Sentiment Analysis of Twitter Comments Using Naive Bayes Classifier

Ziyao Zhang[1,a,*]

[1]Major in Software Engineering, School of Computer Science, China University of Geosciences, Huaian, Jiangsu Province, 430000, China
a. zzyswitch20020322@163.com
**corresponding author*

Abstract: Social media has a significant role in how people express their emotions and elaborate on their opinions in today's culture. There are many new forms of social media, and Twitter is one of them. In this experiment, the sentiment of pre-processed Twitter comment data was examined using naive Bayes and logistic regression techniques. In order to categorise the emotional tendency of text for Twitter comments, a naive Bayesian classifier is created. In processing this material, the Naive Bayes and logistic regression models' benefits and drawbacks are compared and summarised. Naive Bayes can achieve good accuracy with binary emotion analysis. The accuracy of the naive Bayes model is 0.06 points higher than that of logic training under identical processing settings, and the recall rate is 0.05 points higher.

Keywords: naive bayes, sentiment analysis, Twitter comment, natural language processing, machine learning

1. Introduction

Sentiment analysis, which can also be called emotional trend analysis or opinion mining, refers to extracting information from user opinions [1] and analyzing text, audio, images, etc., so as to get people's opinions. The formation of opinions, attitudes and emotions. Sentiment analysis is also related to opinion discovery. In addition, emotional analysis is often used by advertisers, filmmakers and other organizations that wish to acquire their customers. Text sentiment analysis is simply defined as the process of analyzing and summarizing the positive or negative attitudes, opinions, evaluations, etc. that one or more users have towards an entity [2].

The traditional emotion classification methods mainly include dictionary-based models and machine learning models[3]. Sentiment classification can help users accurately analyze huge and complex data information, or help people quickly locate and filter the required information, which has high research significance and application value. In many social media groups, Twitter has always occupied a considerable position. Most of the public opinions are transmitted through Twitter because Twitter is regarded as a medium that can express social wishes and views.

The purpose of this study is to investigate the subjective sentiment of comments, as well as other Twitter-specific sentiment analysis research applied to Twitter data and their outcomes. The method based on machine learning refers to using the marked corpus, using different feature weights [4] formula and feature selection method to process the collected corpus, using a machine learning algorithm to train to get a classifier, and using the trained classifier to recognize the new text. Bayes model, a classification technique based on the Bayes theorem and the presumption of independence of defining circumstances, was used for training. Logistic regression was integrated to demonstrate the model's classification impact.

The classification effect of the naive Bayes model on dichotomous emotion learning is demonstrated in contrast, the Bayes model is studied and elaborated, and the optimization direction of the model is analyzed. Compare the difference between the Bayesian model and the logistic regression training model in handling dichotomous sentiment analysis under the same conditions.

In this study, emotion analysis can judge Twitter users' emotions, which can get good positive feedback in handling public opinion analysis. Combined with naive Bayes model training, analyze the advantages and disadvantages of the Bayes model in processing emotion analysis. The application of emotion analysis has great value in the discussion of public opinion and commercial application.

2. Data Text Processing

In terms of the data set, crawling comments, which is a supervised machine learning process, we know in advance the emotional orientation of each movie review, which can only be positive or negative. Positive Sentiment is going to be 1, negative sentiment is going to be 0.

The data included 25,000 comment data, 10,000 positive emotion data, 10,000 negative emotion data, and 5,000 defined data.

In the experiment of Zhao Jianqiang and Gui Xiaolin[5], six different pretreatment methods were studied to affect the emotional polarity classification in Twitter. The results show that the removal of URL, stop word and number has the least impact on the performance of the classifier, and the replacement of negative and extension of acronyms can improve the accuracy of the classification.

In this experiment, this processing method is used for data cleaning and data reorganization of text data. The processed tweets are stored as a two-dimensional list.

The document processing method also refers to Li Jingmei's [6] and Xianghua Fu's [7] experiments, and on this basis, the combination is expanded. In the direction of data normalization, select the WordNetLemmatizer class to assist. WordNetLemmatizer is a class in the nltk library that converts words to their basic form, which is the root form or lexical form. It uses the WordNet dictionary to look up the basic form of a word and consider the word's part of speech. For example, convert "running" to "run" and "better" to "good." WordNetLemmatizer can be used for natural processing tasks such as text classification, information retrieval, and semantic analysis.

3. Different Models

3.1. Naive Bayes Model

Naive Bayes classification method, a probabilistic model-based classification method, was first proposed by Maron and Kuhns [8] in 1960. Lewis [9] described how to use NB for text classification and information retrieval.

The naive Bayes algorithm uses prior probability, which is the background knowledge that a hypothesis is true, named after the two basic assumptions of conditional independence and location independence, is a classification method based on Bayes' theorem and the assumption of independence, which has a wide range of applications. The term "naive" in the so-called "naive Bayes" model refers to the hypothesis of feature conditional independence, which states that the characteristics used for classification are conditionally independent when the class is known. This premise enables naive Bayes learning.

The Bayesian classifier needs to estimate fewer parameters, is not too sensitive to missing data, the algorithm is relatively easy to handle, strong interpretation. In comparison to other classification techniques, it has the lowest error rate in theory. To develop an attribute model, the text is often categorised, and attributes that are reliant on one another can be handled individually.

The naive Bayes classifier's trainer will estimate the class prior probability P (c) based on the training set D and the conditional probability for each attribute during the training phase.

In this project, a number of naive Bayes method training models are selected among many naive Bayes methods. Polynomial naive Bayes is mainly suitable for probabilistic calculations of discrete features, and Sklearn’s polynomial model does not accept negative input values. To deal with continuous variables, choose the Gaussian model. Polynomial naive Bayes is mostly used for document classification. It can determine the likelihood that a document fits into a particular category. The category of the document is the maximum likelihood kind.

3.2. Logistic Regression Training Model

Logistic regression classification is to make output prediction of marked data through machine learning summary and induction algorithm and classification model or classification decision function [10].

The input discrete or continuous variables are used to categorize the output finite discrete values. It is a statistical learning technique that is primarily applied to binary classification issues. Its main goal is to build a logical function that models the link between input attributes and associated output tags. The logical function usually adopts a sigmoid function, whose output value is between 0 and 1, which can be regarded as a probability value, indicating the probability that the input sample belongs to the positive example.

The training process of the logistic regression model usually uses the maximum likelihood estimation method, that is, to maximize the likelihood function of all samples in the training set, so as to obtain the optimal parameter values. During prediction, the characteristics of input samples are put into the trained logic function to obtain a probability value. If the probability value is greater than 0.5, the prediction is a positive example; otherwise, it is a negative example.

The logistic regression model has the advantages of being simple and easy to understand, fast calculation speed, strong interpretation, and so on. It is widely used in practical applications, such as advertising click rate prediction, credit risk assessment, medical diagnosis and so on.

4. Experimental Design and Implementation

Using the Twitter comment data set in Kaggle, the comments were classified into two categories, positive and negative, by examining user comments and combining them with the twitter_samples data of nltk's public database.

4.1. Word and Data Preprocessing

Segmentation of basic data, normalization of data content, and then data cleaning. In the data normalization section, the steps of part-of-speech tagging, garbage data processing and part-of-speech restoration are included. Using the garbage data processing method mentioned above and combining it with the part-of-speech tagging in the NLP library to complete the normalization part of the data. The results of the comparison of data processing are shown in Table 1.

As can be seen from the table, during data processing, coherent sentence patterns originally containing the name of the tweet publisher and special symbols are deleted, word segmentation and part of speech processing are carried out, and data normalization is achieved through nltk part of speech processing. The final success of processing is the one-dimensional list of words displayed after processing.

Table 1: Data processing result.

DATA	CONTENT
Initial comments	["@metalgear_jp @Kojima_Hideo I want you're T-shirts ! They are so cool ! :D", '@AxeRade haw phela if am not looking like Mom obviously am looking like him :)']
Original positive data	['@metalgear_jp', '@Kojima_Hideo', 'I', 'want', "you're", 'T-shirts', '!', 'They', 'are', 'so', 'cool', '!', ':D'], ['@AxeRade', 'haw', 'phela', 'if', 'am', 'not', 'looking', 'like', 'Mom', 'obviously', 'am', 'looking', 'like', 'him', ':)']
Part-of-speech tagging result	[('haw', 'NN'), ('phela', 'JJ'), ('look', 'NN'), ('like', 'IN'), ('mom', 'NN'), ('obviously', 'RB'), ('look', 'VBP'), ('like', 'IN'), (':)', 'NNS')]
The processed data	['want', 't-shirts', 'cool', ':d'], ['haw', 'phela', 'look', 'like', 'mom', 'obviously', 'look', 'like', ':)']

4.2. Data Preparation

Prepare the model data, and mark the corresponding “positive” and the “negative” label data, after that, the data is scrambled, 7 to 3 assigned to the training set and the test set.That's 14,000 random training sessions and 6,000 random testing sessions, plus an additional 5,000 alternate sessions.

4.3. Training Model

After data processing, the naive Bayes classifier was selected to compare the classification effect of Twitter comment data respectively by Logistic Regression and naive Bayes algorithm. Train the model using the two methods mentioned above.

By determining classification categories, establishing probability for each category, obtaining logarithmic priors, calculating positive probability and negative image probability of words, establishing freqs dictionary, finally obtaining frequency dictionary and training model success, and realizing naive Bayes classifier.

The model accuracy reached 0.99675 with naive Bayes processing results. The precision value under positive labels is 0.9994256, the recall rate is 0.9985652, and the F-measure value is

0.9989952. The precision value under negative labels is 0.9985787, the recall rate is 0.9994310, and the F-measure value is 0.9990046. As shown in Table 2,

Table 2: The naive bayes model evaluates information.

INDEX	VALUE
Model accuracy	0.99675
Pos precision	0.9994256
Pos recall	0.9985652
Pos F-measure	0.9989952
Neg precision	0.9985787
Neg recall	0.9994310
Neg F-measure	0.9990046

Under the Logistic Regression model, the accuracy reaches 0.93278. The precision is 0.9332, recall rate is 0.9367. Through evaluation, it is found that the evaluation data of the logistic regression model is lower than that of the naive Bayes model, and the comparison results are shown in Table 3.

Table 3: Comparison of model evaluation results.

MODEL	ACCURACY	PRECISION	RECALL
Naive Bayes	0.99675	0.9994	0.9985
Logistic Regression	0.93278	0.9332	0.9467

For binary sentiment analysis, the accuracy rate and recall rate of naive Bayes is higher than that of the Logistic regression training model. The accuracy of the naive Bayes model is higher than that of logic training by 0.06, and the recall rate is higher by 0.05.

4.4. The Results of Model Training

It can be found that for binary emotion analysis, naive Bayes can get quite a good accuracy; Based on the experimental results and accuracy values of the model, it is concluded that the Naive Bayes model can be used to predict emotional patterns and has certain applications.

In terms of data volume, logistic regression is a discriminative model that is goal-driven, does not reflect joint probability, and directly predicts the output using training data, whereas Naive Bayes is a generic model that can adapt the data more efficiently under certain situations. As a result, the results obtained using experimental data for training are slightly subpar compared to Naive Bayes.

But at the same time, naive Bayes requires a lot of time and data. Further research can be done by increasing the amount and variety of review data or by other methods to increase the value of accuracy and optimize the naive Bayesian model to obtain more efficient methods.

5. Conclusion

Based on naive Bayes and nltk text processing, this experiment uses language to discuss and study the sentiment analysis of comments on English comment data. This experiment shows a basic method to classify tweets into positive or negative categories based on naive Bayes and expounds language processing and naive Bayes model. On the same basis, the accuracy of the naive Bayes model processing emotion analysis is slightly higher than logistic regression.

Despite the comparatively high accuracy of binary emotion processing, naive Bayes has a poor classification effect for multivariate emotion learning when the number of characteristics or the correlation between attributes is considerable. As well as the input data of the form of processing requirements for higher shortcomings. In addition, this experiment only compares the training gap between naive Bayes and logistic regression under the same condition, and more models need to be trained and compared in the future.

In the follow-up work, further studies can be carried out by increasing the amount and type of review data or by other methods to improve the value of accuracy, as well as to optimize the naive Bayes model to get more efficient methods, compare the treatment of different models and further optimize the experiment.

Natural language processing plays an important role in artificial intelligence learning. In the future, we hope to see more convenient and faster text processing methods and building models.

Acknowledgements

Upon the completion of this thesis, I'd like to offer my heartfelt gratitude to everyone who has provided me with encouragement and assistance throughout my academic career. First and foremost, I'd like to thank Professor Shlomo Ta'asan for his direction during the project. I benefited a lot from his professional knowledge and rigorous attitude, which greatly promoted me academically. Second, I want to express my gratitude to my family and friends for their unwavering support and encouragement as I continue on my academic journey. Finally, I would like to thank all the institutions and organizations that have provided support and assistance to this research, including my university, the China University of Geosciences. Thanks again to all the above people for their support and help. Without their support, I could not finish this paper.

References

[1] Chen Long, Guan Yu, He Jinhong, Peng Jinye. Research Progress in Sentiment Classification [J]. Journal of Computer Research and Development,2017,54(06):1150-1170.

[2] C. Troussas, M. Virvou, K. J. Espinosa, K. Llaguno and J. Caro, "Sentiment analysis of Facebook statuses using Naive Bayes classifier for language learning," IISA 2013, Piraeus, Greece, 2013, pp. 1-6, doi: 10.1109/IISA.2013.6623713.

[3] Lin Jianghao, Yang Aimin, Zhou Yongmei et al. A Naive Bayes based microblog Sentiment Classification [J]. Computer Engineering and Science,2012, 34(09):160-165.

[4] Lu Ling, Wang Yue, Yang Wu. A Naive Bayes-based sentiment Classification Method for Chinese Comments [J]. Journal of Shandong University (Engineering Science),2013,43(06):7-11.

[5] Y. Jianqiang and G. Xiaolin, "Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis," in IEEE Access, vol. 5, pp. 2870-2879, 2017, doi: 10.1109/ACCESS.2017.2672677.

[6] Li Jingmei, Sun Lihua, Zhang Qiarong, et al. A naive Bayesian classifier for Text Processing [J]. Journal of Harbin Engineering University,2003(01):71-74.

[7] Combine HowNet lexicon to train phrase recursive autoencoder for sentence-level sentiment analysis[J]. Xianghua Fu; Wangwang Liu; Yingying Xu; Laizhong Cui.Neurocomputing,2017.

[8] Maron M E, Kuhns J L. On relevance, probabilistic indexing and information retrieval [J]. Journal of the ACM(JACM),1960,7(3):216-244.

[9] Lewis D D. Naive(Bayes)at forty: The independence assumption in information retrieval[C]//Machine learning: ECML-98.Springer Berlin Heidelberg,1998:4-15.

[10] EDITH LAW, BURR SETTLES, TOM MITCHELL.Machine learning and knowledge discovery in databases[M]. Springer Berlin Heidelberg:2010.