Computer Life (CPL) ISSN: 1819-4818
DELIVERING QUALITY SCIENCE TO THE WORLD

Summary and Prospect of Text Sentiment Analysis

Duan Minmin [a], Li Zhen [b]

Faculty of Economics and Management, Xidian University, Xi'an,China

[a]dmin79856@163.com,[b]15114840972@163.com

Abstract: In recent years, with the development of the Internet and social networks, a large number of subjective texts have appeared on the Internet, such as comments published by users on social networks such as blogs and Weibo, which contain a large amount of sentiment information and subjective opinions. Effectively mining information of such texts has important practical value for e-commerce, information prediction, and public opinion monitoring. At present, sentiment analysis has become a research hotspot in the field of natural language processing. This paper describes the technical methods and research applications of text sentiment analysis, summarizes the research difficulties in current text sentiment analysis and looks into possible future research hotspots.
Keywords: text sentiment analysis, dictionary construction, machine learning.

---

1. INTRODUCTION

Text sentiment analysis, also known as opinion mining, focuses on determining the attitude of a speaker or author on a particular topic, that is, the user's perception of a topic or the mining of the review text, so that the view or comment is a right Positive or negative opinions of the matter. Simple face-to-face is the process of analyzing, processing, summarizing and reasoning subjective texts with sentiment color [1]. According to the fine-grained difference of research objects, sentiment analysis can be divided into word level, phrase level and sentence level. Several levels of research, such as chapter level. This is a multidisciplinary cross-disciplinary field involving computational linguistics, artificial intelligence, machine learning, information retrieval and data mining. The text sentiment analysis has important academics research value.

2. CLASSIFICATION OF TEXT SENTIMENT ANALYSIS METHODS

Text sentiment classification is a special text classification problem. Text categorization refers to the process of determining the attribution of a text according to a predefined category, and the sentiment classification is mainly used to discriminate between opinions expressed in natural language words, preferences, and information related to feelings and attitudes. The research on sentiment classification in recent days mainly includes two methods based on dictionary and machine learning. Among them, the dictionary-based technology can be divided into two categories: artificially constructing sentiment dictionary and automatically constructing sentiment dictionary according to different degrees of artificial participation; based on machine learning technology, according to different sentiment

classification methods, it can be divided into naive Bayesian based methods. SVM-based methods, neural network-based methods, decision tree-based methods, and logistic regression-based methods.

2.1 Dictionary based technology

Sentiment dictionary is the basis of text sentiment analysis. Using the constructed text sentiment dictionary, and the polarity and intensity of the sentiment dictionary, and then the text sentiment classification, can effectively analyze the text sentiment. Based on dictionary-based sentiment analysis, building a sentiment dictionary is the key. According to the different degrees of participation in the process of sentiment dictionary construction, it is divided into artificially constructed sentiment dictionary and automatically constructed sentiment dictionary.

2.1.1 Artificial construction of sentiment dictionary

The way of constructing sentiment dictionary by hand is mainly to use a large number of existing sentiment resources to expand the annotation of the sentiment resources summed up by the predecessors, and then form a variety of basic sentiment dictionaries. The advantage is that it facilitates the creation of richer entry information and is easy to control.

Among them, Xu Linhong et al. [2] of Dalian University of Technology constructed the sentiment vocabulary ontology through two methods: manual sentiment classification and automatic strength acquisition. The manual sentiment classification mainly adopts the conversion-driven error-driven learning method. The automatic acquisition of the vocabulary's sentiment intensity uses the Pointwise Mutual Information (PMI) method.

Wang Yong et al [3] constructed a polar dictionary of Chinese microblogs in order to classify Chinese Weibo. Randomly crawling 100,000 microblogs on each microblog website, through multiple manual annotations and proofreading, distinguishing 2199 sentiment words from positive and negative, strong and weak. According to the diversity of microblog expressions, emoji is also constructed. Dictionary, negative dictionary and double negative dictionary.

The currently constructed sentiment dictionary mainly include: General Inquirer Lexicon of Harvard University, Opinion Finder Subjective Emotion Dictionary provided by the University of Pittsburgh, dictionary resources provided by Bing Liu of the University of Illinois, English language dictionary of WordNet built by Princeton University, Chinese emotion of Taiwan University Polar Dictionary (NTUSD), HowNet Sentiment Dictionary HowNet, etc.

The artificial construction of sentiment lexicon has certain advantages in expanding the vocabulary information and handling convenience, but greatly increases the labor overhead, and the scope of expansion is limited, so it is not suitable for cross-domain research. In recent years, the method of automatically constructing an sentiment dictionary has gradually become a research direction.

2.1.2 Automatically build sentiment dictionaries

By automatically constructing sentiment dictionaries, we can reduce labor costs well and enhance domain applicability to a certain extent, so in recent years researchers have been more committed to the automatic construction of sentiment dictionaries. The methods of automatically constructing sentiment dictionaries are mainly based on knowledge base method, corpus-based method and the method of combining knowledge base with corpus.

The method based on knowledge base mainly makes an sentiment dictionary by semantic analysis or extension of the existing knowledge base (such as English WordNetand and Chinese HowNet) to

judge the sentiment tendency of unknown text information. such as the expansion of the WordNet , the addition of nouns, verbs and adverbs, so that the sentiment dictionary more comprehensive.

The corpus-based method mainly obtains the sentiment dictionary by automatically learning from a large amount of corpus, and by extracting the corpus of different fields, the sentiment dictionary of the specific domain can be obtained. For example, Hatzivassiloglou and McKeown [4] proposed a method for automatically retrieving semantic orientation information from indirect information collected from large corpora. The method relies on the corpus, achieves high precision (more than 90%), considers the dependence of sentiment words or phrases and feature word domains, and automatically adapts to the new domain when the corpus changes, and can be directly applied to other word classes. Turney et al. [5] used the PMI method to extend the basic positive and negative vocabulary, and then used the semantic polarity (ISA) algorithm to analyze the sentiment text, and the accuracy rate of the general corpus data was 74%. Considering user behavior, Yang et al. [6] improved the SO-PMI algorithm by using Laplacian smoothing technique and established a Chinese hotel commentary sentiment dictionary.

In recent years, due to the rapid growth of text information in the Internet, some online words have emerged, making it impossible to meet the requirements of existing text information by simply using the original knowledge base or the corpus in the Internet to construct an sentiment dictionary. Therefore, researchers are more inclined to use knowledge. The method of combining library and corpus constructs an sentiment dictionary. By incorporating the extended sentiment knowledge base and the sentiment vocabulary extracted from the corpus into the sentiment dictionary, the constructed sentiment dictionary is enriched. Yang Xiaoping [7] used the Word2Vec tool to train a set of word vectors from a massive corpus, and through the comprehensive screening of the NTUSD dictionary, the HowNet Sentiment Dictionary and the sentiment ontology library developed by Dalian University of Technology, to form the SentiRuc dictionary, through the machine Learning to disambiguate the sentiment color, and synonymous relationship optimization, antisense relationship optimization and sentence-level description force optimization of the dictionary, and achieved good experimental results in the general field dataset.

Text sentiment analysis technique based on dictionary because the dictionary is often only aimed at a certain field, the effect of cross-domain sentiment analysis is not good enough, and the sentiment words in the dictionary may not be rich enough, for short text and specific domain text for sentiment analysis of the effect is better. Therefore, for long text, a better solution is to take advantage of machine learning methods.

2.2 Machine learning methods

Based on machine learning-based sentiment classification, the key lies in feature selection, feature weight weighting, classification model and other three elements. Feature selection is mainly based on information gain, chi-square statistics and document frequency. Common feature weighting methods include: Boolean weight, word frequency (TF), inverted document frequency (IDN), TF-IDF, TFC, entropy weight, and so on. The classifier models include: naive Bayes, support vector machines, K-nearest neighbors, neural networks, decisive trees, logistic regression, and so on.

Pang et al. [8] used Naive Bayes and Support Vector Machines to compare text sentiment analysis in 2002 and found that using SVM for text sentiment analysis can achieve optimal results. The following

summarizes the research results of text sentiment analysis based on naive Bayes and SVM in recent years.

2.2.1 Based on the Naive Bayesian method

Naive Bayes is a probabilistic model that works satisfactorily in many fields. Bayesian classification provides practical learning algorithms and prior knowledge, and the observed data can be combined. In Naive Bayesian technology, the basic idea is to find the probability of a given text document category by using the joint probability of words and categories. This algorithm is widely used in text sentiment analysis.

Govindarajan et al. [9] proposed a new hybrid classification method based on Naive Bayes (NB) and Genetic Algorithm (GA), and compared the effectiveness of the sentiment classification synthesis technique, and conducted sentiment analysis on widely used movie reviews. , proved the feasibility of the method. Text sentiment analysis based on naive Bayesian algorithm can be applied to many fields. Soelistio et al. [10] proposed a simple model for analyzing the sentiment polarity of digital newspapers using naive Bayesian classification, which is applied to digital newspapers. Political sentiment analysis, obtaining positive or negative sentiment information about a particular politician from a digital news article. Wikarsa et al. [11] studied an application of sentiment classification of Twitter users using the naive Bayesian method. Dey et al. [12] used Naive Bayesian algorithm and KNN algorithm to analyze the emotions of movie reviews and hotel reviews, and found that Naive Bayes is better than KNN in movie reviews, but in the hotel reviews, there is little difference in accuracy between them.

Naive Bayesian-based text sentiment analysis technology classifies text emotions by calculating probabilities, which is suitable for incremental training, and the algorithm is relatively simple and performs well on small-scale data. However, this method is sensitive to the expression of the input data, and needs to calculate the prior probability, so there will be an error rate in the classification decision.

2.2.2 Method based on support vector machine

The Support Vector Machine (SVM) was originally proposed by Vapnik [13] and is a relatively new machine learning method. It seeks to minimize the structural risk to improve the generalization ability of the learning machine, and to minimize the empirical risk and confidence range, so as to achieve good statistical rules in the case of less statistical sample size. The following is an introduction to text sentiment analysis by researchers based on SVM in recent years.

Sharma and Dey [14] proposed a mixed sentiment analysis model based on Boosted SVM in 2013. The model uses two techniques, Boosting [15] and SVM, to analyze the sentiment analysis of 2000 movies and hotel review corpora. The results show that the SVM mixed sentiment analysis model based on Boosting algorithm has better performance than the single SVM model.

Hajmohammadi [16] uses standard machine learning techniques SVM and naive Bayes to automatically classify Persian language film reviews into positive and negative, and finds that SVM classifiers achieve higher accuracy than Naïve Bayes in Persian language film reviews. In 2015, Karanasou et al. [17] conducted an sentiment analysis of the metaphors in Twitter, using grammatical and morphological features, annotating the sentiment polarity in metaphorical and non-figurative tweets, and using structured knowledge resources, such as the SentiWordNet sentiment dictionary.

Assign sentiment scores to words and WordNet and calculate word similarity. This experiment achieved the best results with an SVM classifier with a linear kernel function. Based on the characteristics of the financial sector, Haung et al. [18] used the SVM classification method combined with Stanford language dependence to analyze the sentiment generated by users in the financial sector. The text sentiment analysis method based on SVM is considered to be the best sentiment analysis method, which has low generalization error rate, little computation overhead, and can get good sentiment analysis effect for the smaller text of the training sample, and the processing effect of high dimensional data is good, and the low error rate can be obtained. However, this method is sensitive to parameter adjustment and kernel function selection.

3. SENTIMENT ANALYSIS APPLICATION RESEARCH

Natural language processing now has a wide range of applications in information retrieval, social networking, public opinion monitoring, speech recognition, machine translation, and recommendation systems. The following is an introduction to product reviews, public opinion analysis, and information prediction:

Commodity comment analysis, which is one of the most frequently used application points of sentiment analysis technology. Nowadays, e-commerce is developing very fast. More and more people like to shop online, so the number of texts with subjective color product reviews is rapidly increasing. Increase, which contains a large number of user-valued information with commercial value, and extracts the characteristics or attributes of products by using the results of mining and analyzing the subjective comments on the Internet. Consumers can understand people's attitudes toward a certain product and optimize purchasing decisions. Producers and sellers can understand the feedback of consumers on their goods and services, as well as consumers' evaluation of themselves and their competitors. Products, improve services, and gain competitive advantage.

Internet public opinion analysis, public opinion analysis is mainly to analyze the public's views on hot events or news events. The most representative public opinion platform is blog and Weibo. As users are heavily involved in the generation of information, more and more personal content appears on online media such as blogs and forums. These online expressions are aimed at understanding the public's overall situation of news people and news events. Evaluation, mastering the current public opinion information, especially the public opinion information of hot events has an important role. The current direct impact of online public opinion on society is increasing, which is directly related to the information security of the network. Therefore, social managers should promptly give feedback on these grievances. However, it is difficult to deal with the huge amount of information appearing in the network through human means, so the application of automated sentiment analysis technology in this field is very practical.

Information Prediction , with the vigorous development of the Internet , the impact of network information on people's lives has become more and more important. The occurrence of a new event or the hot discussion of an event on the internet has largely shaped people's thinking and actions. For example, when the president or a member of the parliament is abroad, many candidates hope to predict whether they can be elected by summarizing the voters' online comments, so information forecasting

becomes necessary. Sentiment analysis techniques can help users predict the future of an event by sorting out information sources such as news, posts, and so on the internet.

4. SUMMARY AND OUTLOOK

The above is an introduction to the research results obtained in the field of text sentiment analysis in recent years. Although there have been some successful applications, there are more challenges. Here are some of the difficulties and possible hotspots of text sentiment analysis in the coming years:

(1) How to deal with the diversification of the growing sentiment research objects and the complexity of sentiment tasks. With the continuous expansion of the application field, sentiment objects from the tendency to comment on products, services, etc., to the classification of users and topical emotions in social media, the forms of expression are more diverse, the types of emotions are more diverse, and the content of research will also be corresponding. The transition includes more attention to the user's information and changes in the emotions of the event users in social media. Small sample and migration learning in the field of machine learning today may be the research direction to solve this problem.

(2) How to deal with the complexity of sentiment expressions in the text. People's expressions of emotions in texts, especially in short texts, are diversified, expressed in straightforward expressions or expressed in various forms such as rhetorical techniques and even irony. Neural network learning techniques that apply deeper or more complex networks, coupled with the built-in pool of sentiment common sense, may be the way to solve this problem.

(3) Cooperating with cognitive science researchers from the perspective of cognitive science, by comparing the human brain wave waveforms with the detected brain waveforms and reflections of various emotions when reading texts, as a text sentiment analysis The scientific basis will also be a possible research direction.

REFERENCES

[1] Zhao Wei, Qin Bing, Liu Ting. “Text Emotion Analysis”, Journal of Software, 2010, Vol. 21(8), p1834-1848

[2] Xu Linhong, Lin Hongfei, Pan Yu, et al. “The Construction of Sentiment Vocabulary Ontology”, Journal of the China Society for Scientific Intelligence, 2008, Vol. 27(2), p180-185

[3] Wang Yong, Lu Xueqiang, Ji Lianchun, et al. “Chinese microblogging sentiment classification based on polarity dictionary”. Journal of Computer Applications and Software, 2014(1), p34-37

[4] Hatzivassiloglou, V. "Predicting the semantic orientation of adjectives." Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and the Eighth Conference of the European Chapter of the Association for Computational Linguistics, 1997 1997

[5] Turney P D , Littman M L . “Measuring praise and criticism”, ACM Transactions on Information Systems, 2003, Vol. 21(4), p315-346

[6] Yang A M , Lin J H , Zhou Y M , et al. “Research on Building a Chinese Sentiment Lexicon Based on SO-PMI”. Applied Mechanics and Materials, 2012, Vol.263-266, p1688-1693

[7] Yang Xiaoping, Zhang Zhongxia, Wang Liang, et al. “Automatic Construction and Optimization of Sentiment Dictionary Based on Word2Vec”, Computer Science, 2017(1)

[8] Pang B,Vaithyanathans S. “Sentiment Classification Using Machine Learning Techniques”. Conference on Empirical Methods in Natural Language Processing,2002, Acl-02, p79-86

[9] Govindarajan M. “Sentiment analysis of movie reviews using hybrid method of Naive Bayes and genetic algorithm”. International Journal of Advanced Computer Research,2013, Vol.3(4), p139

[10] Soelistio Y E , Surendra M R S . “Simple Text Mining for Sentiment Analysis of Political Figure Using Naive Bayes Classifier Method”,Icts. 2015

[11] Wikarsa L , Thahir S N . “A text mining application of emotion classifications of Twitter's users using Naïve Bayes method”, International Conference on Wireless & Telematics. IEEE, 2016
[12] Dey L,Chakraborty S,Biswas,et al. “Sentiment Analysis of Review Datasets Using Naïve Bayes' and K-NN Classifier”. Information Retrieval,2016, Vol.8(4), p54-62
[13] Vapnik V . “SVM method of estimating density, conditional probability, and conditional density”, IEEE International Symposium on Circuits & Systems. IEEE, 2000
[14] Sharma A , Dey S . “A boosted SVM based sentiment analysis approach for online opinionated text”, Proceedings of the 2013 Research in Adaptive and Convergent Systems. ACM, 2013
[15] Michael J. Kearns， Leslie G. “Valiant. Cryptographic limitations on learning Boolean formulae and finite automata”,Journal of ACM,1994, Vol.41(1),p 433-444
[16] Hajmohammadi M S , Ibrahim R . “A SVM-based method for sentiment analysis in Persian language”, International Conference on Graphic & Image Processing. 2013
[17] Karanasou M, Doulkeridis C. “An svm-based approach for sentiment analysis of figurative language on twitter”, International Workshop on Semantic Evaluation.2015, 709-713
[18] Jin H , Tong R , Ruiquan J . “Sentiment analysis in financial domain based on SVM with dependency syntax”, Computer Engineering & Applications, 2015.