Journal of Computational Science &Engineering 37 (2018) 984-989

Available at http://www.asocse.org ISSN 1710-4068

A survey of research on fine-grained sentiment analysis in Chinese

Yimeng Tang, Peijian Zhang*, Youwei Yu

Qingdao University, Qingdao 266071, China

Abstract: **To review the research progress of fine-grained sentiment analysis, and the classification (namely machine learning classification and classification based on dependency syntax and lexicon). We investigated the key issues, technologies, and application. The research of Chinese fine-grained sentiment analysis in China and the construction of corpus and other resources were introduced. Finally, the application prospect of fine-grained text analysis was introduced. This study helps to understand the key issues and key methods of the current research on fine-grained sentiment analysis.**

Keywords: Fine-grained sentiment analysis; Weight calculation; Evaluation word extraction; Attribute word

Received 20 August 2018, Revised 22 September 2018, Accepted 25 September 2018

*Corresponding Author: Peijian Zhang, peijianzh@126.com

1. Introduction

The popularity of the internet is an important communication platform at present. While promoting people's network communication, it has also produced a lot of commentary information. All kinds of views and comments often have a certain emotional and subjective color, so it also produces the demand for emotional analysis of text generated by the Internet communication platform. For example, public opinion analysis and control, transactions between consumers and businesses, the emotional analysis of the views or comments of participants or consumers is necessary. Micro-blog, circle of friends, forum produce huge amount of speech information every second. The public opinion monitoring technology contains the text clustering analysis, topic extraction, rapid generation of briefings, charts and other analysis results that can provide an analysis basis in order to fully grasp the trend of network public opinion and can make the correct guidance of public opinion.

Before buying the goods, customers often want to get the evaluation of the goods, which is the main reference for the purchase. The merchant also hopes to improve the service or improve the profit by extracting the user's needs and opinions of the comments. However, it is impossible to cope with the emotional analysis task of massive text information by artificial. So it is a hot topic to analyze the emotion of the participants accurately and quickly based on the text data of the massive Internet platform.

2. Definition of emotional analysis

Sentiment analysis is also called opinion mining. It refers to the process of using natural language processing, text mining and other methods to identify and extract the subjective information in the text. The purpose of sentiment analysis is to find out the attitude and intensity of information or comment publishers on certain objects, as well as emotional tag extraction, review summary generation, and so on.

Traditional textual sentiment analysis is mostly coarse-grained sentiment analysis. The emotions contain information (articles, sentences or short texts) are divided into two categories, positive or negative. At the same time, the intensity of sentiment contained in text information is calculated. The traditional text sentiment analysis does not consider the opinions of specific objects on the text information, and it does not further subdivide the polarities. Therefore, it is no longer adapted to the actual needs. In order to make up for the shortcomings of traditional text sentiment analysis methods in emotional subdivision, the researchers proposed a fine-grained sentiment analysis method for text information. At present, domestic research on sentiment analysis is mainly on fine-grained sentiment analysis.

This article reviewed the sentiment classification methods of textual information from the current literature on fine-grained sentiment analysis, and focused on the main issues and methods of fine-grained level sentiment analysis.

3. Process of emotional analysis

There are two ways of sentiment analysis: dependency grammar and dictionary analysis, and machine learning analysis. The analysis steps based on dependency syntax and dictionaries are roughly divided into the extraction of subjective sentences and syntax rules, the identification of emotional words in sentences, and the calculation of emotional scores based on sentiment lexicon for emotional tendencies and emotional strengths. The analysis steps based on machine learning include extracting features, selecting features and getting classification results.

3.1. Text preprocessing: differentiate words, noise reduction and dimensionality reduction

Preprocessing is the premise of expressing text with feature vectors, includes three steps: Differentiate words, Noise Reduction and Dimensionality Reduction. At present, the most popular Chinese word-cutting tools are Jieba, SnowNLP (MIT), Pynlpir, Thula

2018@ The American Computational Science Society. All rights reserved.

(Tsinghua University Natural Language Processing and Social Humanities Computing Laboratory), and PYIPT (Harbin University Language Technology Platform). The text set after cutting the word sometimes contains some frequently used and meaningless words and punctuation such as "IS", "AND", "WITH" and so on. These words are meaningless and increase the similarity between the texts, so that it will cause some interference with the results of the excavation. The way to solve this problem is to build a stop word list or a stop word dictionary to reduce the occurrence of such words. The more used stop word dictionaries include "Harbin Institute of Technology Stop Words Dictionary", "Sichuan University Machine Learning Intelligence Lab Stop Words", "Baidu Stop Word List" and so on. Dimensionality reduction minimizes the dimensionality and computational complexity of vector space without affecting the quality of classification, improves classification speed and classification performance, and finds more meaningful latent variables to help to gain a deeper understanding of the data. The more common methods are singular value decomposition (SVD) and Latent Dirichlet Allocation (LDA) (of course LDA can also be transformed into a model that can directly participate in emotional value calculation).

When constructing a library of supplementary sentiment features for a particular domain, the emotion intensities of the vocabulary have been marked because of the original emotional ontology library. Therefore, the weights extracted from the supplementary emotion feature database are significantly weaker than the original emotion ontology library. Finally, the two are combined to calculate the emotional tendency value.

When calculating the emotional value, the calculation of the weight can better reflect the influence of the important feature words on the document set than using the word frequency alone. At present, there are many applications such as Term Frequency - Inverse Document Frequency (TF-IDF), Word Frequency method (WF), Document Frequency method (DF), Mutual Information (MI), Expected Cross Entropy (ECE), Secondary Information Entropy (QEMI), and Information Gain Method (IG) and so on. However, the traditional weight calculation methods are sometimes not flexible enough. For this reason, Sun[1] proposed a new feature weight calculation method tf *idf *cf. This method quotes the category frequency factor cf, and calculates the representative feature weights in a certain category, which makes the results more objective and accurate. Chen[2] improved the Analytic Hierarchy Process with the G1 method [3]. In order to solve the problem of inconsistency of the judgment matrix, the analytic hierarchy process is improved to increase the accuracy of the weight calculation. The calculation formula of the total weight of the hierarchy is expanded to improves the scope of application of the analytic hierarchy process.

Short texts have the characteristics of uneven distribution of categories, high dimensionality, and sparseness. Therefore, the traditional methods for calculating the weights of long texts are sometimes not suitable. Aiming at this problem, Ma[4] proposed a feature weight calculation method for short texts: a comprehensive category method. This method introduces the concept of anti-document frequency and correlation frequency, and the experimental results have a certain increase in the micro average and macro average. However, the classification performance has not been effectively improved, because the corpus collected from the network contains a large number of network terms and non-standard expression, therefore the system cannot be accurately identified. Therefore, the method of semantic analysis is particularly important. The characteristics of short texts with sparse features and highly unbalanced features make it more necessary to improve the method of feature reduction and feature selection. Better features have great influences on the results of sentiment analysis, so this has also become one of the hot spots in current research.

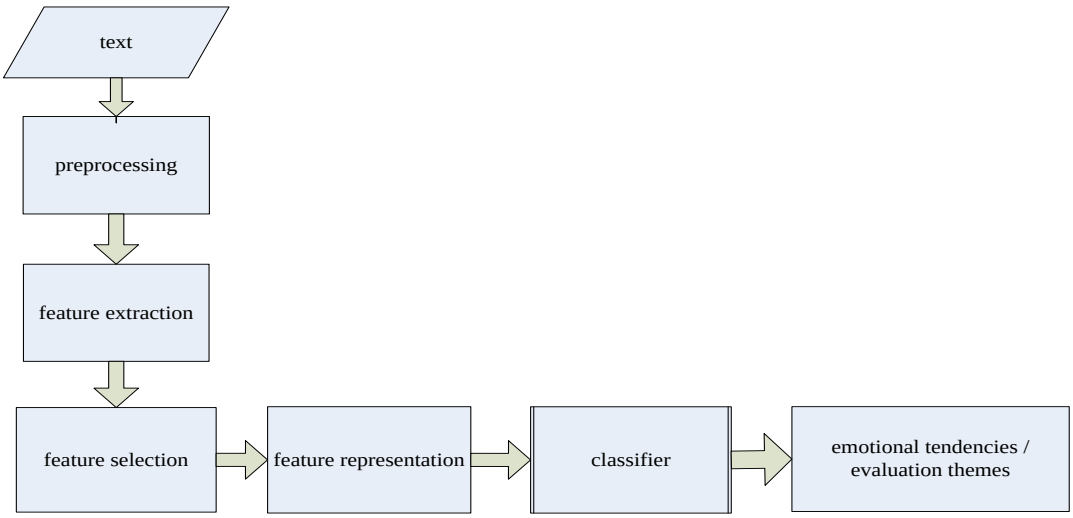

Figure 1. Emotional analysis based on machine learning.

2018@ The American Computational Science Society. All rights reserved.

3.2. Analysis based on machine learning

The classification based on machine learning means that according to the principle of machine learning and training a large number of labeled samples, effective features can be extracted. The classification model can be constructed, then emotional classification will be fulfilled at last [5]. Since data mining has become a hot research point, almost all important machine learning algorithms have been introduced into the field of data mining. However, the classic machine learning algorithms are based on a large number of balanced training samples. These condition settings are too ideal, because a large number of high-quality training samples in practice may not be provided.

For emotional analysis requires a lot of training samples, Su[6] proposed naive Bayes model and Latent Dirichlet Allocation (LDA) to provide appropriate emotional dictionaries and perform progress Emotional tendency analysis without marking the corpus. LDA can be modified to directly conduct emotional digging. Zhang[7] introduces sliding windows in LDA to extract local topics and add emotional layers between the document layer and topic layer so as to fully consider emotional tendency and achieve the purpose of excavation.

Fan[8] proposed a text-based topic and sentiment analysis method basis on a hybrid model. The texts in the training set are labeled with the emotional tendency and the subject category respectively, and are classified with the models of the respective emotions and subject languages. Support vector machine method[9] is a very effective method in the traditional classification, but this method needs to be improved when being adopted in the situation of the Fine Grained Muti Classification. Some researchers have proposed hybrid models, a combination of deep learning and emotional dictionaries, and a combination of machine learning and sentiment lexicon. Ding[10] found a combination of dictionary and LDA, which is higher than the accuracy of that based on dictionary. Based on the entity-based fine-grained sentiment analysis method, Chen[11] provided a double-layer structure Conditional Random Field (CRF) which links the two tasks of emotional entity recognition and emotional tendency classification instead of calculating separately. From the results, the affective entity recognition rate of the double-layer CRF model has been improved relative to the single-layer Linear-chain CRF model. It can be seen that the hybrid model can combine the advantages of machine learning and dictionaries, and it is superior to the performance of only using deep learning or machine learning.

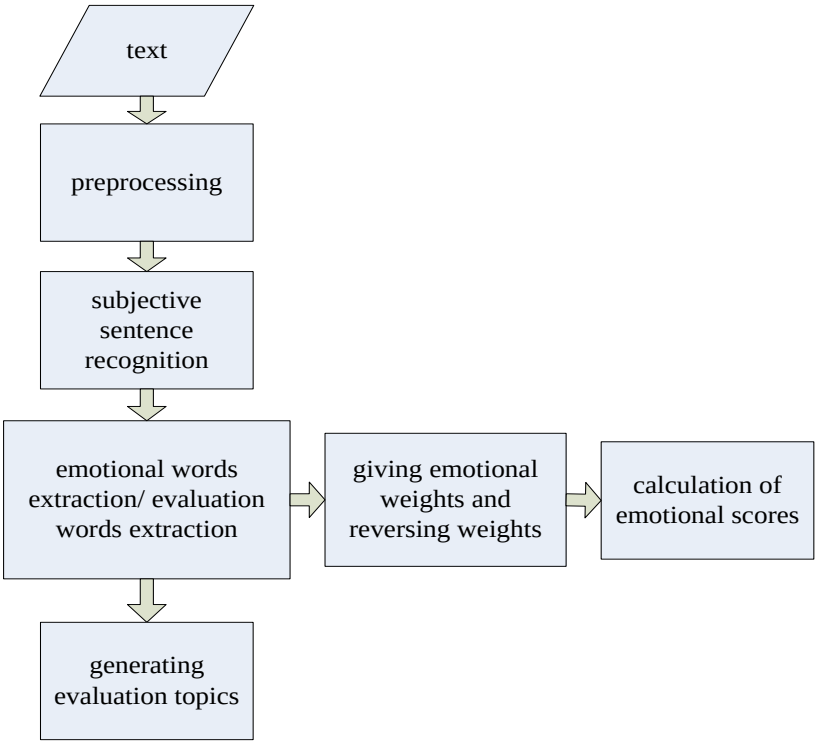

Figure 2. Emotional analysis based on dependency syntax and dictionary.

There are also researchers who use deep learning methods to perform sentiment analysis on feature vectors generated by words. Jiang[12] obtained word vector features, entered the results into Long Short-Term Memory, and used remote monitoring methods to generate a large number of samples to mitigate over fitting. Compared with the MIML-SF model combined with classifier and remote supervision, and the CNN-SF model was constructed from deep learning convolutional neural network. The results show that LSTM has greater advantages in timing information and performance. Although neural

2018@ The American Computational Science Society. All rights reserved.

networks have excellent performance in many fields, neural networks generally have huge data volumes, many parameters, and high performance requirements for running equipment. Therefore, training is time-consuming and labor-intensive. Therefore, fewer researchers use only deep learning methods.

3.3. Emotional analysis based on dependency syntax and dictionary

The sentiment analysis based on dependency syntax and dictionary is mainly divided into steps of establishing emotional dictionary, extracting subjective sentence, dependency parsing, combining dictionary resources and syntax for fine-grained calculation. The method of emotional dictionary, as the name implies, is the process of calculating the emotional tendency and intensity by using the emotional words in the article by the pre-defined emotional dictionary. The sentiment dictionary is more convenient to use because it has defined the polarity of most emotional words before calculation. However, Chinese vocabulary often has a polysemy usage, and even some collocations can cause ambiguity. Therefore, synonyms and ambiguities are also the focus on current research.

3.3.1. Subjective sentence extraction

For the current problems, it is difficult to achieve better results by simply using the dictionary method and the method of syntactic analysis is generally combined with the dictionary. Due to the profound Chinese expression, irony and omission, online buzzwords will interfere with the true meaning of machine-recognized text.

Among the large number of commentary texts, some are subjective sentences with emotional tendencies, such as "I don't like durian, that's too stinky." Some are objective sentences that describe objective facts, such as “One of the characteristics of durian is its strange smell”. Objective sentences have little effect on sentiment analysis and increase the burden of processes. Therefore, the irrelevant objective sentences must be removed before emotional word extraction. There are two ways to identify subjective sentences and objective sentences: semantic-based methods and machine-based learning methods. Semantic-based methods are generally based on part-of-speech calculation scores and threshold comparisons. Exceeding a threshold value is a subjective sentence, and vice versa is an objective sentence. For example, the method proposed by[13] falls into this category. Based on the machine learning method, the judgment of the subjective and objective sentence is generally defined as a binary classification and executed by the classifier for judgment. The key to this method lies in the selection of classification algorithms and classification features. For example, Fang[14] proposed a semi-supervised classifier method, which also overcomes the disadvantages of machine learning that requires a large number of samples to some extent. On this basis, the accuracy and speed of the classifier are also improved.

3.3.2. Emotional word extraction

Emotional word extraction based on sentiment knowledge uses the existing sentiment dictionary to assign emotional sentiment to words or evaluation units with emotional tendencies in the text, and then calculates the emotional tendency of the whole text. The product attribute is the comment point when the user comments on the product. For example, the product attribute in "The color is very beautiful" is "color". Therefore, an accurate and perfect emotional lexicon is needed as a basis to accurately mine product attributes, and accurately analyze the corresponding commentary emotions and emotional strength. First extract the emotional attributes in the sentence, and then calculate the emotional points corresponding to the attributes. Therefore, if the accuracy of the previous step cannot be guaranteed, accurate analysis will be difficult later. For this issue, Huang[15] calculated and analyzed the similarity between words and obtained seven fine-grained emotional lexicons with weight values.

The same words are expressed differently in different professional contexts. For example "the high energy consumption of such a car" and the "high visibility of the light stick at night" are different in different fields. Therefore, when researching different fields, it is necessary to expand the dictionary in a specific field. In response to this problem, Liu[5] proposed a method of expanding the dictionary, and extended the domain sentiment dictionary based on the general sentiment dictionary and the corpus of the online review of the textbook, and designed the domain sentiment orientation analysis algorithm based on the syntax analysis. In the same field, there are sometimes ambiguities in different contexts, so the dictionary of emotional ambiguities is also widely studied. Cai[16] and other contextual categorization produced a collocation dictionary of different types of emotional ambiguous words based on contextual vocabulary, which improved the accuracy of fine-grained sentiment analysis. However, domestic construction of the emotional dictionary didn't exist for a long time, so that resources are not as comprehensive as abroad.

Some scholars have proposed a cross-language emotional classification, that is, using a more complete English sentiment dictionary for Chinese sentiment analysis. Tang[17] a cross-language fine-grained sentiment analysis algorithm based on dependency syntax. Compared with the original emotion evaluation unit extraction method, this method improves the extraction efficiency to some extent. This method first extracts the emotion evaluation unit and then translates it, so that it can reduce the dependence on machine translation, and effectively utilizes the English vocabulary with richer resources. It also tends to translate Chinese emotion units into higher frequency English basic vocabulary through machine translation. This method combines the advantages of synonymy

2018@ The American Computational Science Society. All rights reserved.

and extended emotional lexicon, especially in some languages lacking corpus resources, such as some minority language analysis. The combination of the synonym dictionary can merge some words with similar meanings, so that the dimension of the word vector is reduced.

3.3.3. Evaluation objects extraction

Ontology is the formal expression between concepts and relationships. In product reviews, the focus of reviews is generally to comment on the attributes of the product itself. For example, reviews of meals, dining environments, service attitudes, etc. often appear in restaurant reviews. The meals, the dining environment, and the service attitude belong to the restaurant's natural attributes. These three attributes together with other attributes such as the price of meals constitute the overall concept of the restaurant. A product feature is a product attribute that a user evaluates in a comment. Ontology attribute extraction is the core part of comment mining, including explicit product feature extraction and implicit product feature extraction. An explicit product feature means that the feature word is explicitly mentioned in the comment, and the feature is the explicit feature in the comment. For example: "This dish tastes really good." "Dish" is an explicit feature in the commentary. An implicit product feature means that the user does not directly mention the feature, but instead points to the feature by other words. For example, "this restaurant is a little expensive." At this point, commentators generally do not intend to buy this restaurant, but are evaluating the price of dishes. And we can deduce from "expensive" that we are talking about the characteristics of "dish price". Therefore, the price of dishes is the implicit feature in the reviews.

Implicit feature extraction is more difficult and less research results. But implicit features also have a major impact on sentiment analysis. Lu[18] uses semantic grammar to describe texts containing attribute knowledge and deeply parse sentences to achieve syntactic and semantic analysis. That is, the pattern matching method is used to extract the implicit features. Zhang[19] used the synonym Cilin to expand the corpus that contains only explicit attributes and generate a common word dictionary so that multi-strateential implicit feature extraction algorithms are used to extract implicit features based on dictionaries. However, some common words can be matched with many features, resulting in inability to identify features and reduce accuracy. And lack of corpus can lead to inaccurate results. Tang[20] used the improved TF-IDF method to calculate the fixed collocation relationship between emotional words and feature words for microblog product reviews, and used this relationship to extract implicit product features. Tao[21] analyzes the components of implicit sentences and calculates the characteristic clusters for each case of whether the sentences contain emotional words, verbs, and names. The representative words with the highest probability are used as feature words. Compared with literature[22-23], it has more advantages in performance. The same words are different in different contexts. For example, “high” is in derogatory sense when describing “price” and it is in complimentary sense when describing “price/performance ratio”. Therefore, one of the next research directions is to study the emotional expression in different situations.

3.3.4. Emotional tendency calculation

The basic idea of emotional tendency calculation is: according to the emotional word and emotional modifier weight and sentence weight, the emotional score of the emotional object is calculated according to the emotional dictionary. The initial value of the calculation value of most researchers set the negative word as -1, the default is 1, and the emotional strength of degree adverbs ranges from -0.5 to 2.0. According to the effect of a particular sentence pattern on sentence meaning, different calculation weights are set to calculate the emotion score. For example, Cai[24] used the the weight of emotional modifiers and then calculating the emotional polarity values of the viewpoints. Huang[25] proposed the lexicon weights, sentence weights, and modified weights. The combined method makes the calculation result more accurate and reliable. The accuracy of the method based on the sentiment dictionary is limited to the quality of the corpus. Therefore, the future research direction lies in the development of higher-quality corpus, multi-domain sentiment dictionary generation methods, dictionary expansion methods, and the improvement of cross-language sentiment analysis methods.

4. Conclusions and future work

As network information infiltrates into everyday life, it is important for individuals, businesses, and governments to grasp the direction of public opinion. Fine-grained text analysis as a new type of technology is also continuing its rapid development. This paper summarizes the development trends and research hotspots in this field by discussing the research methods and latest developments of Chinese fine-grained sentiment analysis in recent years. The best method is not a single model or algorithm, but a combination of multiple algorithms and dictionaries. At the same time, the expansion of the emotional dictionary is also imperative. Future research directions include cross-domain sentiment analysis, ambiguity’s solution of different domains’ semantic, and implicit emotional object extraction.

Text sentiment analysis can be combined with other technologies to achieve the real-time monitoring of the trend of network public opinion. For example, it can dominant company's brand negative monitoring of government information. The system automatically and timely finds information that may damage the reputation of the company or the government from the massive data. Therefore, the public relations treatment can be carried out in time. In the product reviews word

2018@ The American Computational Science Society. All rights reserved.

of mouth analysis, market research data analysis when it comes to the feedback from consumers on product services. The design and processes can be optimized by extracting from millions of user feedback. In financial investment, text sentiment analysis mainly includes the following points: financial news sentiment analysis, analyst research sentiment analysis, public stock sentiment analysis. Emotion influences investment and through emotional analysis of textual information from different sources, strategies can be developed by acquiring the emotional distribution of different subjects in the market.

References

[1] T. Sun, G. Geng, M. Zhou. Journal of Zhengzhou University (SCIENCE EDITION), 04 (2008) 48-51.
[2] Z. Chen, Z. Wang. Intelligence Science, 05 (2012) 759-763.
[3] Y. Guo, D. Pan. Decision and decision support system, 3 (1992) 56-62.
[4] W. Ma, Y. Deng. Computer application, 08 (2013) 2280-2282+2292.
[5] R. Liu, M. Nian, Z. Fan. Application of computer system, 10 (2017) 144-149.
[6] Y. Su, Y. Hu, B. Hu, X. Tu. Computer application, 06 (2016) 1613-1618.
[7] F. Zhang. Business economy research, 24 (2017) 55-57.
[8] N. Fan, W. Cai, Y. Zhao. Journal of Huazhong University of Science and Technology (NATURAL SCIENCE EDITION), 01 (2010) 31-34.
[9] C.H. Chang, C.J. LinACM Transactions on Intelligent Systems and Technology, 27 (2011) 1-27.
[10] W. Ding. Xi'an University of post and Telecommunications (2017).
[11] B. Chen, Z. Hao, R. Cai, W. Wen, I. Wang, H. Huang, X. Cai. Journal of Guangdong University of Technology, 03 (2017) 8-14.
[12] H. Jiang. Zhejiang University, (2017).
[13] Q. Ye, Z. Zhang, Z. Luo. Journal of information system, 01 (2007) 79-91.
[14] Y. Fang, D. Chen, Y. Wu. Computer application research, 07 (2014) 2035-2039.
[15] G. Huang, X. Zhou. Computer Engineering, 11 (2014) 211-214.
[16] X. Cai, P. Liu, Z. Wang. Journal of Zhengzhou University (SCIENCE EDITION), 02 (2017) 49-54.
[17] X. Tang, Y. Liu. Information theory and Practice, 06 (2018) 124-129.
[18] http://kns.cnki.net/kcms/detail/11.1762.G3.20180315.1523.004.html
[19] Y. Lu. jiangsu university of science and technology, (2016).
[20] L. Zhang, X. Xu. Modern library and information technology, 12 (2015) 42-47.
[21] X. Tang, Y. Lan. Library and information work, 16 (2016) 121-127.
[22] Y. Tao. Chongqing University of Posts and Telecommunications, (2017).
[23] Z. Hai, K. Chang, J. Kim. International Conference on Computational Linguistic sand Intelligent Text Processing. Berlin, Heidelberg: Springer-Verlag, (2011) 393-404.
[24] X. Cai. Shandong Normal University, (2017).
[25] G. Huang, X. Zhou. Computer application and software, 04 (2015) 239-242.

2018@ The American Computational Science Society. All rights reserved.