ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202209_5(9).0004

A Review of Text Sentiment Analysis

Weizhi Zhang, Xiuxia Li and Wei Liu

Qufu Normal University, Rizhao, Shandong, 276800, China

Abstract

[Objective] To collect and sort out the data of Chinese and foreign literature on sentiment analysis, summarize the relevant methods and technologies, research content, discipline fields, application scenarios and development trends of sentiment analysis, and present them in a visual way. [Methods] In this paper, journal papers in CNKI and Web of Science core databases were used as literature sources, relevant concepts and methods of text sentiment analysis were used to construct retrieval formulas, and quantitative statistics were made for research papers on text sentiment analysis methods from 2011 to 2021. It also summarizes, analyzes and summarizes the main techniques, methods and application scenarios of text sentiment analysis from multiple dimensions such as time, keywords and disciplines, and discusses its current situation and shortcomings on this basis.

Keywords

Text emotion analysis method technology application scenario development research.

1. The Introduction

Sentiment Analysis, also known as opinion mining and Sentiment orientation Analysis, refers to the process of analyzing, processing and extracting subjective texts with emotions by using natural language processing technology and text mining technology. This concept was proposed by Nasukawa[1]et al in 2003. Pang[2]et al and Turney[3]et al realized emotion classification from supervised learning and unsupervised learning respectively. Wiebe[4][5]et al realized the extraction of subjective information from text.

According to the granularity of text processing, sentiment analysis can be roughly divided into three research levels: discourse level, sentence level and word level. In the early stage of sentiment analysis, Pang et al. And Turney mostly conducted sentiment analysis at the level of discourse. Discourse level sentiment analysis assumes that each sentiment text represents an object independently. Sentence level and word level are of a more detailed level, which judge the emotion expressed by a sentence or word, and constitute the emotional tendency of the whole text together with these more fine-grained sentences.

In order to comprehensively analyze literatures related to sentiment analysis, CNKI and Web of Science core databases were used as literature sources to construct retrieval formulas based on related concepts and methods of sentiment analysis, and to conduct quantitative statistics on research papers on text sentiment analysis methods from 2011 to 2021. This paper summarizes, analyzes and summarizes the main models, methods and application scenarios of text sentiment analysis from the dimensions of time, keywords and disciplines, and discusses the current situation and shortcomings on this basis.

2. Data Sources and Analysis Methods

In this paper, CNKI and Web of Science core database are used as retrieval platforms. Since the literature in the Web of Science core database only dates back to 2010 at the earliest, the retrieval time of both platforms is set from 2011 to 2021. First, construct the retrieval formula

TS=(sentiment analysis* OR sentiment classification OR opinion mining)AND(Lexicon OR rule OR model OR Algorithm), and the retrieval results were filtered by Web of Science core collection and English language, and 7209 English literatures with high correlation were obtained. After removing irrelevant literatures, 7156 literatures were obtained, which were exported as plain text files, full records and references. Advanced retrieval was selected in CNKI, subject retrieval was carried out with SU=' sentiment analysis 'OR' sentiment classification 'OR' opinion mining 'as the retrieval type. 17,463 Chinese literatures were also selected after screening and exported to Refworks format. Based on the above exported literature data, 59 articles were selected as extensions after reading relevant literatures through references. The results were imported into Citespace5.7.R5 to construct a data set, and then the data were analyzed by means of co-occurrence analysis between national institutions and authors, co-cited literature analysis, keyword co-occurrence analysis and other methods, and the conclusions were drawn.

3. Literature Analysis

3.1. Analysis of Posting Trend

Since 2011, the research in the field of sentiment analysis has been increasing year by year, with the most obvious growth in 2015-2016 and 2018-2019 compared with the previous year. In 2020, the most relevant research results are 2,898, and in 2021 (as of December 31), 1,369, the same as that in 2016. The number of publications of Web of Science core database is roughly the same as that of CNKI, with the most obvious growth period from 2015 to 2016 and from 2018 to 2019. In mid-2008, the concept of big data was put forward for the first time, which also promoted the research on sentiment analysis. The explosive growth of data greatly promoted the research on sentiment dictionary and machine learning methods in sentiment analysis, and the attention index of sentiment analysis began to rise sharply in this year. In 2013, after the release of Word2Vec, sentiment analysis methods based on deep learning began to be widely used in the field of natural language processing, and sentiment analysis entered a stage of rapid development.

3.2. Analysis of Discipline Distribution

From the CNKI data, the top 20 disciplines with the number of published papers are selected. In terms of their discipline distribution, they are mainly computer software and computer application, automation technology and other disciplines. Among them, computer software and computer application are the most concentrated with 6,642 papers, almost equal to the sum of the other 19 disciplines. The publications in the core database of Web of Science involve 25 disciplines, and the ranking of discipline distribution is shown in Figure 4. Different from the distribution of domestic disciplines, the ranking of library and information has risen from 11th to fourth, but the research is also concentrated in the field of computer Science, followed by enterprise economics and engineering.

3.3. Analysis of Sentiment Analysis Research Content

Keywords co-occurrence analysis is carried out in CiteSpace based on the exported literature data, through which the main technologies, methods and application scenarios in the field of sentiment analysis can be identified. In order to make the co-occurrence map more accurate, keywords directly related to sentiment analysis are deleted, such as sentiment analysis, opinion mining, sentiment analysis, opinion mining, text classification, etc., as well as other interfering words. Such as Weibo, online comments, social media, Social Medical, Model, impact, etc. The synonyms are combined, such as Sina Weibo, Weibo, SVM and support vector machine, LDA, LDA model and LDA theme model, emotion dictionary and polarity dictionary, etc.

Through the analysis, it can be seen that the main methods and technologies applied in the field of emotion analysis are: method based on emotion dictionary, method based on machine learning and method based on deep learning. There are many foreign researches on machine learning and deep learning methods, while domestic researches tend to emotion dictionary method, followed by deep learning method.

The core of the method based on emotion dictionary is to construct a marked emotion dictionary or domain ontology, analyze the semantic association between text units through dependency syntax, and judge the semantic similarity by using semantic rules, so as to complete the judgment and analysis of text emotion tendency. The methods based on machine learning mainly include Naive Bayes algorithm (NBM), support vector machine (SVM), LDA topic model, conditional random field (CRF), K-means, etc. The methods based on deep learning mainly include convolutional neural network, long and short term memory network, attention mechanism, recurrent neural network and two-way long and short term memory network. Word2Vec, GloVe and BERT and other pre-training models also improved the effect of emotion classification..

3.4. Application of Sentiment Analysis

Both at home and abroad, the application objects of sentiment analysis are all focused on social media, which tend to analyze and predict user emotions, monitor and guide online public opinion, and create a good online community environment. At present, sentiment analysis has made some achievements in various fields, such as politics, public health, business and so on. Add time line to the analysis result, get the development context of application and technical method of sentiment analysis.

Sentiment analysis was initially applied to social media to help government departments understand people's views and attitudes towards policies, obtain feedback and adjust relevant decisions to adapt to economic and social development. The most prominent application was the collection of public emotions in the 2008-2009 financial crisis and opinions on presidential and government elections. And to predict election results by analyzing political sentiment expressed in tweets and evaluating candidates. The initial application method of sentiment analysis is to obtain viewpoints and contents from social media platforms on the Internet, and to extract them manually, which has a large workload and low efficiency. In this case, it is extremely unrealistic to obtain useful information, so automatic information extraction tools are spawned. This, coupled with the growth of the Internet and social media, has greatly promoted the expansion of sentiment analysis applications.

In the business world, sentiment analysis of information from blogs, microblogs and forums has been widely used in building brand image, tracking customer feedback, and automated dialogue systems to deal with customer needs and complaints. The Pulse system researched and developed by Gamon et al. can extract users' opinions on product details from a large amount of text data by using text clustering technology, and mine the polarity and intensity of emotions in automobile evaluations. Gupta et al. proposed a sentiment analysis method for customer complaint emails to better manage customer relationship. Bollen et al. obtained the sentiment index by emotion analysis of all the Twitter information in a period of time, and found that the "CLAM" sentiment index was surprisingly consistent with the Dow Jones Industrial Average (DIJA) after the "CLAM" sentiment index moved three days later. Therefore, they proposed that the stock market could be predicted by twitter sentiment analysis. Mishne et al first applied the sentiment analysis method into the prediction of film box office in 2006 and achieved good results, which was widely recognized and promoted the follow-up research.

In the field of public health, Cherry et al. developed a system to prevent suicide through emotion analysis. Chen et al. used emotion analysis to identify online violence. Opinion Parser developed by Bing Liu covers more than 40 fields, including catering, medical, lifestyle products, film and television, computer, finance, and politics. Today, individuals, factories, enterprises, institutions and government departments are using information from the media to help them make decisions.

Domestic sentiment analysis applications are more diversified. In addition to weibo, it is also popular to find out consumers' preferences and provide personalized push by analyzing product reviews on e-commerce platforms, douban book reviews and movie reviews. Enterprises can adjust production and sales strategies accordingly to obtain the maximum economic and social benefits. For example, in 2006, yao Tianfang's opinions on automobile reviews were mined, and the average accuracy rate reached 60%. Zhang then applied sentiment analysis to business in 2007 to help consumers make better purchasing decisions. Around 2010, the application of sentiment analysis expanded to film reviews, book reviews and other fields; Taobao cashed in on Singles Day in 2012, the same year sentiment analysis research began involving e-commerce platforms. Due to the epidemic since 2019, COVID-19 and public health events have appeared with high frequency in the past three years and become the focus topic of major platforms. Sentiment analysis method has been applied in the medical and public health fields on a large scale. 2.5 Sentiment analysis tools and techniques

In the data acquisition stage of sentiment analysis, the main method is crawler tool. Now most network platforms have set up anti-crawling mechanism, but the anti-crawling mechanism adopted by each major platform is different. Therefore, in line with the principle of simplicity and efficiency, different crawler tools will be used when acquiring data from different platforms. For example, douban, Weibo and CNKI with relatively simple anti-crawling mechanism can use crawler applications such as Octopus, Houyi and Spiderman2 with relatively simple operation, and can also use Python and Java to write crawler programs. Platforms such as Dianping, Taobao and Jingdong with complex anti-crawling mechanisms can only be crawled by Python and Java crawlers. Special coding is also required to reduce the crawl frequency. The main methods used are Scrapy, Pillow image processing, Selenium and Session verification code, fontTools font mapping processing, JavaScript decryption, etc. Twitter is a platform that allows crawler to crawl. The anti-crawl mechanism is relatively simple, but it needs to apply for the corresponding API and is only allowed to crawl the data within a week. If more data is needed, it needs to write a very complex crawler program to bypass the API. In addition, you can also use the WebScrape plugin from Google Chrome when you're crawling into your ecommerce data.

In the stage of data preprocessing, the application methods at home and abroad are roughly the same. In the stage of stopping words and word segmentation, English text only needs to identify some fixed phrases with a relatively small amount of work. NLTK tool is mainly called by Python. There is no clear boundary between words in Chinese texts, and they are more complex and changeable. Therefore, professional dictionaries need to be built on the basis of general dictionaries for different application fields. Word segmentation tools are also more diversified. At present, the most commonly used Chinese word segmentation is Jieba word segmentation, followed by HanLP, which are mostly called through Python. You can also use Pangu word segmentation software and online word segmentation tools, such as webmaster tools; In addition, there are researchers to achieve the whole process of emotion analysis in the form of outsourcing, such as Ali Cloud NLP, Baidu NLP and so on. Data cleaning stage, there are mainly manual cleaning and automatic cleaning two methods, manual cleaning accuracy is high, but the speed is slow, low efficiency; The automatic cleaning can be implemented using the Python tool, such as Pandas, or application software, such as dataWrangle.

In the final stage of emotion analysis, the methods are more diversified. Emotion analysis methods of microblog data are mainly realized through emotion dictionary, the main methods include LDA topic clustering, TF-IDF word frequency analysis, and the use of ROST.EA software or SnowNLP algorithm for emotion analysis. The dictionaries used mainly include Hashong

Lexicon and Dalian Institute of Technology lexicon. Sentiment analysis methods based on e-commerce data mainly involve Word2vec model, SVM for text classification, and CNN based on deep learning for sentiment analysis. Sentiment analysis methods based on Twitter data mainly classify text by Word2vec model or naive Bayes algorithm plus decision tree, and conduct sentiment analysis by machine learning plus dictionary. Some researchers also use TextBlob for sentiment analysis, and achieve good results with an accuracy of 71-79%.

Artificial intelligence (AI) was first proposed in 1956 and has received a lot of attention and attention. At the same time, domestic computer technology is still underdeveloped, let alone artificial intelligence technology. Therefore, in the development process of foreign sentiment analysis research methods, since the birth of sentiment analysis research, methods based on machine learning and deep learning have occupied the main position, and a series of corresponding algorithms and technologies have been born. Such as LDA algorithm, TF-IDF algorithm, Word2vec model, support vector machine (SVM), convolutional neural network (CNN) and so on. Among them, machine learning has the earliest birth and the most applications, while deep learning can be regarded as a branch and deep application of machine learning, appearing relatively late and its development has not reached the level of machine learning. Compared with the above two methods, there is less mention of sentiment dictionary-based method in foreign sentiment analysis research, so its development and utilization is not so prominent. However, the sentiment analysis research method based on sentiment dictionary appeared first in China and still plays an important role throughout the development of sentiment analysis research. The method based on deep learning is a hot topic in the research of sentiment analysis method, and has a certain connection with the method based on emotion dictionary. The research on machine learning is relatively less popular, which may be related to the late start of this kind of research in China and the existence of negative voice of machine learning algorithm. Deep learning algorithms can complete more complex tasks with higher accuracy, but they need to be built on the basis of massive data, so the threshold is relatively high. In contrast, traditional machine learning algorithms are unnecessary, simpler, easier to implement, and perfectly capable of performing some everyday tasks.

4. Conclusion

This paper is based on cnKI and Web of Science core database to retrieve related papers on sentiment analysis, and uses CiteSpace to analyze the obtained data. This paper takes the three main methods of sentiment analysis as the main line, and expounds and analyzes the techniques, methods, applications and development process of sentiment analysis research. As can be seen from the development trend of sentiment analysis research, with the development of big data, the research on sentiment analysis has made significant progress in technology, methods and applications in the past decade, and shows a rapid growth trend. A large number of scholars and enterprises have flooded in, giving birth to a large number of emotion analysis systems and technologies, providing more and more high-quality services; At the same time, with the in-depth development of sentiment analysis, whether individuals, institutions or government departments, the application of sentiment analysis is becoming more and more strong demand; The research and demand of sentiment analysis promote each other, which makes this research field full of vitality. The existing technical conditions provide a great space for the development of sentiment analysis research. In addition, the improvement of emotion analysis techniques is also a research focus in this field.

There is still a lot of room for development of emotion analysis methods based on machine learning. With the development of various social media platforms, a variety of new online words emerge one after another. Therefore, the expansion and reconstruction of emotion dictionary is also very important. Finally, current sentiment analysis studies mostly focus on

Chinese and English languages. In recent years, with the emergence of multilingual natural language processing, sentiment analysis studies on minority languages and minor languages have begun to emerge, which is bound to become another focus of sentiment analysis research.

References

[1] Nasukawa, Tetsuya and Jeonghee Yi. Sentiment analysis: Capturing favorability using natural language processing. in Proceedings of the K- CAP-03, 2nd Intl. Conf. on Knowledge Capture.2003.

[2] Pang B, Lee L, Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques [C]//Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002: 79-86.

[3] Turney P D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews[C]//Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002: 417-424.

[4] Wiebe, Janyce. Learning subjective adjectives from corpora. in Proceedings of National Conf. on Artificial Intelligence (AAAI-2000). 2000.

[5] Wiebe, Janyce, Rebecca F. Bruce, and Thomas P. O'Hara. Development and use of a gold-standard data set for subjectivity classifications. in Proceedings of the Association for Computational Linguistics (ACL-1999). 1999.

[6] Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012). Detecting offensive language in social media to protect adolescent online safety. In Privacy, Security, Risk and Trust (PASSAT), 2012, International Conference on and 2012 International Confernece on Social Computing (SocialCom), pp. 71–80. IEEE.