**COMP90049 Knowledge Technologies, Semester 1 2019**

Project 2: tweets r mad, or r they!?

1. INTRODUCTION

The growing popularity of social networking has changed the way that people express their feelings. It is important to analyze the emotions from the content that people post on the Internet, especially for those websites that highly depend on customer feedback. This report aims to analyze the people’s attitudes based on the text that is extracted from Twitter through machine learning, and category them into positive, negative and neutral. However, it can be very difficult to determine true attitude because there can be emoticons or slangs involved, and people express their feelings in different ways. Thus it is necessary to pre-process the datasets.

The datasets used in this project are extracted from train-tweets, evaluation-tweets and test-tweets files [1]. To simplify the project, all non-alphabetic characters are removed and there are 45 words selected to be the features that determine the sentiment behind the tweet text. All feature numbers along with tweet id are put into three csv files [1]: training data, evaluation data and test data; and the labels from two txt files [1]: train-labels and evaluation-labels are added to training data and evaluation data csv files corresponding to tweet id.

Based on the pre-process above, training data will be used to create models so that evaluation data can be put in to test how effectively the method works compared with other methods.

One relative research on sentiment analysis of tweets has focused on two techniques: symbolic techniques and machine learning techniques. Much of the research in unsupervised sentiment classification using symbolic techniques makes use of available lexical resources [2]. For machines learning method, it consists of unsupervised learning and supervised learning. The former does not consist of a category and does not provide with the correct targets at all and therefore conduct clustering. While the latter is based on the model during the process and these labeled datasets are trained to produce reasonable outputs when encountered during decision-making [3].

1. METHODOLOGY
   1. Evaluation

Accuracy is defined as the ratio that the number of sentiment identified correctly out of total number of tweets. It shows the effectiveness of each method explicitly. Since the precision is calculated for each category, accuracy will be the only measurement in this project. The equation of accuracy is shown below:

* 1. Methods Applied

This project mainly studies two machine learning methodologies, Naïve Bayes and Decision Tree. The basic mechanism of Naïve Bayes is: for the given sample features to be classified, calculate the probability of occurrence of features under the condition of each class, and the class with the largest probability is considered to be the one that the sample belongs to. For Decision Tree, there are three models implemented: ID3, J48 and CART. For each of those, different measurement is considered as criterion.

In ID3, information gain is the key when decision making.

J48 has adopted gain ratio as criterion:

For CART, the gini index is used when selecting a split node.

Before applying the training dataset to create models, it is essential to remove tweet id. Because in decision tree method, no matter what measurements above three are used, id would be always the first split node. For example, the entropy of id feature is 0 as there would be only one label for each id. Furthermore, for either Naïve Bayes or Decision Tree, the ids in training data never repeat in evaluation data according to comparison program written by Python, which means the evaluation data would never fit the models based on training data with ids in. Thus tweet id dose not contribute to the decision making.

* 1. Results Analysis

The results of Naïve Bayes and Decision Tree with each measurement are shown below:

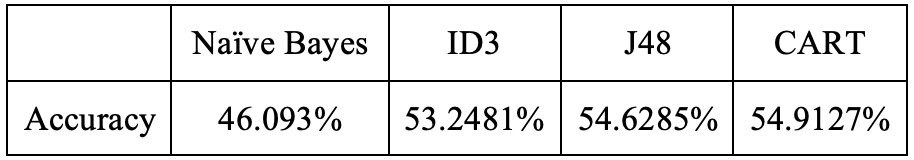


Table 1: Accuracy of each method

Since this is a three-way classification, the chance of each class should be 33%. As it is shown, Naïve Bayes and Decision Tree are able to predict the sentiment based on tweet content. However, the accuracy is quite low, and the average accuracy of machine learning methodologies is around 80%. According to the previous work, Naïve Bayes gives 88.2% of accuracy [3], while for Decision Tree, the accuracy is 80.08% based on the work from Shivaraju Kethavath [4].

An explanation of Naïve Bayes that leads to much lower accuracy than average one, is that all features are considered to be independent in Naïve Bayes model, which means it calculates the probability of each feature separately. For example, the tweet “John cent vs Seth Rollins: am I the only one who's not gonna be happy Sunday in the John cena vs Seth Rollins ...” is labeled to be negative in evaluation dataset, but it is predicted to be positive. Apparently the algorithm did not take “not” into consideration, and this causes a completely opposite result.

Another possible reason is due to the limitation of features selected when building up models. Insufficient features can make the predictions incorrect. A convincing evidence in this project, is that when applying Naïve Bayes to the training data, some key features that determine the sentiment of the tweet are missing. The results of Naïve Bayes are shown below:

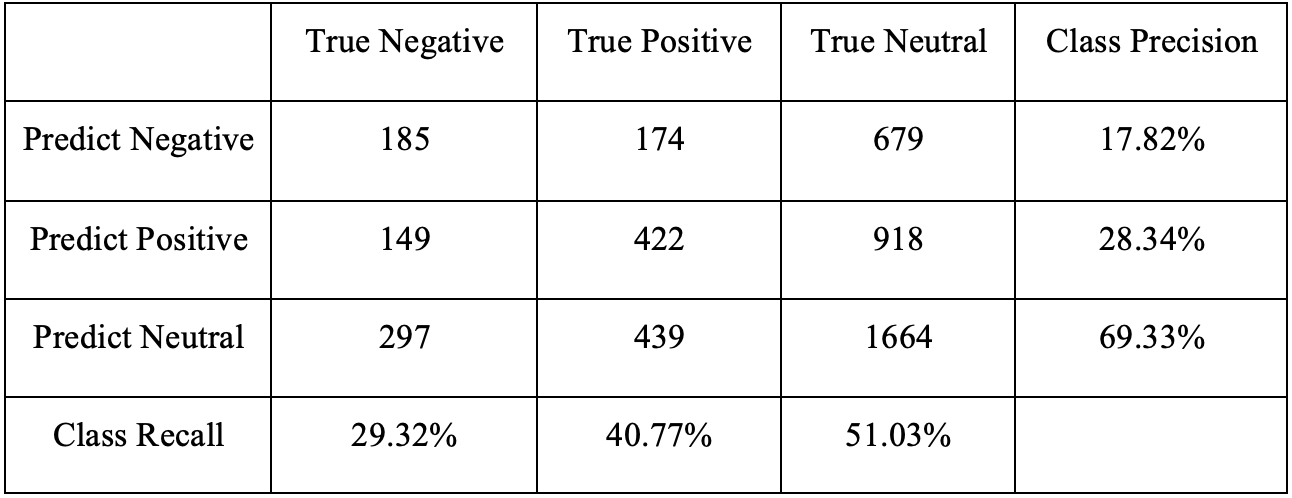


Table 2: Results of Naïve Bayes

As the table shows, nearly half of positive and negative tweets are predicted to be neutral incorrectly. For example, the tweet “My teeth hurt”, and “I am thankful for @bitchy\_antics for pointing out how all the animals in the shop sat up when Thor strided in there!!!! ROFL”, these two tweets are manually labeled to be negative and positive because the words “hurt” and “thankful” are obvious the sentiment tone words. However, they are missing in the features instead the words “my” and “the” appear in the content, so these two tweets are classified into neutral.

For Decision Tree, there is an increase in accuracy even though it is still lower than the average rate which is normally around 80%. This is because there are 11454 “neutral” tweets in training dataset when building up trees, which takes up 49.82% of total training data. Due to the insufficient features, most leaves lead to “neutral” in the tree. When apply the tree models to evaluation dataset, each tweet generally contains one to two features that are attributes within the tree, which is not enough to give an accurate prediction. That means the limited features would follow the nodes and are more likely to lead to “neutral” label. From the table below, it is obvious that around 70% of both “negative” and “positive” tweets are predicted incorrectly to be “neutral”.

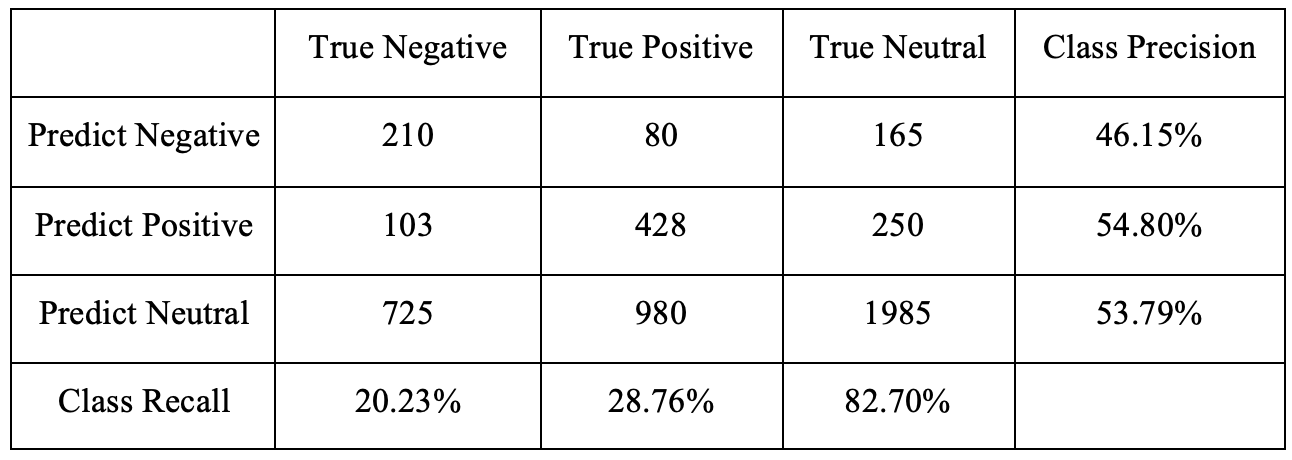


Table 3: Results of ID3 Model

However, it is noticeable that the J48 model and CART give a slightly higher accuracy from table 1. In ID3, it is very sensitive to attributes with many values. For example, labels in our dataset is basically different for different instances, or even more extreme, it is unique for each instance such as one single sentiment label corresponding to one unique id in this case. If we use this attribute (i.e. tweet id) to divide the data set, it will get very high information gain, but the results are not what we want. This is also one of the reasons that tweet ids are removed in pre-processing. To get rid of the shortcoming of ID3, the information gain ratio is used in J48 when decision making for sub-tree. Thus it gives higher accuracy.

1. CONCLUSION

Since the baseline of random guessing is 33.33% for 3-way classification, it is apparently useful to analyze twitter text to predict sentiment through Naïve Bayes and Decision Tree. The main reason of low accuracy is poor quality of limited features. The better solution to improve accuracy may be giving more sentiment words in training data and remove some conjunctions and prepositions such as “and”, “to”, etc.

REFERENCE

[1] 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.

[2] Suchdev, Riya, et al. “Twitter Sentiment Analysis Using Machine Learning and Knowledge-Based Approach.” *International Journal of Computer Applications*, vol. 103, no. 4, 2014, pp. 36–40., doi:10.5120/18066-9006.

[3] Gautam, Geetika, and Divakar Yadav. “Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis.” *2014 Seventh International Conference on Contemporary Computing (IC3)*, 2014, doi:10.1109/ic3.2014.6897213.

[4] Kethavath, S. (n.d.). *Classification of Sentiment Analysis on Tweets using Machine Learning Techniques* (Unpublished master's thesis). Thesis / Dissertation E