Sentiment Classification of Tweets
COMP90049 Report

Anonymous

1 Introduction

With the rise of online social networks, a large number of users use words to express their emotions on the Internet, including emotional expression of life and events, as well as experience and evaluation of all aspects of products. Twitter is a popular microblogging service where users can create status messages which are called tweets. These tweets sometimes express opinions on different topics. (Huang and Bhayani, 2009)

This report refers to the resource datasets published by Vadicamo et al. (2017) and Go et al. (2009). This project can evaluate the performance of some machine learning models in predictive text sentiment analysis. This article will focus on the following models: Naive Bayes model, logistic regression model and neural network model. Through experiments, it is found that using machine learning technology, we can obtain a higher accuracy rate for the sentiment classification of Twitter messages.

2 Literature review

At present, there are many papers on sentiment analysis mainly cover the field of microblog and comment, and the field of Twitter is relatively less. This may be because Twitter is a social platform that has become popular in recent years. But in general, using machine learning is a good area for research. (Manning and Schuetze, 1999) Pang and Lee pointed out in the Emotion Classification using Machine Learning Techniques published in 2002 that various Machine Learning Techniques Naive Bayes, Maximum Entropy, and support Vector Machine have been compared and adjusted in specific areas of film criticism. It is found that a high accuracy rate is obtained by using support vector machine and unary model. (Pang and Li, 2002)

In addition, detecting emotions in texts is currently a popular research direction, and in 2002, Turney proposed a "semantic orientation" algorithm to detect emotions. In 2004, Pang and Lee proposed a layered method. First, the text was divided into two types containing emotion and not containing emotion, and then the text containing emotion was divided into two types, positive and negative. In 2005, researchers also completed the label classification of emoticons. This research is very important for Twitter's sentiment analysis, because many Twitter will use emoticons in their published content. Vadicamo et al. (2017) and Go et al. (2009) publish articles about Twitter sentiment classification. This was a very important milestone, and a lot of subsequent research on Twitter sentiment analysis has been based on them. At the same time, these two articles also provide data set training. The data set is divided into three parts, which are 159,253 test sets, 19,894 test sets, and 19,906 development sets. Among them, the model model is trained through the training set. After getting the trained model, use the development set to test, and finally use the model on the test set to predict the label. In addition, each data set contains four csv files, respectively full, glove, count and tfidf. The full data set contains a string of original tweets. The data in the other three data sets is generated by some algorithms from the data in the full data set.

3 Methods

3.1 Naive Bayes

Naive Bayes (NB) is a simple model for classification. (Frank and Bouckaert, 2006) It is simple and works well on text categoration. Therefore, it is very suitable for this project. This project adopts three bayes

classifier models: Gaussian NB, Multinomial NB and Bernoulli NB. Multinomial NB is usually used in the event model of document classification, where the event represents the number of occurrences of a word in a single document .Unlike Multinomial NB, the value of each element in the Bernoulli NB model can only be 0 or 1. If we only need to judge whether there are keywords, then Bernoulli NB is a better choice.

3.2 Logistic regression

Logistic Regression classifiers are used to solve multi-class classification tasks. (Pranckevičius and Marcinkevičius, 2017) Although the Logistic Regression is called Regression, it is actually a classification model and is often used for dichotomies. Logistic Regression is favored by the industry for its simplicity, parallelization and explicability. The essence of Logistic regression is to assume that the data obey this distribution, and then use maximum likelihood estimation to estimate the parameters. Features are not required to be independent from each other, and the results are more universal. Therefore, logistic regression is a good choice in this project.

3.3 Neural network

Neural network is a mathematical model formed by referring to the working principle of biological neural network. Neural network is one of many algorithms in machine learning. It can be used for supervised tasks, such as classification and visual recognition, as well as unsupervised tasks. At the same time, it can deal with complex nonlinear problems. Its basic structure is neuron, as shown in the following Figure (See Figure 1) : x1, X2, x3 represent the input, the middle part is neuron, and the final Hw, B (x) is the output of neuron. The whole process can be understood as input to processing to output.

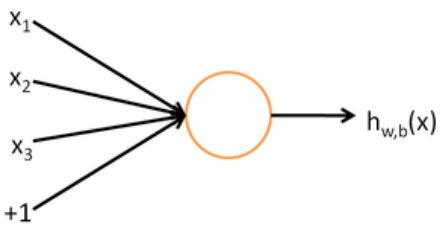

Figure 1-The basic structure of a neuron network

A single neuron contains two partial weights and biases. Each input value entering the neuron will be similar to y=wx+b, where w is the weight, b is the bias, x is the input value, and y is the result of the single input value. After the activation function is activated, the output result is f(wx+b), where f is the activation function. When there are multiple input values using the activation function, the output value is f(w1x1+w2x2+...+wnxn+b).

A neural network is composed of multiple neurons. In the nodes of the neural network, each node corresponds to a weight vector W.(Severyn and Moschitti, 2015)

Nowadays, with the increasing amount of data, more and more abundant computing resources, as well as the improvement and optimization of algorithms, the layers of neural network become more and more, and the learning effect becomes better and better, which is very suitable for the processing of large data.

4 Results

4.1 Calculation of baseline

The zero-R classifier is the simplest classifier. This method only selects a category with the highest probability as the classification result of the unknown sample based on historical data statistics. That is to say, for any unknown sample, the classification result is the same. The zero-R classifier simply uses the category of the majority class as the predicted value. Although this classifier does not have any predictive power, it can be used as a contrast classifier with other classifiers. That is to say baseline performance. Zere-R is also the most

commonly used baseline measurement method in the field of machine learning.After measurement, the accuracy obtained by zero-R is 0.4066613.

4.2 Results

Visualizing the experimental results (see Table1), and a histogram of the comparison of results was obtained(see Figure2). The blue line is the baseline.

According to the experimental results, it can be found that the accuracy of Count and TFIDF calculated by Multinomial NB and Bernoulli NB is very close. The accuracy measured by Gaussian NB is lower than Multinomial NB and Bernoulli NB.

It is intuitively see that the logistic regression model has the highest accuracy among these models.

The neural network model has the highest accuracy and stability by using different data.

Regardless of which machine learning model was used, the count and TFIDF datasets were generally more accurate than the GLOVE datasets.

Comparing all experimental results to the baseline, all machine-learning models produced higher accuracy than zero-R. It can be proved that machine learning contributes to the improvement of prediction accuracy.

Model	Data	Accuracy
BNB	count	0.73379885
BNB	tfidf	0.73379885
BNB	glove	0.58891791
MNB	count	0.73078469
MNB	tfidf	0.72561037
GNB	count	0.58314076
GNB	tfidf	0.58560233
GNB	glove	0.44936224
LR	count	0.74068128
LR	tfidf	0.74153522
LR	glove	0.67793632
NN	count	0.69727727
NN	tfidf	0.70159751
NN	glove	0.70446097

Table 1 – Results of each model

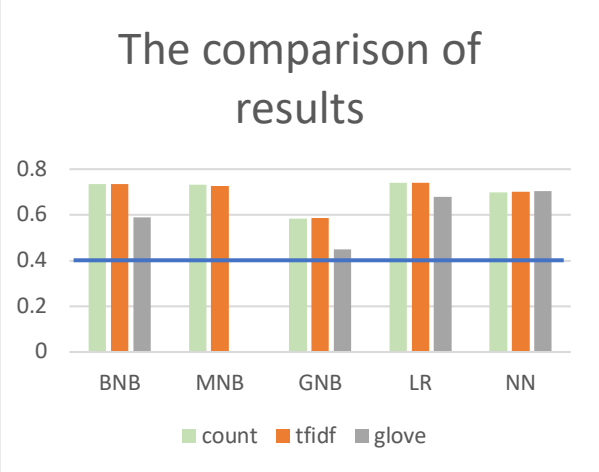

Figure 2 – The comparison of results

5 Discussion

5.1 Contextualise

5.1.1 Naive Bayes

Naive Bayesian model originated from classical mathematical theory, has a solid mathematical foundation, and stable classification efficiency. Therefore, the measured results for the two data sets of count and tfidf are very close. It has a higher speed when training and querying a large number of. In this experiment, the time required for NB operation is the shortest. At the same time, NB is not very sensitive to missing data, and the algorithm is relatively simple. It is often used for text classification, so the prediction accuracy of NB in this project is relatively high.

NB also has disadvantages. For example, there is an error rate in classification

decision. Due to the use of the assumption of the independence of sample attributes, the effect is not good if the sample attributes are related.In addition, it is very sensitive to the expression form of the input data. According to the Russian experimental results, although they are all NB classification models, the accuracy of the prediction results using glove data is significantly lower than that of count and tfidf data.

Multinomial NB is often used for text classification, characterized by words and valued by the number of occurrences of words. In Bernoulli's NB model, the values of each feature are Boolean, that is, true and false, or 1 and 0. In text categorization, it is whether or not a feature appears in a document. Observing the data, it is not difficult to find that many words appear 1 time, which explains the fact that the results of this experiment are very similar. Multinomial NB does not accept the non-negative values created during the LSA stage. Multinomial NB cannot process data values from glove**.** In Gaussian NB, every feature is continuous and normally distributed. Because the distribution of data in this time does not present a good normal distribution, the accuracy of Gaussian naive Bayes is lowee than that of other models.

5.1.2 Logistic regression

Logistic regression is easy to implement, without scaling features and adjusting parameters. In this experiment, the implementation is very convenient and the running time is reasonable, not particularly long. Logistic regression can directly model the classification probability without realizing the hypothetical distribution, avoiding the problem of inaccurate hypothetical distribution. It can not only predict the category, but also get the predicted probability, which is very useful for some tasks that use probability to assist decision-making. Therefore, the logistic regression in this experiment has the highest accuracy.
Logistic regression actually uses the predicted value of the linear regression model to approximate the log probability of the true label of the classification task. Therefore, the processing effect of non-linear data is not good, such as image data. The algorithm is relatively simple and easily surpassed by other algorithms. In addition, logistic regression highly dependent on correct data representation. All important variables/characteristics should be identified so that they can work well. Because this project uses processed data, the results of the experiment are not greatly affected by the data.

5.1.3 Neural network

Neural network is an algorithm model inspired by biological neural network.
It is a pattern matching that is often used for regression and classification problems, but has a huge subdomain consisting of hundreds of algorithms and variants of various kinds of problems. Neural networks perform extremely well in many tasks, and their algorithms can be quickly adapted to new problems.

In theory, the results from the neural network should have the highest accuracy, but the experimental results are inconsistent with the prediction. The main reason for this result is the absence of tuning parameters. According to the prediction principle of neural network, we need to adjust parameters such as weight. However, due to the complexity of the neural network model, which takes a long time to run, and the limited experimental time and conditions, no parameter adjustment was carried out in this experiment, resulting in poor experimental results. However, even in the case of no tuning parameters, there is still a good prediction result, which shows that the neural network model has a strong prediction ability. At the same time, it can also be found that the adjustment of parameters will have an impact on the experimental results.

The disadvantage of neural network is that it needs a lot of data training, and the

hardware configuration of training is very demanding. It takes more than an hour to run the neural network model in this project at a time. In addition, since the model is in a black box, it is difficult to understand the internal mechanism.

5.2 Ethical issues

According to Damasio's theory, emotions are invested in decision-making. When analyzing and making decisions, people will become emotional, and ultimately emotions will guide people to make choices and actions. Sentiment analysis is useful for consumers who are trying to research products or services, or for marketers to study the public's perception of their company. They can influence user decisions from steps such as user identification, sentiment, analysis, and guidance. Therefore, it is an important research to analyze the user's emotion based on the content posted by the user.

However, sentiment analysis based on tweets also has some moral hazards. First of all, it is morally not allowed to analyze the user's emotions privately without the user's knowledge, and it will cause the user's privacy to be leaked. In addition, if the results of the analysis are not used properly, it will have a huge negative impact. For example, malicious guidance based on the user's emotions leads to the loss of the user's property and damage to the physical and mental health of the user.

Therefore, it is necessary to make rational use of the technology of sentiment analysis. And to ensure the security of the results after analysis to avoid problems such as huge damage and privacy leakage. It is best to keep the analyzed user informed when using this technology.

6 Conclusions

To conclude, this report reviews the effectiveness of Naive Bayes, logistic regression and neural networks. Based on the given data and experimental conditions, logistic regression has the highest accuracy, while Gaussian naive Bayes has the lowest accuracy. In addition, parameter setting plays an important role in the prediction accuracy of the model. In the future, the accuracy of tweet-based sentiment analysis can be further improved by continuing to adjust parameters and optimize experimental data.

References

Go, A., Bhayani, R., & Huang, L. (2009). *Twitter sentiment classification using distant supervision. CS224N project report, Stanford, 1(12)*

Vadicamo, L., Carrara, F., Cimino, A., Cresci, S., DellOrletta, F., Falchi, F., & Tesconi, M. (2017). *Crossmedia learning for image sentiment analysis in the wild. In 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), pages 308-317.*

Pang, L. Lee, S. Vaithyanathan. (2002) *Thumbs up? Sentiment Classification using Machine Learning Techniques*

Manning and H. Schuetze.(1999) *Foundations of Statistical Natural Language Processing*

Frank, E. and Bouckaert, R.R. (2006) *LNAI 4213 - Naive Bayes for Text Classification with Unbalance Classes*.

Huang, L. and Bhayani, R. (2009) *Twitter Sentiment Analysis*. Available at: http://cli.gs/9ua6Sb.

Pranckevičius, T. and Marcinkevičius, V. (2017) “Comparison of Naive Bayes, Random Forest, Decision Tree,Support Vector Machines, and Logistic Regression Classifiers for TextReviews Classification,” *Baltic Journal of Modern Computing*, 5(2). doi:10.22364/bjmc.2017.5.2.05.

Severyn, A. and Moschitti, A. (2015) “Twitter Sentiment Analysis with deep

convolutional neural networks," in *SIGIR 2015 - Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, Inc, pp. 959-962. doi:10.1145/2766462.2767830.

Shi, L. *et al.* (2010) *Rough Set Based Decision Tree Ensemble Algorithm for Text Classification*, *Journal of Computational Information*. Available at: http://www.JofCI.org1553-9105/.