. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In Proceedings of the 2006 conference on empirical methods in natural language processing. Association for Computational Linguistics, 355–363.