A Comparative Analysis Of News Categorization Using Machine Learning Approaches

Nabamita Deb, Vishesh Jha, Alok K Panjiyar, Roshan Kr Gupta

Abstract: The rapid growth of print and digital media increased the reach of one and all in terms of information, resulting in more amount of Text data to be mined. This data is nothing but a heap of unclassified information which when kept together means nothing. This means that there is a need to tag all these data i.e. News Classification. News classification is the task of automatically classify the news documents into their predefined classes based on their content with the confidence learned from the training news dataset. This research evaluates some most widely used machine learning techniques, mainly Naive Bayes, Random Forest, Decision Tree, SVM and Neural Networks, for automatic news classification problem. To experiment the system, a dataset from BBC that have two columns, one has the news headlines and the other contains the type it belongs to. There are 2225 rows in the data set is used.

Index Terms: Machine Learning, News classification, Naive Bayes, Support Vector Machine, Neural Networks, Random Forest, Decision Tree.

1. INTRODUCTION

THE news world is bustling every second with news entering from all sources. There are multiple news channels, online news portals that let out the daily proceeding every minute. Different types of news make it to these portals. Whether it is print media or electronic media, news stories are important and flowing everywhere. So, it is important to have an efficient system of segregating news into different categories. Technology can be used to enhance and better this system by the proper use of machine learning . The news headlines will be used to train the model and later the machine will be able to predict the category of the news item very fast and accurate. This will be helpful for all news channels and apps as it will give them an efficient and speedy way to do their job. Large data can be segregated easily which is a very good thing for them. We have a dataset from BBC that have two columns, one has the news headlines and the other contains the type it belongs to. There are 2225 rows in the data set. In this work, five supervised machine learning algorithms, Naive Bayes, Support Vector Machine, Neural Networks, Random Forest and Decision Tree with the same features are compared in terms of their accuracy.

2 RELATED STUDIES AND INTRODUCTORY CONCEPT

The main task of news classification is to automatically classify the news documents into their predefined classes based on their content. Many machine learning techniques has been developed for classification of news. Classification is a challenging task in the field of text mining as it requires preprocessing steps to prepare the textual data into structured form which is initially available in unstructured form. Classification requires pre-processing steps to convert the data into structured form from the unstructured form. The different steps involve in news classification are collection of news, preprocessing of collected news, feature selection , different classification techniques to classify news and evaluating performance measure for different classification technique.

2.1News Collection

In this work, we used a dataset from BBC that have two columns, one has the news headlines and the other contains the type it belongs to. There are 2225 rows in the data set.

2.2 News Pre-processing

After news collection, pre-processing is done as this information is originating from variety of sources and its cleaning is required so that it could be free from corrupted file. Information after pre-processing should be separated from random words like semicolon, double quotes, full stop, bracket, special characters and so on and also the information is converted to lower case character to get better outcomes. Information is made free from those words which show up generally in content and are known as stop words

.

2.3 Feature Selection

The feature selections in this research are done by using TF-IDF. Term Frequency-Inverse Document Frequency (TF-IDF) [10] is a very common algorithm which is used to transform text into a meaningful representation of numbers.TF-IDF can be used for stop-words filtering in various subject fields including text summarization and classification.

2.4 News Headlines Classification

The next most important phase after feature selection are classification where the news headlines are classified with an aim to assign to their respective classes. The most common news headlines classification methods used in research up till now are Naïve Bayes, Support Vector Machine, Neural Network, Random Forest and Decision Tree.

NAÏVE BAYES

Naive Bayes classifier is a classification algorithm which is based on Bayes' Theorem. Instead of a single algorithm, it is a family of algorithms where all of them share a common principle, where every pair of features being classified is independent of each other. The Naïve Bayes classifier is a simplest approaches to the classification task that is still

- *Nabamita Deb is a faculty in the Department of IT,Gauhati University deb.nabamita@mail.com*
- *Vishesh Jha, Alok K Panjiyar, are bachelor degree graduates in Information Technology in Gauhati University, E-mail: jha.vishesh@gmail.com ,panjiyar.alok.26@gmail.com*
- *Roshan Kr Gupta is currently pursuing bachalorsdegree program in information technology e in Gauhati University E-mail: roshandec252@gmail.com*

capable of providing reasonable accuracy. It is a probabilistic classifier which is based on probability models that incorporate strong independence assumptions.

Support Vector Machine

A Support Vector Machine (SVM) is an algorithm of supervised learning which is used for fast and dependable classification that performs very well with a limited amount of data. But, if the size of text document is large then there will be a number of dimensions in hyper-space which may increase computational cost of the process.

Neural Networks

Artificial neural-networks work on the concept of human brain consisting neurons. It consists of a layered arrangement of neurons where the input vectors are converted into the some form of output. ANN is considered to be a good classifier because it can better handle multiple categories and work well on it. It supports fast testing phase. In this research, Multilayer Perceptron Neural Network with 20 Hidden Layer is used for the classification learning and prediction.

Random Forest

Random forest algorithm is a supervised classification algorithm which creates the forest with a number of trees. it is also one of the most used algorithm because of its simplicity and also for the fact that it can be used for both classification and regression problem. Random Forest has nearly the same hyper parameters as a decision tree and it builds multiple decision trees and merges them together to get a more accurate and stable prediction.

Decision Tree

Decision Tree classifier is an algorithm which belongs to the family of supervised learning algorithms which can be used for solving regression and classification problems too.. It is represented in a tree form of structure where the branches of tree represent weight and each leaf is a different class. The main aim of using Decision Tree is to create a training model which can be used to predict class or value of target variables by learning decision rules inferred from prior data (training data)

3. RELATED WORK

Various techniques have been proposed to classify the text: Neural Network , Decision Trees , Support Vector Machines, Naïve Bayes, and Random forest. The basic concept of these techniques is the classification of news type using the trained classifier that can automatically predict an incoming news type to some of the predefined classes. In this paper, review of classification of news on basis of their headlines is performed. Variety of news headlines classifications work have been taken place in the past, few of them include emotions classification on basis of news headlines, financial news classification [1], automatic news headlines classification [2], news headlines classification using N-gram model [3], classification of news headlines for providing user centered e-newspaper [4], emotions extraction from news headlines [5], and short news headlines classification of twitter [6]. Seyyed M. H. Dadgar et al. (2016) [8] point is to group news to various classes, utilizing SVM order procedure. In this work, they utilized TF-IDF and SVM classifier. They utilized the procedure of text preprocessing, feature extraction based on TF-IDF and toward the end grouping by utilizing SVM classifier. In this work, they utilize two unique datasets of BBC news and 20newsgroup and assessed their outcome and their accuracy was 97.84% and 94.93% for both the datasets respectively. Sandeep Kaur et al. (2016) [9] proposed a method of classification of online news using neural networks which increased the accuracy of the classification up to 99%.

4.EXPERIMENTAL SETUP

The news classification system pipeline is given in Figure 1. It consists of Preprocessing, Feature Extraction, Machine Learning Algorithm for Classification, Training of Classifier, Test Classifier with Trained Model and Evaluation phase.

5.

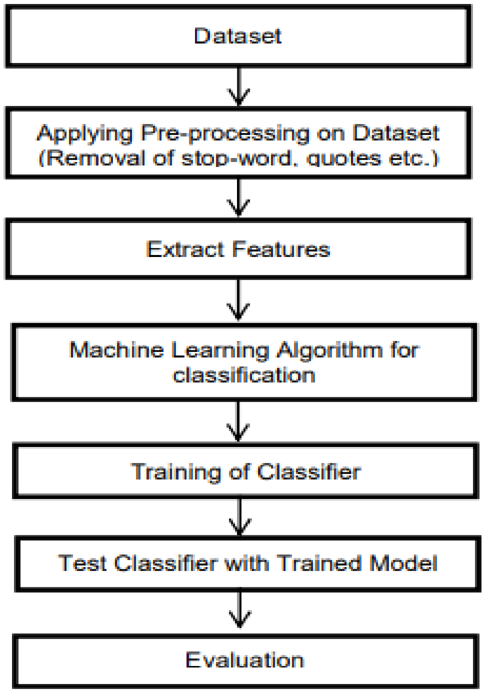

Dataset
Applying Pre-processing on Dataset (Removal of stop-word, quotes etc.)
Extract Features
Machine Learning Algorithm for classification
Training of Classifier
Test Classifier with Trained Model
Evaluation

Fig1: News Classification System Pipeline

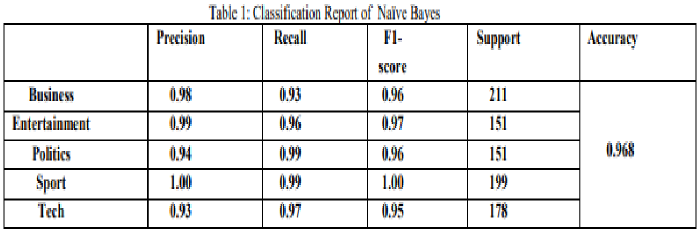
Evaluation And Result

The experimental result was analyzed for three evaluation parameters: Accuracy, Precision and Recall. Accuracy: Accuracy tell us how comfortable the model is with detecting the positive and negative class. Precision: Precision states us about the success probability of making a correct positive class classification. Recall: Recall states how sensitive the model is towards identifying the positive class.

5.1 Evaluation Using Naïve Bayes

The Confusion Matrix and the Classification Report obtained from Naïve Bayes are described below.

Table 1: Classification Report of Naïve Bayes



	Precision	Recall	F1-score	Support	Accuracy
Business	0.98	0.93	0.96	211	0.968
Entertainment	0.99	0.96	0.97	151	
Politics	0.94	0.99	0.96	151	
Sport	1.00	0.99	1.00	199	
Tech	0.93	0.97	0.95	178	

Confusion Matrix:

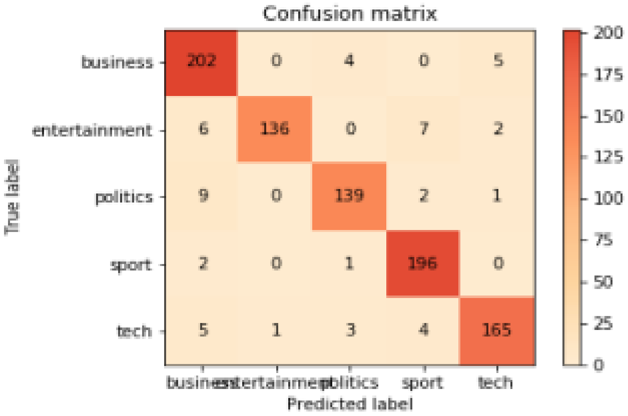


Fig 2: Confusion Matrix of Naïve Bayes Classifier

5.2 Evaluation Using Support Vector Machine

The Confusion Matrix and the Classification Report obtained from Support Vector Machine are described below.

Table 2: Classification Report of Support Vector Machine

	Precision	Recall	F1-score	Support	Accuracy
Business	0.95	0.95	0.95	211	0.964
Entertainment	0.97	0.98	0.98	151	
Politics	0.96	0.93	0.95	151	
Sport	0.98	0.99	0.99	199	
Tech	0.96	0.96	0.96	178	

Confusion Matrix:

5.3 Evaluation Using Multilayer Perceptron (MLP) Neural Network

The Confusion Matrix and the Classification Report obtained from Multilayer Perceptron Neural Network with 20 hidden layers are described below.

Confusion Matrix:

5.4 Evaluation Using Random Forest

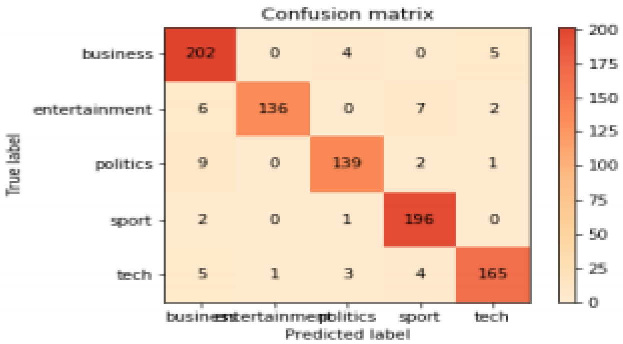


Fig 4: Confusion Matrix of Multilayer Perceptron (MLP) Neural Network

The Confusion Matrix and the Classification Report obtained from Random Forest are described below.

Confusion Matrix:

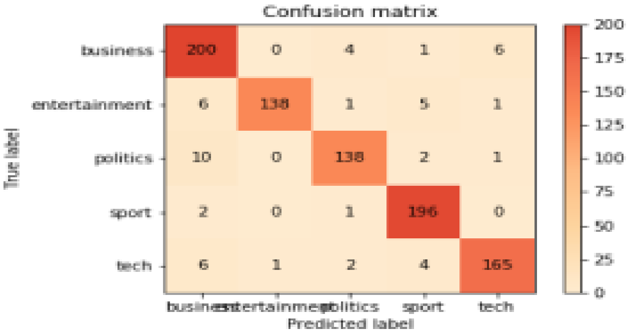


Fig 5: Confusion Matrix of Random Forest

5.5 Evaluation Using Decision Tree

The Confusion Matrix and the Classification Report obtained from Decision Tree are described below.

Table 5: Classification Report of Decision Tree

	Precision	Recall	F1-score	Support	Accuracy
Business	0.75	0.85	0.80	211	0.832
Entertainment	0.87	0.77	0.81	151	
Politics	0.78	0.79	0.78	151	
Sport	0.91	0.88	0.90	199	
Tech	0.88	0.85	0.87	178	

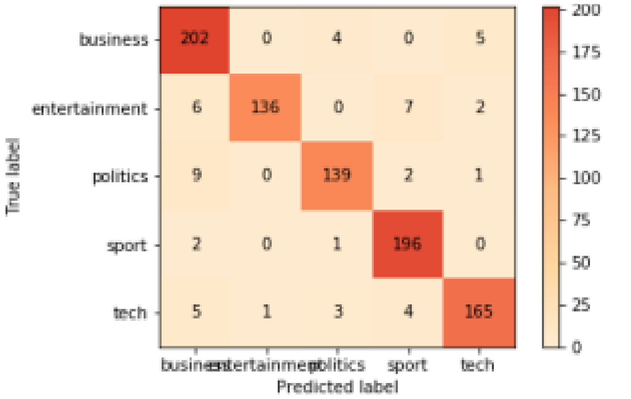


Fig 3: Confusion Matrix of Support Vector Machine Classifier

Table 3: Classification Report of Multilayer Perceptron (MLP) Neural Network

	Precision	Recall	F1-score	Support	Accuracy
Business	0.94	0.95	0.95	211	0.964
Entertainment	0.97	0.97	0.97	151	
Politics	0.97	0.94	0.96	151	
Sport	0.98	0.99	0.98	199	
Tech	0.96	0.96	0.96	178	

Confusion Matrix:

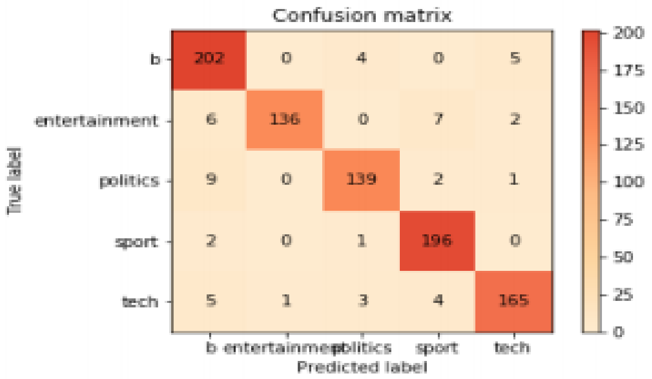


Fig 6: Confusion Matrix of Decision Tree

RESULTS

The average results show that the Naive Bayes is performing better than the other four algorithms with the classification accuracy of 96.8 % .Then follows the Random Forest with accuracy 94.1 % , Support Vector Machine (SVM) with accuracy 96.4 %, Neural Networks with accuracy 96.4 % and the Decision Tree with accuracy 83.2 % as shown in fig-7.

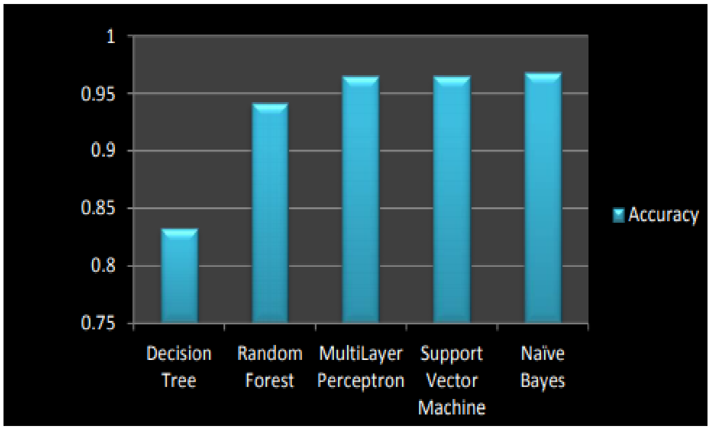


Fig 7: Different Classifier Algorithm Vs Accuracy curve

ACKNOWLEDGMENT

We would like to express our gratitude to BBC for sharing their data which gave us the golden opportunity to do this wonderful project on the topic "A comparative Approach for News categorization using Machine Learning".

REFERENCES

[1] Drury, B., Torgo, L., and Almeida, J.J. —Classifying News Stories to Estimate the Direction of a Stock Market Index. Paper presented at the Information Systems and Technologies (CISTI) 6th Iberian Conference, 2011

[2] M. W. Pope, "Automatic Classification of Online News Headlines," 2007. School of Information and Library Science of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Science in Information Science (November 2007).

[3] L Xin, R Gao, and L Song. Internet News Headlines Classification Method Based On The N-Gram Language Model. International Conference on Natural Language Processing and Knowledge Engineering, 2012.

[4] Dr. R. R. Deshmukh, Mr D. K. Kirange. Classifying News Headlines for Providing User Centered ENewspaper Using SVM. International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)

[5] D. K. Kirange, R. R. Deshmukh, "Emotion classification of news headlines using SVM," Asian Journal of Computer Science and Information Technology, pp. 104-106, 2012

[6] Dilrukshi, I., De Zoysa, K., Caldera, A.: Twitter news classi cation using svm. In: Computer Science & Education (ICCSE), 2013 8th International Conference on, IEEE (2013) 287-291

[7] Seyyed Mohammad Hossein Dadgar et al "A Novel Text Mining Approach Based on TF-IDF and Support Vector Machine for News Classification" 2nd IEEE International Conference on Engineering and Technology (ICETECH), 17th& 18thMarch 2016.

[8] Sandeep Kaur, Navdeep Kaur Khiva "Online news classification using Deep Learning Technique" International Research Journal of Engineering and Technology (IRJET) Volume: 03 Issue: 10 ,Oct -2016

[9] Vandana Korde,C Namrata Mahender "Text classification and classifier:A survey" International Journal of Artificial Intelligence & Applications. 2012 [10]H. Wu and R. Luk and K. Wong and K. Kwok. "Interpreting TF-IDF term weights as making relevance decisions". ACM Transactions on Information Systems, 26 (3). 2008