

News Classification: A Data Mining Approach

Dipak Ramchandra Kawade^{1*} and Kavita S. Oza²

¹Department of Computer Science, Sangola College, Sangola – 413307, Maharashtra, India; dipakkavade@gmail.com

²Department of Computer Science, Shivaji University, Kolhapur – 416004, Maharashtra, India; skavita.oza@gmail.com

Abstract

Objectives: Text classification is one of the important applications of data mining. Text classification classifies text documents on the basis of words, phrases, combination of words etc. into predefined class labels. **Method/Analysis:** Present study classifies news data into four predefined classes namely Business, Entertainment, sports and Technology. For text classification WEKA an open source data mining tool is used. Different classification algorithms are applied on News data set. A comparative study of these algorithms is done based on Accuracy, Time, Errors and ROC to predict the best algorithm for news data set classification. **Findings:** Present study analyzed result on the basis of accuracy, time, error and ROC curve. Present work concludes that NaïveBayes Multinomial algorithm is best for news classification.

Keywords: Classification Algorithms, Data Mining, Text Classification, WEKA

1. Introduction

In this digital era, numbers of digital documents are grown rapidly with ease of availability. This is possible due to high capacity of hardware, software and powerful computing and storage device which are easily available in affordable prices.

Data mining is useful for extracting or discovering new relation, hidden knowledge and important patterns from huge amount of data¹. Data mining is also known as Knowledge Discovery in Databases (KDD). Data mining uses different technique for knowledge discovery such as classification, clustering, summarization, associations etc.

Text mining is one of technique used in data mining for analysis of large volume of textual data. Analysis helps in discovering useful patterns and extracts information². Text classification is one of the key technologies used in text mining. It is useful in various applications such as automatic indexing of research and scientific articles, tracking customers email for Customer Relationship Management (CRM), categorizing web documents, detection and identification of terrorist activity etc.³. It is useful

for document filtering, information extraction, machine translation, language identification and document classification.

Vast availability of any type of documents needs its effective organization and retrieval. To have easy access to these documents these documents are classified using text classification techniques. It is one of the important techniques for categorization of documents in supervised way³.

Present study uses text classification technique for classification of news. In this study News data is classified as per the types of news such as Business, Sports, Entertainment and technology. Text classification is useful to assign unlabeled class document to predefined class. This technique works in two stages. In first stage, it can extract future terms or effective keywords which are useful for identifying class in training phase. In next stage i.e. testing phase actual classification of document is carried out using future terms of keywords. For effectiveness and efficiency purpose these documents are pre-processed.

In² discusses on automatic classification of text documents using machine learning techniques. Due to automatic classification; extensive saving will be done

*Author for correspondence

on different factors such as labour power, portability of documents etc. Application domain of automatic text classification are various and important in digital medium. It is also essential for many applications where number of documents might be classified within short period of time. Article discusses importance of automatic classification by improving productivity of human classification problems. This technique increases effectiveness and efficiency of automatic text classification technique.

Text is classified using keyword extraction technique^{3,8}. The data is pre-processed by removing stop words and used stemmed techniques. After pre-processing frequency for each term in text document is calculated and TF-IDF is found. Experiments have shown that decision tree algorithm gives higher accuracy.

Automatically extracted keyword technique is used to improve text categorization and also to identify impact of keywords on text categorization⁴. Predictions are made on the basis of keywords and are distinguished in either unigrams or intact. Higher weights are assigned to those keywords which are full text and extracted keywords also. This study shows that unigram's performance is higher. But this experiment does not consider keywordness technique which may increase performance.

An efficient framework for identification of knowledge from text categorization by using supervised learning technique has been developed earlier⁵. Experimental result shows that background lexical information with supervised learning gives better performance than lexicon and unlabelled data.

WEKA is used for classification of text data which is in the form of SMS dataset⁶. Different WEKA algorithms are used on SMS dataset and analysis these results are presented. The analysis of result on the basis of accuracy, time and errors is also presented. Experimental result of the study shows that Filtered Classifier with Naïve Bayes algorithm is good for SMS spam dataset.

In⁷ have considered evolution of case based reasoning technique for long text to short text. For these purpose appropriate future type id determined and also identified representation of short message and then compare the performance of the algorithm. For classification purpose, Naive Bayes and support vector machine algorithms are used.

According to¹⁴, user can classify papers on the basis of journal subject with accuracy rate of 100%. This article also discusses difference between KSCI and SCI.

2. Experimental Work

2.1 Working Environment

Present study uses environment with Intel Core i3, 3.3 GHz processor with 2 GB RAM and Windows XP operating system. This experiment is carried out in WEKA tool.

2.2 Dataset

All the data required for present study was collected between 20-07-2015 to 05-08-2015. These data were collected from Times of India, Indian Express and The Hindu websites. In the data set, present study considers four different types of news namely Business, Entertainment, Technology and Sports. All these news data are stored in text file with specific domain as mentioned above. Final dataset consist of total 483 news distributes shown in Table 1.

2.3 WEKA-data Mining Tool

WEKA is an open source data mining tool developed by University of Waikato in New Zealand. Java Language is used to develop this tool. This tool consists of number of data mining algorithms and different filters for data preprocessing. For simplicity and ease of use, these algorithms are grouped on the basis of rules generated by algorithm. This tool is very handy and simple to use and also it is freely available on the internet. Due to these reason, we use WEKA for present study. In present study we used different classification algorithms for Text classification⁹.

2.4 Data Pre-processing

Collected data is not suitable for experimental work. We need pre-process these data and make clean and suitable data for experimental work. For present study all news are stores in folder named as news. By using TextDirectoryLoader class of WEKA, present study load data for pre-processing. When data is loaded in WEKA, it shows two attributes namely text and class. For text classification, it is needed to convert text attribute into collection of words i.e. word vector. For this purpose, WEKA's StringToWordVector unsupervised filter is used. This filter convert string attribute into set of attributes rep-

Table 1. Text document contains in each class

| Business | Entertainment | Technology | Sports |
|----------|---------------|------------|--------|
| 101 | 115 | 110 | 157 |

representing word based on which tokenizer used. Proposed work uses WordTokenizer with delimiters as „;”()?!. Also we set lower case token to true which will convert all word in lower case latter. After applying this filter on the dataset, text attribute was converts into 2946 attributes with 483 instances. For efficiency purpose, proposed study removes stop words and numeric values. Stop words such as is, the, an, a, these etc. are not useful for text classification. For better efficiency eliminate frequent usage words such as conjunctions, numbers, prepositions, names, base verbs, etc. these types of words have poor characterizing power so it has useless for the text classification. Also there is no effect of any numeric values in text classification. Therefore, numeric values are also removed. For removing these values, present work uses unsupervised Remove filter. After removing these attributes, 2768 attributes are available for experiment purpose.

2.5 Text Classification Process

Text Classification has been done on the basis of words, phrases and word combinations with respect to set of pre-defined class labels. Text Classification processes consists of training phase and testing phase. In training phase, dataset is loaded and different classification algorithms are applied on this dataset. After finishing the training phase, performance of classifiers is analyzed and the classifier with the best performance is selected. This can be chosen on the basis of different factors such as accuracy, time and error rate etc.¹⁰.

2.6 Experimental Result

Present study uses WEKA's classification algorithm namely J48, NaïveBayesMultinomial, MultiClass

Classifier(Logistic), SMO and HyperPipes, JRip for text classification purpose. Experimental result of said algorithms are summarizing in following tables.

Present study use “10-fold cross-validation” as a Test mode for all algorithms and “full training set” as classifier model. In 10-fold classification method, dataset is divided into 10 folds. Each fold is used as test dataset while other nine folds are used as training dataset. In each fold and test set, performance is measure for different case based configuration⁶.

Table 1 shows result based on accuracy and time taken to build model for each algorithm. Table 2 shows different error rates obtained by algorithms. Table 3 displays different errors gain by different algorithm and Table 4 display ROC curve information values obtained by different algorithm with respect to different class labels. Figure 1 shows accuracy and time obtained by each algorithm. Figure 2 represent different errors obtained by algorithms. Figure 3 shows average ROC values for each class obtained by

Table 2. Table of accuracy and time

| Algorithm | Correct Instance | Percentage of Correct Instance | Time Requires to Build Model (in Second) |
|---------------------------------|------------------|--------------------------------|--|
| NaiveBayesMultinomial | 448 | 92.7536 | 0.03 |
| J48 | 364 | 75.3623 | 26.46 |
| JRip | 360 | 74.5342 | 66.92 |
| MultiClassClassifier (Logistic) | 443 | 91.7184 | 649.59 |
| SMO | 434 | 89.8551 | 0.46 |
| HyperPipes | 447 | 92.5466 | 0.14 |

Table 3. Simulation of error rate

| Algorithm | Kappa statistic | Mean absolute error | Root mean squared error | Relative absolute error (in %) | Root relative squared error (in %) |
|---------------------------------|-----------------|---------------------|-------------------------|--------------------------------|------------------------------------|
| NaiveBayesMultinomial | 0.9023 | 0.0375 | 0.1837 | 10.1164 | 42.6576 |
| J48 | 0.6669 | 0.1328 | 0.3365 | 35.7817 | 78.1145 |
| JRip | 0.6518 | 0.1651 | 0.3245 | 44.4967 | 75.3316 |
| MultiClassClassifier (Logistic) | 0.8884 | 0.3126 | 0.3631 | 84.2523 | 84.3093 |
| SMO | 0.8634 | 0.2597 | 0.3265 | 69.9752 | 75.81 |
| HyperPipes | 0.8994 | 0.3744 | 0.4323 | 100.894 | 100.3681 |

Table 4. ROC curve values

| Algorithm | Business | Entertainment | Sports | Technology | Average Value |
|---------------------------------|----------|---------------|--------|------------|---------------|
| NaiveBayesMultinomial | 0.977 | 1 | 0.999 | 0.981 | 0.99 |
| J48 | 0.82 | 0.913 | 0.915 | 0.765 | 0.85 |
| JRip | 0.846 | 0.874 | 0.89 | 0.84 | 0.86 |
| MultiClassClassifier (Logistic) | 0.962 | 0.998 | 0.998 | 0.971 | 0.98 |
| SMO | 0.947 | 0.979 | 0.991 | 0.919 | 0.96 |
| HyperPipes | 0.936 | 0.988 | 0.991 | 0.946 | 0.97 |

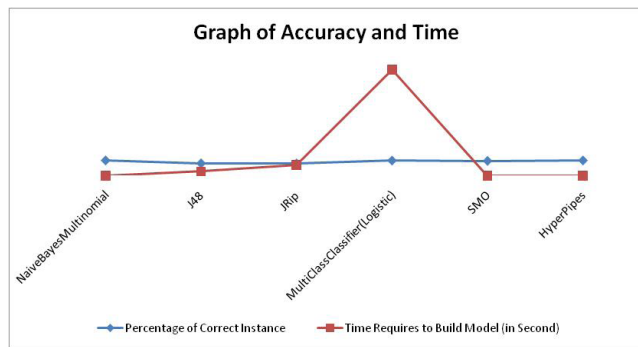


Figure 1. Accuracy and time obtained by algorithm.

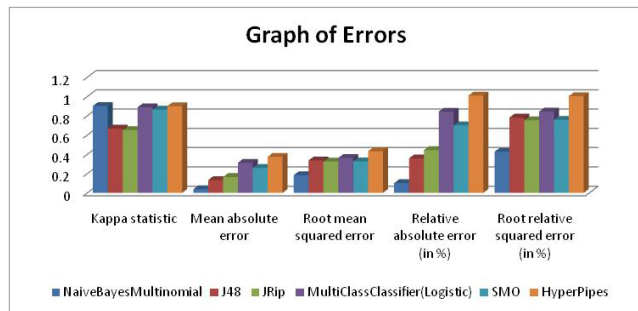


Figure 2. Errors obtained by algorithm.

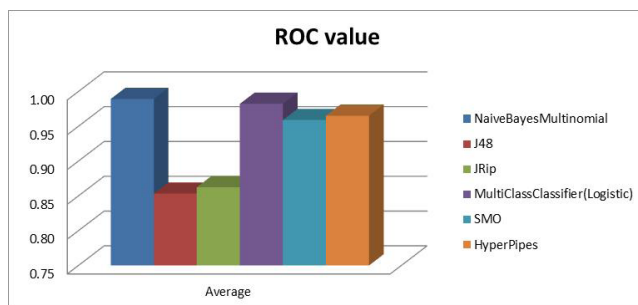


Figure 3. Average ROC value for each algorithm.

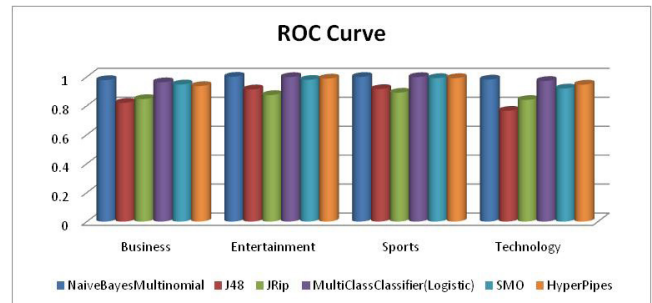


Figure 4. ROC value for each class.

different algorithm. Figure 4 shows ROC values for each class obtained algorithm.

3. Observations

Tables 2-4 shows summarized result of text classification on news dataset. Present study uses J48, NaiveBayesMultinomial, MultiClassClassifier(Logistic), SMO and HyperPipes, JRip algorithms from WEKA. These algorithms are selected randomly from each group. Figures 1-3 shows graphical representation of Table 2-4.

Table 2 and Figure 1 shows result based on accuracy and it also display time, in second, required to build model. Table 2 shows correctly classified instances out of totally 483 instances. It also shows percentage of correctly classified instance. From this table it is clearly observed that NaiveBayesMultinomial algorithm gains highest accuracy among all other algorithm followed by HyperPipes and MultiClassClassifier(Logistic). These three algorithms' accuracies are very close (i.e. nearly equal to 92%) to each other. Therefore, on the basis of accuracy; for news classification dataset, these three algorithms work best. J48 and JRip algorithms are very poor accuracy as compare to these algorithms. If we consider time factor; then Table 2 and Figure

1; it is clearly observed that NaiveBayesMultinomial require very less time i.e. 0.03 seconds to build model. HyperPipes requires little more time i.e. 0.14 seconds to build model. From Table 2 and Figure 1; present study observed that NaiveBayesMultinomial algorithm is best for news dataset take in present paper and HyperPipes is secondrunnerup algorithm for this dataset. As MultiClassClassifier(Logistic) accuracy is good but it requires more time i.e. 649.59 seconds. Therefore on t,he basis of time MultiClassClassifier(Logistic) algorithm is not suitable for present dataset.

Kappa statistics are useful to differentiate between reliability of data collected and validity of the data. It is used to access particular measuring cases¹¹. The average Kappa score from the selected algorithms is around 0.812033. According to Kappa Statistic, the accuracy of this classification purposes is substantial¹¹. Highest value of Kappa statistics is 0.9023 which is near to 1 for NaiveBayesMultinomial algorithm. Therefore, according to statement of Kappa statistics NaiveBayesMultinomial algorithm is best for present news dataset.

Table 3 shows a very commonly used error indicator for classification techniques which are mean of absolute errors, root mean squared errors and the relative errors. It is discovered that the highest error is found in HyperPipes with an average score of around 0.744 followed by MultiClassClassifier(Logistic) then SMO. JRip and J48 algorithms error rates are nearly equal i.e. 0.46. But NaiveBayesMultinomial algorithm has very less error rate which is 0.33. An algorithm which has a lower error rate will be preferred as it has more powerful classification capability [SMS paper]. Here in this case; from Figure 2 and Table 3; NaiveBayesMultinomial is less error rate. On the basis of accuracy, time and error rate, NaiveBayesMultinomial algorithm is best suited for News classification purpose.

Table 4 shows ROC value obtained for each class for different algorithm. From experimental study it is observed that when value of ROC is near to 1 for specific class; it means that misclassification for that class is less^{12,13}. From Table 4 and Figure 4 it is clearly observed that NaiveBayesMultinomial algorithm's average ROC value is 0.99 which is approximately equal to 1. It shows that misclassification in this algorithm is very less.

From Tables 2-4 and Figures 1-4, present study shows that NaiveBayesMultinomial algorithm works well for news classification technique. NaiveBayesMultinomial algorithm has highest accuracy, less time and errors and having highest value of ROC.

4. Conclusions

Text classification is one of the important applications of data mining technique. Present study uses text classification technique for classification of news. This news dataset is created from Times of India, Indian Express and The Hindu newspaper websites. This dataset consists of four types of news namely Business, Entertainment, Technology and Sports. For classification of news; present study uses WEKA data mining tool which is open source tool. For present work J48, NaiveBayesMultinomial, MultiClassClassifier(Logistic), SMO and HyperPipes, JRip algorithms from WEKA are used. In this work we apply these algorithms on news dataset. These datasets are in text form. For experimental purpose, the present study uses WEKA data mining tool. Experimental result is analyzed on the basis of accuracy, time, error and ROC curve. Analysis of present work shows that NaiveBayesMultinomial algorithm is best for news classification. NaiveBayesMultinomial has highest accuracy i.e. 92.7536 in 0.03 seconds. Also value of Kappa statistics for this algorithm is near to 1. Similarly, average error rate for this algorithm is less i.e. 0.33 and average ROC value for NaiveBayesMultinomial algorithm is 0.99.

5. References

1. Han J, Kamber M. Data mining - concepts and techniques. Third Edition, India
2. Sebastiani F. Machine learning in automated text categorization. *Journal ACM Computing Surveys*. 2002 Mar; 34(1):1-47.
3. Menaka S, Radha N. Text classification using keyword extraction technique. *International Journal of Advanced Research in Computer Science and Software Engineering*. 2013 Dec; 3(12).
4. Hulth A, Megyesi BBA study on automatically extracted keywords in text categorization. *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, Sydney*; 2006 Jul. p. 537-44.
5. Melville P, Gryc W, Lawrence RD. Sentiment analysis of blogs by combining lexical knowledge with text classification. *KDD '09 Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data mining*; 2009. p. 1275-84.
6. Kawade DR, Oza KS. SMS spam classification using WEKA. *International Journal of Electronics Communication and Computer Technology*. 2015 May; 2015. p. 5.
7. Delany, SJ, Cunningham P. An analysis of case-base editing in a spam filtering system. *Advances in Case-based Reasoning*. Springer Berlin Heidelberg; 2004. p. 128-41.

8. Li-gong Y, Jian Z, Shi-ping T. Keywords extraction based on text classification. Proceedings of the 2nd International Conference On Systems Engineering and Modeling (ICSEM-13); 2013.
9. WEKA [Internet]. [cited 2015 Aug 21]. Available from: <http://www.cs.waikato.ac.nz/~ml/weka>.
10. Junaid MB, Farooq M. Using evolutionary learning classifiers to do mobile spam (SMS) filtering. GECCO'11, Dublin, Ireland; 2011 Jul.
11. Kappa [Internet]. [cited 2015 Oct 12]. Available from: <http://www.dmi.columbia.edu/homepages/chuangj/kappa>.
12. Available from: <http://gim.unmc.edu/dxtests/roc3.htm>.
13. Available from: https://en.wikipedia.org/wiki/Receiver_operating_characteristic.
14. Kang MY, Shin J-D, Kim B. Automatic subject classification of korean journals based on KSCD.