Project 2 Tweets r mad, or r they!?

Anonymous

1 Introduction

Nowadays people usually like to express their opinions on social media. The various interactions such as shares, comments, recommendations that users use to generate a large amount of data are called user-generated content. These data contain a large amount of information, reflecting the user's inherent behavior. For example, on Twitter and Microblog, every tweet of users on these social media may contain the user's mood and opinions at the time, which brings us a challenging topic to predict user's emotions by analyzing tweets. Sentiment analysis is also a challenging problem in natural language processing (NLP).

This report will apply specific evaluation metrics to discuss the performance of different classifiers (Naïve Bayes, Decision Tree, Random Forest) on the dataset and the features of different datasets. Finally, we will demonstrate if it is feasible to conduct sentiment analysis on tweets.

2 Related Work

In this task, we have nine files in our dataset including the original tweet files, different format tweet files, the tweet files which contain labels. And for the issue of tweet sentiment analysis, many scholars have proposed their own methods and assumptions.

2.1. Dataset

The dataset mainly contains 33k tweets which divided into three parts: a training set, an evaluation set, and a test set. And training and evaluation set has been labeled the sentiment which involved in positive, negative and neutral. And the label of the test set is left for us to finish. Data is provided by Rosenthal, Sara, Noura Farra, and Preslav Nakov. (2017)

2.2. Related Research

In recent years, with the widespread promotion of social media, the amount of data on social media has become more and more huge, which makes the analysis of [1]these tweets more important. Zhao, J., Dong, L., Wu, J., & Xu, K. (2012) analysis the emoticon of each tweet, predicting tweets sentiment by emoji and some emoticons. Bakliwal, A., Foster, J., van der Puil, J., O'Brien, R., Tounsi, L., & Hughes, M. (2013) conduct sentiment analysis of political tweets. They select 1000 discriminative words and other features to analyze tweets.

3 Feature Analysis

3.1. Original Features Analysis

The original dataset has 47 features. In this part, using Naïve Bayes, Decision Tree, and Random Forest as the classifier to analyze this dataset. Meanwhile, the file "train.arff" is used to train the model and the file "eval.arff" is used to evaluate the performance of each method.

3.1.1. Naïve Bayes, Decision Tree,

Naïve Bayes classifiers are one of the methods of machine learning which are based on applying Bayes' theorem with strong independence assumptions between the features. Decision Tree is also a common classification method in machine learning. In this report, the training data is pre-processed by WEKA and then classified by Naïve Bayes classifiers which are provided by WEKA. And in WEKA we use J48 which is one of the method in the Decision Tree. The result of this method is as follows.

Table 1. The accuracy of NB, J48

	Naïve Bayes	J48(unpruned)
Accuracy	46.4989%	45.9712%

From the accuracy of each method, we find that Naïve Bayes has the higher score and

[1] Weka: http://www.cs.waikato.ac.nz/ml/weka/

[2]the accuracy of J48 is lower than Naïve Bayes, one of the reason may be that the J48 method we used is unpruned, and there may be many unrelated features in the data that will cause this method overfitting. In order to verify this hypothesis, we will compare Naïve Bayes and J48(pruned). The result is below.

Table 2. The accuracy of NB, J48(pruned)

	Naïve Bayes	J48(pruned)
Accuracy	46.4989%	47.4122%

The result shows that J48(pruned) makes a great improvement in accuracy. However, these two methods with 47 features still don't get a satisfactory accuracy result. So we analysis more in-depth with these two methods. The evaluation metric is as below.

Table 3. The evaluation metric of NB, J48(pruned)

	P(NB)	R(NB)	P(J48)	R(J48)
Positive	40.7%	27.5%	42.0%	22.0%
Negative	31.3%	19.9%	32.6%	19.4%
Neutral	51.4%	69.8%	51.2%	75.3%

From the result of the evaluation metric, we find that the recall of the Neutral class is higher than the recall of the Positive and Negative class in both Naïve Bayes and J48 classifiers. There may be two main reasons for this problem. One of the reasons is that some features in the dataset are not related to the sentiment analysis. For instance, "big", "we" are not having a strong relation with sentiment class. And another reason we think is the features are not enough, some strongly related sentiment features are not included in the dataset. For example, some Emoticon features can predict the sentiment directly.

3.2. Deleted Unrelated Features

As mentioned in 3.1.1, some features may not have a strong relation with the sentiment. In response to this situation, we decided to delete the 46 features one by one, delete each of the 46 features, and then compare each accuracy by experiment. The results are as follows.

Table 4. The accuracy of each method after deleting feature.

	Original	big	do	happy	ice
NB	46.49%	46.51%	46.55%	46.07%	46.49%
J48	45.97%	46.01%	46.21%	45.99%	45.82%

Through experiments, we found that after deleting some features, the accuracy will not only decrease but will increase. This can explain to some extent that these features are not related to sentiment. So we will delete these unrelated features and leave the features related to the sentiment (such as the red part of the table). According to our statistical analysis, a total of 40 features apply to Naïve Bayes, and 23 features apply to J48. The accuracy after deleting these unrelated features are as follows.

Table 5. The accuracy after deleting features

	Original features	Deleted features	Total number of deleting
Naïve Bayes	46.4989%	46.9657%	7
J48(unpruned)	45.9712%	48.0414%	24

From the result from Table 5, we believe that these unrelated features will affect the accuracy of the two methods to some extent. However, we find another problem that the accuracy of Naïve Bayes is much lower than J48. For this problem, the assumption we make may be because the Naïve Bayes method is more suitable for each feature which is conditionally independent, and in our dataset may have many features associated with each other.

3.3. Deleted associated features

[2] http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/J48.html

As we mentioned above, we decided to re-define each feature one by one according to the analysis and analyze the accuracy obtained after deletion. The feature of reducing the accuracy after deletion is retained, and all the features of retaining and increasing the accuracy after deletion are removed. Based on this hypothesis, the result of the experiment is as follows.

Table 6. The accuracy of delating features

	Original	Deleting unrelated	Deleting associated	Number of deleting
Naïve Bayes	46.4989%	46.9657%	47.1484%	26

From the result, we can find that these features may be associate with others in our dataset and affect the accuracy.

4 RESULT

From the experiments, we have designed and the analysis of the results obtained, it is achievable to predict the user's sentiment based on the tweets. Because people want to express emotions by using specific vocabulary to strengthen their tone. In the content of the tweet, you can find the features that can express emotions through continuous screening of features. The tweet is verified by the above-mentioned verified methods. The analysis is carried out to obtain the prediction model, and the accuracy rate will be improved correspondingly with the accurate extraction of the features, and finally, the purpose of accurately predicting the sentiment can be achieved.

5. Conclusions

This report is to make assumptions and verify the hypothesis for a problem which is “Can we use tweet text to help us to identify people sentiment on Twitter?” For this problem, we use two classifiers to do the experiment. And from each result, we propose a new hypothesis and design a new experiment to verify our hypothesis. Finally, based on the result of these experiments we think if we can accurately extraction the features of tweets, we will be able to build an accurate model to predict the sentiment of tweets.

6. References

Timothy Baldwin, Marie-Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. 2015. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 126–135, Beijing, China.

Martin Kay. 1986. *Parsing in Functional Unification Grammar.* In "Readings in Natural Language Processing", B. J. Grosz, K. Sparck Jones & B. L. Webber, ed., pages 125-138, Morgan Kaufmann Publishers, Los Altos, California.

Rosenthal, Sara, Noura Farra, and Preslav Nakov (2017). SemEval-2017 Task 4: Sentiment Analysis in Twitter. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval ’17). Vancouver, Canada.

Frederick Mosteller and David Wallace. 1964. Inference and Disputed Authorship: The Federalist. Addison-Wesley, Reading, Massachusetts.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, 28(1):11-21.

Zhao, J., Dong, L., Wu, J., & Xu, K. (2012, August). Moodlens: an emoticon-based sentiment analysis system for chinese tweets. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1528-1531). ACM.

Bakliwal, A., Foster, J., van der Puil, J., O'Brien, R., Tounsi, L., & Hughes, M. (2013, June). Sentiment analysis of political tweets: Towards an accurate classifier. Association for Computational Linguistics.