Analysing Performance Of Text Classification Models For Sentiment Analysis Of Movie Reviews

N Nagajothi, Dr A R Nadira Banu Kamal

Abstract: Text mining is the process of extracting interesting, non-trivial extraction of implicit, previously unknown and useful information from huge volume of textual data. It has become a vibrant research area. It deals with machine supported exploration of text. It uses the techniques from information retrieval, and extraction along with natural language processing. With the increasing amount of data being available on the web, this tremendous volume of data mostly unstructured text, it is a tough task to analyse their content within a short span of time. The crucial challenge in text classification techniques is the overall performance. With the growth of computer technology, there have been many classification algorithms. Each classification algorithms will get different result at speed and efficiency due to the various features of test data. In this paper, we discuss performance measures such as accuracy, recall and precision for text classification model using Random Forest Classifier and Naïve Bayes classifier. We have done pre-processing before applying classification methods. It has been found that Random Forest classifier has a higher accuracy compared with Naïve Bayes classification for sentimental analysis of movie reviews.

Index Terms: Text mining, pre-processing, classification, information retrieval, Naive-Bayes, Random Forest, Precision, Recall.

1. INTRODUCTION

TEXT mining is same as text analytics which meant as process of extracting high quality information from text. High quality information is extracted through the creation of patterns and trends by using statistical pattern learning. Text mining can be performed with the process of structuring the input text, evaluating and interpreting output. This process is used to give relevance, novelty and interest in the output pattern. Text mining comprises text pre-processing, text transformation, feature extraction, pattern recovery and evaluation. Text mining includes text classification and clustering, extraction, analysis, summarization and modelling. Before the text classification on documents, text pre-processing has been applied to documents. Text pre-processing transforms a text into a structured format. The text pre-processing is achieved into the following steps:

- Tokenization: It eliminates all punctuation, whitespaces, spaces and commas etc.
- Filtering: Stop word filtering is a standard filtering method. It eliminates content information like html or xml tags, articles, conjunctions, prepositions etc.
- Lemmatization: It maps verb forms to the infinite tense and nouns to the singular form Stemming: It build the basic forms of words e.g. like, liked, liking, likes, dislike belongs to like
- Stemming: It build the basic forms of words e.g. like, liked, liking, likes, dislike belongs to like

With the rapid growth of information technology, text classification has become one of the key techniques for handling and organizing text data [1]. It plays an important role in the fields of information retrieval, machine learning, natural language processing, data mining and others. So far, there have been a large number of text classification algorithm techniques.

This paper is organized as follows: In Section II, related work of this paper is given. In section III, two principles of classification algorithm are given. They are Naive Bayes classification and Random Forest classification. Section IV describes the experiment discussion. Classifiers are made to classify Movie Reviews in Natural Language Toolkit [2] based on the principles of classification algorithm. It also shows the analysis of the experimental data in this part. Finally, section V draws a conclusion about classification performance of the two classification algorithms in this type of text by comparing accuracy and rate.

2 RELATED WORKS

Data Mining is an interdisciplinary field, the convergence of a set of disciplines, together with database systems, statistics, machine learning, visualization and information science. Depending on the kinds of data to be mined or on the given data mining application, the data mining system may also mix techniques from spatial data analysis, information retrieval, pattern recognition, image analysis, signal processing, computer graphics, web technology, economics, business, bioinformatics or psychology. It is necessary to offer a clear classification of data mining systems, which may support potential users distinguish between such systems and identify those that best match their needs. Data mining system can be considered according the various norms as follows: Classification belongs to supervised learning model. Two steps can be performed in this learning process. First step is a classifier is built to describe a predetermined set of classes or concepts which is called as learning or training model. Classifier is built by any classification algorithms. Classifiers are used to analyse or learn from a training set consisting of database tuples and associated class labels. Second step is to estimate predictive accuracy of the classifier; this is very much used for classification. Estimated accuracy of a classifier on a

- *N Nagajothi is Research Scholar,Department of Computer Science,Alagappa University,Karaikudi and Assistant Professor in Department of Computer Science,Thassim Beevi Abdul Kader College for Women,KilakaraiRamanathapuram District,Tamil Nadu,India.. E-mail: nnjothi@gmail.com*
- *Dr A R Nadira Banu Kamal is Principal , Mohamed Sathak Hamid College of Arts and Science for Wome,Ramanathapuram and Research Guide in Department of Computer Science,Alagappa University, Karaikudi,Tamil Nadu,India,E-mail: nadirakamal@gmail.com*

given test set denotes percentage of test set tuples. This measure denotes that classifier correctly classifying the models. Most of the text classification methods derive from the pattern classification. There are three types of common methods: (1) Based on statistical: such as Naive Bayes, KNN, Support Vector Machine and so on. (2) Based on connection: Artificial neural networks. (3) Rule-based methods: such as decision tree, association rules [3]. In the process of text classification, feature extraction is an essential section. The methods of feature extraction are such as Term Frequency Inverse Document Frequency (TFIDF), information gain (IG), mutual information (MI), statistics, expectation cross-entropy (CE), weight of evidence for text (WET), odds ratio (OR) and so on [4]. As shown in the previous studies [5, 6, 7, 8], the accuracy is the most used evaluation metric in practice either for binary or multi-class classification problems. Through accuracy the quality of produced solution is evaluated based on percentage of correct predictions over total instances. The complement metric of accuracy is error rate which evaluates the produced solution by its percentage of incorrect predictions. Both of these metrics were used frequently by researchers in practice to separate and select the optimal solution. The advantages of accuracy or error rate are, this metric is easy to compute with less complexity; applicable for multi-class and multi-label problems; easy-to-use scoring; and easy to understand by human. In this paper we have used two classification algorithms such as Random Forest which is rule based and Naive Bayes classification algorithm which is based on statistical for evaluating the performance accuracy of movie reviews with the help TFIDF feature extraction method.

3 CLASSIFICATION ALGORITHMS

As demonstrated in this document, the numbering for sections upper case Arabic numerals, then upper case Arabic numerals, separated by periods. Initial paragraphs after the section title are not indented. Only the initial, introductory paragraph has a drop cap.

3.1 Random Forest

Random Forests (RF) [9] algorithm is a famous integrated learning algorithm by taking the decision tree as basic categorizer. The algorithm does not require a priori knowledge, and has high categorization accuracy without over fitting problem. First we briefly discuss the Random Forest [10]. It is composed of a predefined number of binary decision trees. Individual trees in the forest are grown using a bootstrap sample [9] from the training data set. Consider that there are M features associated with each feature vector. During the growth of a random decision tree, a subset of f (f<M) features are randomly chosen in each node and calculate Gini score for determining the best split based on the f features in the training set.

3.2 Naïve Bayes

Naive Bayes algorithm is one of the most active methods in the field of text classification. It is a simple and very effective probabilistic classification method and shows a good performance in some areas [11]. The principle of naive Bayes classification algorithm is: for the given content to be classified, calculating the emergence probability of each label under the conditions of the content features appearing, and then taking the label with largest probability as this content's label. For text classification, it is to see the text feature words. The probability whose features word appears relatively large will match the characteristics corresponding label [12]. The advantages of Bayesian method are efficient and fast. Naive Bayes is regarded as a good classification method, but consuming too much time and space.

4 IMPLEMENTATION AND RESULTS

These two classifiers have been designed in Natural Language Tool Kit (NLTK) based on the principles of categories mentioned above. They are Naïve Bayes and Random Forest classifier. The testing data is the text of movie reviews in NLTK. We have used 2000 text documents of movie reviews for analysis. From this set of documents 20% is used as testing data set and 80% is used as training data set. Testing and training data set has been well categorized. Before applying above said classifiers to the data set, pre-processing has been done. First step of pre-processing is tokenization. Tokenization is a process of splitting an input text into meaningful chunks and that chunk is actually called token. This token is a useful unit for semantic processing. We may want the same token for different forms of the word. So process of token normalization performed here. This is the second step. The process of normalizing the words is called stemming and lemmatization. Stemming is a process of eliminating and changing suffixes to get to the root form of the word, which is called the stem. Lemmatization means to doing things correctly with the use of a vocabulary and morphological analysis and returns the base or dictionary form of a word which is known as the lemma. It uses the WordNet database to lookup lemmas. Stemming process fails on irregular forms and produces non-words. So we have used lemmatization for token normalization. After tokenization process, next one is to transform tokens into features which are called feature extraction. The best way to do that is bag of words. We have to count occurrences of a particular token in our text. And for each token we introduce a new feature or column that particular text. We actually replace the text with a huge vector of numbers and each dimension of that vector corresponds to a certain token in our data set. This process is called text vectorization. Here we can actually come up with a huge number that can exponentially grow with the number of consecutive words that we want to analyse. We want to overcome that problem; we can actually remove some n-grams. We removed n-grams from features based on their occurrence frequency in documents of our corpus. High frequency n-grams which are called stop words that won't help us to discriminate texts and we removed them. Low frequency n-grams are typos which are typographical mistakes that also removed. Medium frequency n-grams are good frequency n-grams. We found the problem is there are a lot of medium frequency n-grams. So we introduced some notions which are term frequency (TF), inverse document frequency (IDF) and Term frequency-inverse document frequency (TF-IDF).

Term Frequency is the frequency for term t. The term is an n-gram, token or anything like that in a document d. We have options how we can count that term frequency. It shows in Table 1.

TABLE 1TERM FREQUENCY OPTIONS

Weighting Scheme	TF Weight
Binary	0,1
Raw count	$f_{t,d}$
Term Frequency	$f_{t,d} / \sum f_{t',d}$ 'ted
Log normalization	$1 + \log(f_{t,d})$

Term Frequency is calculated as

TF =No. of Occurrences of a word / Total words in the document

Inverse Document Frequency (IDF) which measures how important a term is. Following equation is used to weigh down the frequent terms.

IDF=log(Total no. of documents) / No. of documents containing the word

Information retrieval and text mining are using Tf-idf which is a numerical statistic. This numerical statistic is also called as weight used to assess how important a word is to a document in a corpus. It is calculated as follows:

$$Tfidf(t,d,D) = tf(t,d) \cdot idf(t,D)$$

Precision and recall are two common methods to measure the efficiency of the classifier. Precision is if we classify a random document d_i under category c_i, the probability that is true. Recall is that if a document di belongs to a category c_i, the probability that the classifier makes such a decision.

After pre-processing, we have applied Naïve Bayes classifier to the movie review data set. Table 2 shows performance measures of Naïve Bayes classifier for movie review documents.

Table 3 shows the results obtained by Random Forest classifier for the same data set. We have calculated precision, recall and f1 score value for evaluating the accuracy.

TABLE 2. PERFORMANCE MEASURES OF NAIVE BAYES CLASSIFIER

Class Labels	Precision %	Recall %	f1-score %	Support
0	76	83	79	286
1	83	77	80	314
avg / total	80	80	80	600
Accuracy :79.5 %				

TABLE 2. PERFORMANCE MEASURES OF NAIVE BAYES CLASSIFIER

Class Labels	Precision %	Recall %	f1-score %	Support
0	76	83	79	286
1	83	77	80	314
avg / total	80	80	80	600
Accuracy :79.5 %				

TABLE 3 PERFORMANCE MEASURES OF RANDOM FOREST CLASSIFIER

Class Labels	Precision %	Recall %	f1-score %	support
0	86	87	86	208
1	85	84	85	192
avg /total	85	85	85	400
Accuracy :85.500				

We have obtained precision and recall as 85% for random forest and precision and recall as 80% for Naive Bayes. From this result, we have found that Random Forest classifier gives more performance accuracy than Naive Bayes classifier for sentimental analysis of movie reviews. The overall accuracy of RF classifier is 85.5% which is higher than the Naive Bayes classifier. Figure 1: shows the performance evaluation by Random Forest classifier. In this experiment,we found that overall accuracy of Naive Bayes classifier is 79.5% in terms of precision,recall and f-score.Figure 2: shows the performance measures by Naive Bayes classifier.

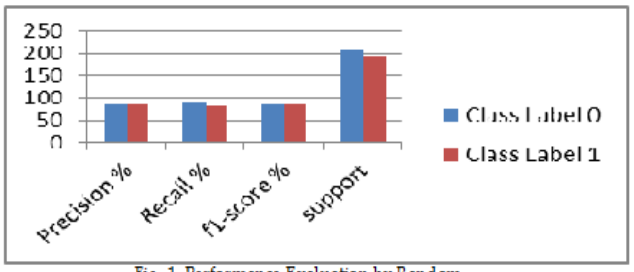

Fig. 1. Performance Evaluation by Random Forest Classifier

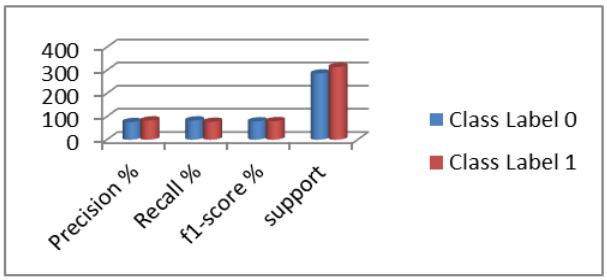

Fig. 2. Performance Evaluation by Naive Bayes Classifier

5 CONCLUSION

Text classification is one of the most commonly used natural language processing tasks. In this paper, we have taken the data set from movie review. The text documents must pass through some set of steps such as tokenization, normalization of tokens, vectorization, term frequency, inverse document frequency and tfidf calculation and finally we applied two classifiers such as Naive Bayes which is based on statistical and Random Forest which is based on rule to same training and test data set. We evaluated the performance accuracy in terms of precision, recall and f1-score using Naive Bayes and Random forest classification methods for sentimental analysis of movie reviews. No single method has been found to be the superior over all others for all data sets. But we found that Random forest classifier gives more accuracy than the Naive Bayes for our data set.

REFERENCES

[1] F. Sebastian, "Machine learning in automated text categorization," ACM Computing Surveys, 34(1), pp. 1-47, 2002.
[2] Information on http://www.nltk.org/# natural-language-toolkit.
[3] ZHU Zhen-fang, LIU Pei-Yu, Lu Ran," Research of Text Classification Technology Based on Genetic Annealing

Algorithm," International Symposium on Computational Intelligence and Design, pp. 265-269, 2008.

[4] Zhou Faguo, Zhang Fan, Yang Bingru, Yu Xingang. "Research on Short Text Classification Algorithm Based on Statistics and Rules," Third International Symposium on Electronic Commerce and Security, pp. 3-7, 2010.

[5] N.V. Chawla, N. Japkowicz and A. Kolcz, "Editorial: Special issue on learning from imbalanced data sets," SIGKDD Explorations, 6, pp. 1-6, 2004.

[6] Q. Gu, L. Zhu and Z. Cai, "Evaluation Measures of the Classification Performance of Imbalanced Datasets, " in Z. Cai et al. (Eds.) ISICA 2009, CCIS 51. Berlin, Heidelberg: Springer-Verlag, pp. 461-471, 2009.

[7] M. Hossin, M. N. Sulaiman, A. Mustapha, N. Mustapha and R. W. Rahmat, "A Hybrid Evaluation Metric for Optimizing Classifier, " in Data Mining and Optimization (DMO), 2011 3rd Conference on, pp. 165-170, 2011.

[8] R. Ranawana and V. Palade, "Optimized precision-A new measure for classifier performance evaluation," in Proc. of the IEEE World Congress on Evolutionary Computation (CEC 2006), pp. 2254-2261, 2006

[9] Dong Shishi, Huang Zhexue. A Brief "Theoretical Overview of Random Forests," Integrated Technologies, 2(1), pp. 1-7, 2013

[10] L. Breiman, "Random forests," Machine learning, vol. 45, no. 1, pp. 5–32, 2001.

[11] He Youquan, Xie jianfang and Xu cheng, "An improved Naive Bayesian algorithm for Web page text classification," Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD) pp. 1765-1768, 2011.

[12] Yaguang Wang, Wenlong Fu, Aina Sui, Yuqing Ding, "Comparison of Four Text Classifiers on Movie Reviews, "3rd International Conference on Applied Computing and Information Technology / 2nd International Conference on Computational Science and Intelligence, pp. 495-498, 2015.

[13] S.Raman, V.Kuma,S.Venkatesan, "Performance Comparison of Various Information Retrieval Models Used in Search Engines, " IEEE Conference on Communication, Information and Computing Technology, Mumbai, India, pp. 78-89, 2012.

[14] K. R. Bindu, L. Parameswaran, K. V. Soumya, "Performance Evaluation of topic Modelling Algorithms with an application of Q & A Dataset, "International Journal of Applied Engineering Research, vol. 10, pp. 23-27, 2015.