A Study of Human Emotion Analysis Based on Social Media

Hanzun Zhang(✉)

School of Journalism and Communication, The Chinese University of Hong Kong, Hong Kong, China
155187089@link.cuhk.edu.hk

Abstract. Social media, a new platform for online interaction, has dramatically changed the way people communicate, interaction and think, while facilitating an explosion of user-generated information. Large number of user-generated social media texts have become one of the most representative data sources of big data in recent years, and mining and analyzing these user-generated information has far-reaching implications for social development. Social media text sentiment analysis, as an information processing technique for analyzing, processing, generalizing and reasoning about subjective texts with emotional overtones, has received widespread attention from academia and industry in recent years, and has been widely applied in many fields of the Internet. Traditional research work on text sentiment analysis focuses on analyzing sentiment from text, but ignores the individual differences in users' expression of sentiment, thus affecting the quality of the analysis results.

Keywords: sentiment analysis · social media · sentiment lexicon · deep learning

1 Introduction

Many current sentiment analysis methods utilize sentiment vocabulary as prior knowledge to varying degrees, whether it is a seed word or a sentiment lexicon. Since there is no need to use large number of manually labeled samples, lexicon-based sentiment analysis methods can be easily and conveniently applied to problems in different domains [1]. Therefore, the quality of the sentiment dictionary directly affects finally result of sentiment analysis. Manual methods for building lexicons are of high quality, but are time-consuming and dependent on the expertise background of the builder. Therefore, mainstream research tends to adopt automatic approaches to construct sentiment dictionaries, including semantic lexicon based on approaches and corpus based on approaches. However, due to the limited size of semantic dictionaries, this greatly limits the capacity of constructed dictionaries and leads to low coverage of dictionaries and difficulty in scaling their use [2].

M. F. b. Sedon et al. (Eds.): SSHA 2023, ASSEHR 752, pp. 174–180, 2023.
https://doi.org/10.2991/978-2-38476-062-6_23

2 Social Media Platform and Sentiment Analysis

2.1 Social Media Platform

With the rise and development of various social media platforms, more and more people tend to express their subjective emotional views and opinions on the Internet about a news event or topic, such as love, agree, dislike, sad, etc. Research on how to use effective technology to analyze this user-generated content and empower computers with sentiment analysis has attracted a lot of attention from both academic and business communities in recent years. The Dictionary of Psychology gives a definition of emotion: "Emotion is the experience of a person's attitude toward whether something objective satisfies his or her needs." The first person to venture into the idea that computers could have artificial emotions or even artificial mental processing was Minsky of MIT, who pointed out in his monograph "The Society of the Mind" in 1985 that the question was not whether intelligent machines could have any emotions, but how they could achieve intelligence without emotions. In response to the explosive growth of user-generated content on the Internet, sentiment analysis is an information processing technology that has emerged in recent years. It is the process of analyzing, processing, summarizing and reasoning about subjective text with emotional overtones, and has a wide range of applications in production and life [3].

2.2 Emotional Analysis

One of the most basic tasks of sentiment analysis is to classify the emotional disposition of a subjective text, i.e., positive or negative sentiment. However, as the research of sentiment analysis continues to deepen, this rough binary polarity classification method can no longer meet the needs of reality, because people often express one or more subtle feelings or emotions, such as "surprise", "disgust", "sadness" and so on. Currently, more and more work considering sentiment analysis and viewing it as a natural, fine-grained evolution of binary emotions. To more clearly illustrate types of emotions and their corresponding voting percentages, the emotional voting statistics are presented in Table 1. As shown in Table 1, the Rappler News website considers eight emotions: fear, anger, disgust, sadness, disinterest, delight, amusement, and inspiration. The last four contain negative emotional tendencies, while the last two have positive emotions and "don't care" is more of a neutral emotion. This division takes in account both subjective and objective considerations as well as fine-grained emotional divisions, providing a more nuanced perspective on the views of news readers [4].

In this paper, we regard news readers' sentiment voting statistics as the natural sentiment label of the news item, so it can be represented as an 8-dimensional sentiment vector in the order of appearance in Table 1 $F \quad =<$

Table 1. Voting results for the news "10 people killed on US campus"

	Afraid	**Amused**	**Angry**	**Annoyed**	**Don't care**	**Happy**	**Inspiration**	**Sad**
Vote	12%	0%	30%	1%	1%	2%	1%	53%

0.12, 0.00, 0.30, 0.01, 0.01, 0.02, 0.00, 0.53 >, where the weight value of each dimension of the vector represents the proportion of votes containing the corresponding sentiment category, while the sum of all sentiment values is 1, so it can also be seen as a distribution of sentiments.

The expression of emotion is usually closely related to the context or domain in which it is used. For example, the Chinese translation of the word "predictable" means "predictable", which is a "happy" emotion when describing the trend of the stock market, and a "disappointed" or "angry" emotion when describing a new movie. When it describes the stock market, it means "happy", while when it describes a new movie, it means "disappointed" or "angry". Therefore, considering the influence of potential topics offers theoretical possibilities for constructing a more accurate emotion lexicon. In this paper, a novel joint non-negative matrix decomposition model is proposed (Joint Nonnegative Matrix Factorization, JNMF), Finally, the sentiment lexicon can be generated from the word-one topic non-negative matrix and the sentiment-one topic non-negative matrix by a combined semantic approach. In general, the contributions of the methodological study of emotion lexicon construction in this chapter are shown below:

(1) Joint non-negative matrix decomposition method is proposed to construct a fine-grained sentiment or emotion lexicon;
(2) Uilding a large-scale, well-formed corpus based on crowdsourced emotionally annotated news corpus;
(3) A high-quality, high-coverage sentiment lexicon was generated, and the effectiveness of the method and the usability of the lexicon were demonstrated by the sentiment subtask of the news headline review data.

3 Related Jobs

Semantic lexicon-based approaches to building sentiment lexicons are often limited by the size of the lexicon, which presents a disadvantage for practical use, so this chapter focuses on corpus-based lexicon studies. Current corpus-based sentiment lexicon research methods rely more and less on the selection of seed words. Xu et al. proposed a lexicon construction algorithm based on a graph model in which seed words are first manually selected for six emotions, and the emotion values of candidate words are calculated by the propagation of the emotions of the seed words on the graph. As mentioned earlier, the method relies heavily on the quality of the seed word selection and thus poses difficulties for practical use. Song et al. and Feng et al. further extended the method by introducing emoticons as seed words. Microblog texts often contain emoticons related to users' sentiments or emotions, and by manually selecting some emoticons with obvious sentiments to replace or supplement the seed words, the quality of the lexicon can be improved while avoiding the influence of too many human factors. In contrast to the above research method. Proposed a method to construct an emotion lexicon based on combinatorial semantics by constructing a word-document matrix and a document-emotion matrix from the news corpus, and generating a word-emotion matrix, i.e., an emotion lexicon, by combining the products of the two matrices. However, this method has a problem in that it ignores the fact that words have diverse and variable emotions under different topics or domains, and that it is difficult to accurately identify the emotions of words while ignoring the topic. Some current research work attempts to model topic and sentiment simultaneously, leading to joint topic- and sentiment-based analysis. However, there is no current work that introduces topic factors into the lexicon construction process [5].

4 Emotional Lexicon Construction Framework

As shown in Fig. 1, this chapter proposes a general framework diagram of the sentiment dictionary construction method, which contains three main components: news crawling and corpus construction, news data pre-processing and sentiment dictionary generation, which are described in detail below.

(1) News crawling and corpus construction.

First, a crawler is used to automatically crawl the news corpus from news websites, where each news item contains sentiment information crowdsourced and labeled by readers. Take the English news corpus data from the Rappler news website as an example, and extract useful information from the news pages, including: news headlines, body content and sentiment voting information [6].

(2) Pre-processing of news data.

Secondly, the news text and headlines are divided into words, lexical reduction and lexical annotation, respectively, an off-the-shelf natural language analysis toolset. In order to filter out emotionally irrelevant words and at the same time ensure that the words in the emotion dictionary have more obvious emotions, the candidate words are filtered using the existing emotion dictionary Corpus, and the processed data are saved in the desired format for dictionary generation [7].

(3) Emotional lexicon generation.

This chapter proposes a two-stage approach to building sentiment lexicon, firstly, non-negative decomposition of word-one document matrix and document-one sentiment matrix by SSNMF or JNMF to obtain topic-related non-negative sub-matrices, and then synthesize the topic-related non-negative sub-matrices into sentiment lexicon by combining semantic model CS [8].

4.1 Experimental Evaluation

This section constructs fine-grained sentiment (or emotion) dictionaries for crowd-sourced annotated sentiment English web news data from Rappler, and evaluates the quality of the dictionaries comprehensively with a sentiment classification task. Meanwhile, the method in this chapter is scalable and can also be directly applied to news annotation data in Chinese (e.g., Sina Social News) to generate a Chinese sentiment dictionary.

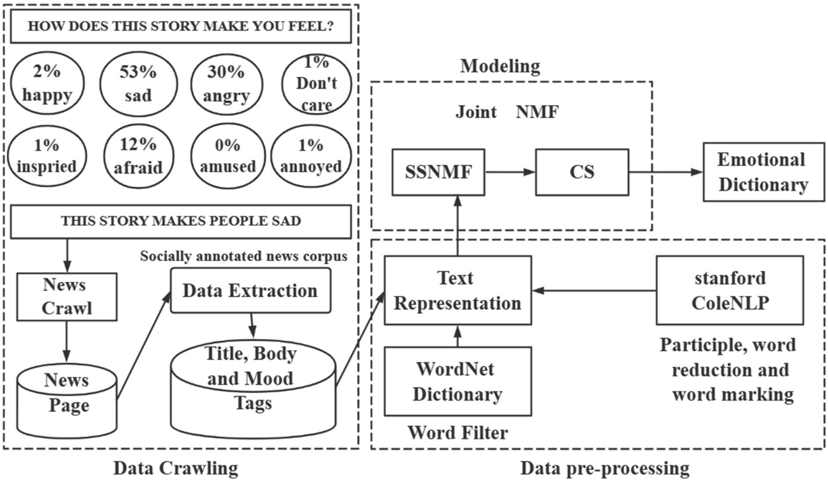

Fig. 1. Framework diagram of the joint non-negative matrix decomposition method

4.2 Experimental Design

Experimental data: To construct the sentiment dictionary, 31, 107 English news articles published before 2015–11-06 were first automatically collected from the Rappler news website using a custom crawler program as a corpus for building the sentiment dictionary. In this paper, the natural processing tool is used to perform preprocessing operations such as word division, lexical annotation (including adjectives, nouns, verbs and adverbs), word form reduction and case conversion. The lexicon represents each word in the form *lemma#POS*, where *lemma* represents the word after morphological reduction and POS is the lexical property. This form of words is also used by WordNet, due to the fact different word forms, may lead to different semantic and emotional expressions. The results of the statistical analysis of the dataset are presented in Table 2, which gives the statistical results of the average number of votes for all news documents in the corpus under each sentiment category. The Rappler news corpus data used in this paper and the sentiment dictionary built on it are available online and can be freely downloaded, studied and researched [9].

As shown in Table 2, the proportion of user votes corresponding to the "happy" emotion is significantly higher than the proportion of user votes for other emotion categories, which indicates that news readers are more inclined to express the "happy" emotion, which is consistent with the real observation. The experiments in this chapter

Table 2. Statistics of sentiment distribution on the Rappler dataset.

	Afraid	**Amused**	**Angry**	**Annoyed**	**Don't care**	**Happy**	**Inspiration**	**Sad**
Average vote	7.8%	10.6%	10.9%	5.9%	5.9%	34.1%	10.3%	14.5%

used only the body text of the news and the sentiment labels to construct the sentiment lexicon, while the headlines of the news were used as the validation set and test set for the lexicon quality evaluation [10].

Experimental environment: The crawler program used for news data collection in this chapter is based on the Java programming language and mainly uses the Jsoup tool to extract the headline, body and sentiment tag information from news pages, while the dictionary construction method in this chapter is based on the Matlab programming language and runs on a high-performance Linux cluster.

4.3 Experimental Results and Evaluation

The emotion dictionary evaluation experiment in this chapter contains three parts: (1). Evaluation by means of an emotion classification task on news headlines; (2). Comparison of the emotion dictionary constructed in this chapter with the best existing emotion dictionaries of similar capacity; (3). Finally, visualization of the advantages of the emotion dictionary constructed in this chapter by listing some emotion words in the dictionary.

5 Conclusion

This chapter proposes a novel joint non-negative matrix decomposition method to construct a fine-grained sentiment or mood lexicon, which is used in a sentiment analysis task. The method generates topic-related non-negative sub-matrices for crowdsourcing labeled sentiment news corpus to generate sentiment lexicon by combining semantics. Finally, by conducting sufficient experiments on the standard evaluation dataset and the large-scale Rappler news headline test dataset constructed in this chapter, the results show that the sentiment dictionary construction method proposed in this chapter can achieve better classification performance on sentiment analysis tasks. At the same time, a comparison with the best extant emotion lexicon shows that the emotion lexicon constructed in this chapter helps to achieve better results in emotion analysis. Future work will consider deep learning methods to construct word vectors containing emotions and use them in fine-grained sentiment or mood analysis tasks.

References

1. Zhang M.X., Chang M.Z.. The impact of media use on political expression among youth groups-examining the moderating effect of political knowledge and media trust [J]. Journalism and Communication Review, 2023,76(01):76-86. DOI:https://doi.org/10.14086/j.cnki.xwycbpl.2023.01.006.
2. Wang Che. Study on the role of UGC in social media communication and media reporting [J]. Journal of Culture,2022(08):129-132.
3. Cai, Bingyi, Wang, Jinjin, Mao, Xianxian. A study of college students' travel information search behavior in social media context[J]. Market Week,2022,35(04):65-70.
4. Sun Yuqing. Research on the key technology of text stance detection for social media[D]. Northeastern Forestry University, 2022. DOI:https://doi.org/10.27009/d.cnki.gdblu.2022.001226.

5. Chen Yuanwen, Wang Xiao, Li Lingxiang, Wang Feiyue. Research and progress of traffic situational awareness based on social media data enhancement[J]. Journal of Intelligent Science and Technology, 2022, 4(01):1-13.
6. Li JG. Systematic reshaping and value interpretation of media governance in digital society [J]. Modern Communication (Journal of Communication University of China), 2022, 44(01): 37-42.DOI:https://doi.org/10.19997/j.cnki.xdcb.2022.01.005.
7. Xu Haoran. Privacy Preference and Risk Analysis Based on Social Media User Behavior[D]. Shandong University, 2021. DOI:https://doi.org/10.27272/d.cnki.gshdu.2021.005916.
8. Fa Hui, Xu Xiao-ting,Zuo Wen-ge. A study on the academic performance of young scholars from a multidimensional perspective: the example of science, technology, agriculture and medicine scholars[J]. Intelligence Inquiry,2021(11):81-88.
9. Ding Haixin, Wang Qin, Li Zhigang. Analysis of public emotional changes and distribution of emotional indices and countermeasures under the influence of social media opinion during the epidemic[J]. New Media Research,2021,7(21):11-14. DOI:https://doi.org/10.16604/j.cnki.issn2096-0360.2021.21.003.
10. Li Yang, Research on social media information recommendation method based on user interest modeling. Heilongjiang Province, Northeast Forestry University, 2021–09–06.