A Review About Sentiment Analysis of Short Texts Based on Machine Learning

Lisheng He(✉)

Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China
200111108@stu.hit.edu.cn

Abstract. In today's Internet era, the number of netizens has grown rapidly, which has spawned a large number of comments expressing their own opinions and views. How to extract valuable emotional information from these texts and make use of them is becoming more and more important. Therefore, text sentiment analysis comes into being pregnancy. As an vital branch in the field of natural language processing. It is universally used in network public opinion monitoring and analysis, semantic network analysis, knowledge graph, content recommendation, etc. It is worthy of in-depth research by scholars. Depending on the methods of use, It can be divided into three methods, which are based on sentiment dictionary, traditional machine learning and deep learning. Among them, the classification methods based on machine learning are used commonly. Through analysizing and studying this method, summarizing the advantages and disadvantages of the specific methods, and reviewing the research results of scholars, This paper puts forward some prospects for future research directions.

Keywords: text sentiment analysis · machine learning

1 Introduction

With the popularization of the Internet, an increasing number of netizens have become accustomed to expressing their views and opinions on public platform. Netizens can convey their opinions on various matters and express their emotional tendencies on Weibo, twitter and other public platforms. The value of text sentiment analysis is also reflected. In terms of public opinion analysis, through sentiment analysis of comments on hot events, it helps the government to continue to pay attention to the emotional attitudes of the masses or the trend of public opinion, so as to avoid events that have adverse social benefits; in terms of comment analysis on e-commerce websites, Combining the evaluation objects and extracting their comments, analyzing and identifying the emotional attitudes in the comments, on the one hand, it helps to correctly guide consumers to choose products better, and on the other hand, it can also promote the merchants to improve their products. Therefore, it is a significant task to study and effectively apply text sentiment analysis.

C. F. Peng et al. (Eds.): MMET 2022, ASSEHR 703, pp. 368–372, 2023.
https://doi.org/10.2991/978-2-494069-51-0_52

2 Introduction to Text Sentiment Analysis

Text sentiment analysis, which is also known as tendency analysis, etc. [1], refers to the analysis of subjective texts with emotional tendencies, mining the emotional attitudes contained in them, and dividing them. The process of text sentiment analysis is roughly divided into the following steps: original text acquisition, text preprocessing, feature extraction, sentiment classification using the model, and result output. The original text acquisition is generally obtained through web crawlers to obtain relevant information. Text preprocessing mainly refers to removing invalid characters, unifying English data, capitalizing English letters, serializing data, and performing word segmentation. This process is completed by specialized mature tools. of. The purpose of feature extraction is to express text into a data structure that can be processed by an algorithm. The sentiment classification output obtains the sentiment polarity of the text. Commonly used classifier methods include SVM, KNN, Naive Bayes, etc.

3 Introduction to Machine Learning Research Methods

Machine learning refers to training and evaluating models through collected data sets, to the extent that results can be more accurately predicted through models. So far, this method has got great achievements. It first performs feature processing on text information, and then conducts supervised learning and training of the model. The trained model is used to predict the sentiment polarity of the text information that needs to be concerned. Machine learning-based sentiment classification methods are mainly divided into three categories. In the supervised method, the data that has been manually labeled is usually used to obtain the mapping from input to output, and then this mapping relationship is applied to the unknown data to predict the unknown data. Common supervised methods are: KNN, Naive Bayes, SVM, maximum entropy, etc.

3.1 K-Nearest Neighbor Algorithm (KNN)

KNN classification algorithm is a lazy learning method, which does not need to establish a classification model in advance, but simply stores training examples [1]. The idea of this method is to determine the category to which the required classification samples belong based only on the categories of the nearest samples. The KNN algorithm is simple and easy to implement, does not require parameter estimation and algorithm training, and is very suitable for multi-classification problems. But when the samples are not balanced, the prediction result may has a large deviation.

3.2 Naive Bayes Algorithm

The basic idea of Naive Bayes is to find the probability of a given text document category by using the joint probability of words and categories. This method is widely used in sentiment analysis. The algorithm The space-time overhead in the classification process of this method is relatively small. It is also very suitable for training with large numbers of sets; And the algorithm supports incremental operations. And it performs best when attribute correlations are small. However, it may not work well when sample attributes are associated.

3.3 Support Vector Machine (SVM)

As a supervised learning model that can effectively analyze data, SVM is a Novel machine learning methods for the regression analysis and classification applications related to machine learning algorithms [2]. The algorithm can solve many small-sample machine learning problems; And it has excellent adaptability to new samples. However, It is usually hard to solve multi-classification problems. It also appears less likely in the face of large samples.

3.4 Maximum Entropy Algorithm

The maximum entropy classifier belongs to the probabilistic classifier of the exponential model class. Based on the principle of maximum entropy, from all models that fit the training data, the model with the maximum entropy is selected [3]. The algorithm has high accuracy and can flexibly set constraints. Researchers can adjust the model's fitness to data which is still unknown through the number of constraints. But it is computationally expensive.

3.5 Related Research

By combing through many literatures, it can be found that there have been great progress and achievements in text sentiment analysis. In the early 21st century, pang et al. took the lead in applying machine learning methods to study movie reviews, and achieved relatively good performance [4]; in 2007, Xu Jun et al. used naive Bayes and maximum entropy methods to conduct sentiment classification research on news and comment corpus, results indicated the machine learning method could get good classification performance in sentiment-based text classification, and the accuracy rate can reach about 90%. As a feature item weight, it could make accuracy of sentiment-based text classification higher [5]; Then, Fan Na et al. applied a two-layer CRF model to get local texts' sentiment, and then used the weighted nearest neighbor algorithm to get texts' global sentiment. As a result, this sentiment analysis method significantly improves the accuracy of emotion recognition [6]. In 2014, Li, Tingting et al. considered words and part-of-speech features and proposed a sentiment analysis method based on the combination of SVM and CRF multi-features. It was applied to the corpus provided by COAE2014 and achieved good results. It could be found that the accuracy of the SVM model was 88.72%, the correct rate of the CRF model is as high as 90.44% [7]; in 2020, Xu, Linhong et al. used machine learning algorithms such as SVM to achieve good results in automatically analyzing the sentiment tendency of citations, with an accuracy rate of 93.4%, which can meet the citation requirements. The basic requirements of sentiment analysis [8]; Li, qiao et al. used a model of naive Bayeset.al to analyze the position analysis of twitter rumors parameter [9]. Wang, Wentao et al. built a sentiment classification model based on sentiment dictionary and SVM, and obtained that the accuracy of the classification model reached 0.96, and concluded that the classification model has better results and can be used for sentiment classification of comment data [10]. In 2021, Xia, Changyu implemented four machine learning models of SVM, KNN, XGBoost and Naive Bayes for the comments on the new crown pneumonia epidemic on Weibo. At the same time,

a comparative experiment was designed in terms of feature extraction method and feature dimension selection, and the optimal feature extraction method and optimal feature dimension were obtained. Finally, it was shown that the best machine learning model under this condition was XGboost [11].

4 Conclusions

At present, more and more scholars apply machine learning to analyze text sentiment, and scholars can appropriately apply them to the required conditions according to the advantages and disadvantages of various algorithms in machine learning. It is worth noting that, in many cases, the data analysis of a single algorithm cannot achieve the optimal effect. It is an important future research direction to combine multiple algorithms in a timely manner to improve the effect. At the same time, some sarcasm or sentences with multiple meanings in the text are still difficult to deal with, and some false comments are difficult to identify one by one before conducting research. Subsequent research can try to filter out obviously untrue comments by improving relevant algorithms, which may be a direction of action in the future.

References

1. Zhong, Jiajia, Liu, wei, Wang, Sili & Yang, heng. (2021). A Review of Text Sentiment Analysis Methods and Applications. Data Analysis and Knowledge Discovery (06), 1-13.
2. Li, Mengnan & Wang, Mingyan. (2021). A review of sentiment analysis methods and applications based on machine learning. Software engineering (09), 21-23+8. doi: https://doi.org/10.19644/j.cnki.issn2096-1472.2021.09.005.
3. Liu, shuang, Zhao, Jingxiu, Yang, Hongya & Xu, Guanhua. (2018). A Review of Text Sentiment Analysis. Software Guide (06), 1-4+21.
4. Bo Pang, Lillian Lee & Shivakumar Vaithyanathan. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. *CoRR*.
5. Xu, jun, Ding, Yuxin & Wang, Xiaolong. (2007). Automatic Sentiment Classification of News Using Machine Learning Approaches. Chinese Journal of Information (06), 95-100.
6. Fan, na, An, Yisheng & Li, Huixian. (2012). Research on text emotional analysis methods based on the K-near-neighboring algorithm. Computer engineering and design (03), 1160-1164. doi: https://doi.org/10.16208/j.issn1000-7024.2012.03.053.
7. Li, Tingting & Ji, Donghong. (2015). Sentiment analysis of microblog based on multi-feature combination of SVM and CRF. Application Research of Computers (04), 978-981.
8. Xu, Linhong, Ding, kun, Lin, yuan & Yang, yang. (2020). Research on Automatic Recognition of Citation Sentiment Based on Machine Learning Algorithm——Taking the Field of Natural Language Processing as an Example. Modern intelligence (01), 35-40+48.
9. Li, qiao & Liu, yu. (2019). Research on Twitter Rumor Stance Analysis Based on Machine Learning. Electronic Design Engineering (21), 36-39+44. doi: https://doi.org/10.14022/j.cnki.dzsjgc.2019.21.009.

10. Wang, Wentao & Zhang, Shibao. (2021). Sentiment Analysis of Weibo Netizens Based on Sentiment Dictionary and SVM. Modern information technology (24), 24–27+31. doi: https://doi.org/10.19850/j.cnki.2096-4706.2021.24.007.
11. Xia, Changyu. (2021). Sentiment Analysis of COVID-19 Weibo Sentiment Based on Machine Learning and Deep Learning (Master Thesis, Jiangxi University of Finance and Economics). https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD202102&filename=1021612757.nh