International Journal for Modern Trends in Science and Technology, 7(11): 40-45, 2021

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST0711008
Available online at: http://www.ijmtst.com/vol7issue11.html

Design and Implementation of NEWS Classification Predictor using Mechine Learning

S. Rahamat Basha | Dr. T. Bhaskara Reddy

Dept. of Computer Science and Technology, S. K. University, Anatapuramu, AP, India

To Cite this Article
S. Rahamat Basha and Dr. T. Bhaskara Reddy. Design and Implementation of NEWS Classification Predictor using Mechine Learning. *International Journal for Modern Trends in Science and Technology* 2021, *7*, pp. 40-45. https://doi.org/10.46501/IJMTST0711008.

Article Info
Received: 16 September 2021; Accepted: 02 November 2021; Published: 05 November 2021

ABSTRACT

This work deals with document classification. It is a supervised learning method (it needs a labeled document set for training and a test set of documents to be classified). The procedure of document categorization includes a sequence of steps consisting of text preprocessing, feature extraction, and classification. In this work, a self-made data set was used to train the classifiers in every experiment. This work compares the accuracy, average precision, precision, and recall with or without combinations of some feature selection techniques and two classifiers (KNN and Naive Bayes). The results concluded that the Naive Bayes classifier performed better in many situations. The documents of the self-made corpus were collected online articles from CNN, Washington Post, and New York Times.category predictors are developed by training Naïve Bayes, Random Forest, K-Nearest Neighbor (KNN), and Support Vector Machine (SVM) models on the same dataset. Then, each category predictor's performance is evaluated by analyzing the confusion matrix and quantifying the test dataset's precision, recall, and overall accuracy. In the end, the performance of all category predictors is studied and compared. The results show that all category predictors have achieved satisfactory accuracy grades.

Keywords–Text Classification, Feature selection, Machine learning, accuracy

INTRODUCTION

In all its forms, the technology has a significant impact on society and has significantly changed the way people access information. The News is a well-known and standard platform that serves as an information disseminator for local and global masses. The recent technological advancements have considerably changed; the way News is produced, consumed, and disseminated. It has enabled the more frequent and on-spot News reporting that smartphones can access anywhere and anytime. Therefore, people now expect to receive News of their interest in real-time, and they have an array of choices—business, sports, technology, politics, and entertainment. The news sources are already flooded with colossal information. Therefore, it is essential to automatically classify the News in specific categories based on the information content to allow timely and efficient information dissemination. Automatic document classification can be used to efficiently manage the text-based information (i.e., News) [1-3]; it allows timely and efficient information retrieval in the search phase. ATC can assign a relevant category to a news from a predefined set of reference categories based on the text feature extraction by correctly understanding the meaning and context of words. The time required to categorize the News

correctly is directly proportional to the quantity of text. In the newspaper's archive, the comprehensive range of articles starts from business to technology, so it is inconceivable that humans could manage this abundant content of information in a reasonable time frame. The manual document classification is cumbersome and resource-exhaustive.

The news category predictor aims to recognize and categorize different news articles based on content/information type. The automatic news classification plays a vital role in processing a massive amount of news content. It can classify and label the news articles by analyzing the content (i.e., extracting feature values) to quickly access what they are interested in, allowing efficient and speedy news dissemination.

Additionally, news websites can also increase their visibility by developing a recommendation system that suggests/recommends the relevant news to attract more attention. Several studies have been carried out to study modeling and performance evaluation of news category predictors using machine learning algorithms over different datasets (i.e., datasets differ in languages and range of categories) [4-10]. In these studies, well-known machine algorithms, i.e., Naïve Bayes, SVM, Random Forest, etc. are used to model news category predictors. The findings/results show that the category predictor's performance can vary with the machine algorithm deployed and dataset used to train the model. In contrast, machine learning is envisioned to solve problems in various related domains [11,12]. For a given machine learning algorithm, prediction performance can vary significantly depending upon the dataset. To quote a few, the Naïve Bayes algorithm's precision in categorizing the NEWS articles is reported to be 0.92 in Ref. [3] and 0.88 in Ref. [5]. In both cases, different datasets are used to train the same machine learning model, but the prediction performance is observed to be different. In the past few years, lots of research is carried out using different machine algorithms in natural language processing (i.e., text/news classification [13-17]. However, this review paper is more focused on evaluating and comparing category predictors' performance based on well-known machine learning algorithms. We choose dataset of self-made news having five categories —business, sports, technology, politics, and entertainment. This is a balanced dataset and quite different from traditional datasets that usually contain the biases. The main novel contribution of this research is as follows, and we are the first who developed multi-class NEWS category predictor by training four well-known machine learning algorithms (i.e., Naïve Bayes, Random Forest, K-Nearest Neighbor (KNN) and Support Vector Machine (SVM)) on the same dataset.Then, eachcategorypredictor's performance is evaluated by analyzing the confusion matrix and quantifying the test dataset's precision, recall, and overall accuracy. Finally, the performance of all category predictors is studied and compared. The rest of this paper is organized as follows. Section 2 discusses Methodology and Dataset, Feature Engineering, and Class Encoding are presented in section 3 and section 4, respectively. Section 5 details the machine learning algorithms, Performance Evaluation of Category predictor and Results and Discussions are given in section 6 and section 7, respectively. Finally, the conclusion of the whole analysis is presented in section 8.

METHODOLOGY AND DATASET

The ultimate aim of this study is to classify the news into specific categories and analyze the performance of the category predictor. Initially, datasets are collected and preprocessed, then the content of text document (Dj) is converted into useful features (w1j... wkj) by feature extraction algorithms such as unigrams. The extracted features are transformed into numeric data that act as inputs for machine learning algorithms or classifiers (Naïve Bayes and Random Forest). Finally, the ML models are trained on these transformed features, and the performance is evaluated on the test data set. The research methodology/ work flowchart is given in Fig.1 below.

Dataset

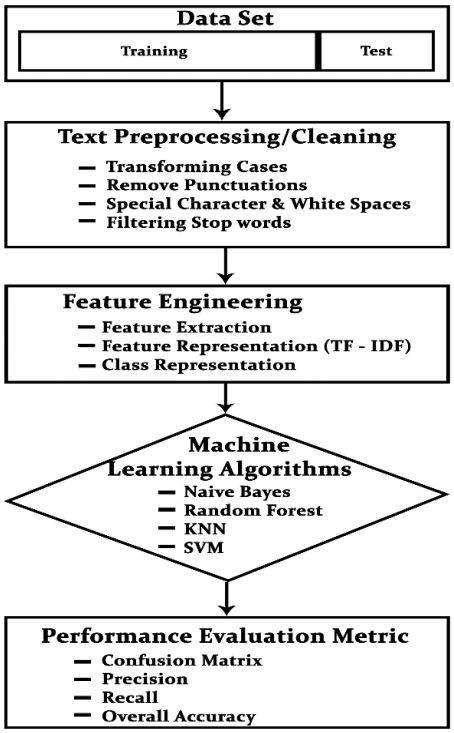

Fig 1. Methodology or Workflow

In this study, the Self-made news data set is used, which is obtained from Kaggle. It consists of 1490 documents (news) from the Self-made news website corresponding to stories in five typical areas —business, sports, technology, politics, and entertainment. However,maximum samples, such as 23.2%, belong to the sports category, and the tech category has a minimum share in the dataset. The distribution of classes plays an important in classification, and balanced datasets result in better learning models. In this study, the dataset is broken into 1,192 (80%) records for training and 298 (20%) testing.

Text Cleaning/ Preprocessing

Text preprocessing or cleaning is a preliminary and crucial step of news classification, which reduces space to make the classification more efficient [18-22]. Most of the time dataset is unstructured and combinations of useful and useless data. The unnecessary information such as stop words, punctuations, special characters, irrelevant sentences, quotations, and dates do not add any predictive power to the classifier/model. It only consumes space and can distort the ML model; therefore, before extracting any feature from the raw dataset, we should perform a cleaning process to minimize distortions introduced to the model. In this paper, we have followed these steps to preprocess the news text:

Transforming text

Transforming text in the same case (i.e., lower case) to eliminate homologous words that are different only in their case. For instance, the words “Fruit” and “fruit” are the same in a real sense and should not be considered as separately for prediction.

Removing punctuations and special characters

The characters such as "?", "!", ";" and "." are disposed of, this process simplifies computations in the next steps. Any special character and unnecessary whitespaces are also removed because they don't contribute to prediction power.

Filtering stop words

This technique is mainly used to remove unnecessary words or words with no specific meaning, such as "the, an, a, what, etc." so that classifier cannot co-relate stop words and important class features. Furthermore, the most frequent or rarely used words do not contribute to the predictive power model. Therefore, they must be removed from the training set. In this study, we have downloaded a list of English stop words from the nltk library and then removed them from the dataset.

CLASS REPRESENTATION/ENCODING

The news category prediction is a multi-class classification. For instance, the dataset used in this study corresponds to five classes — business, sports, technology, politics, and entertainment. Each class is labeled to make it more understandable and often labeled in words. For ML models, label encoding is used to transform labels into numeric values. It can be done by Label Encoder, which converts class labels into values between 0 and n-1, where n is the number of unique class labels. The actual and encoded labels of the dataset used in this study are given in Table I below.

TABLE I. CLASS REPRESENTATION ENCODING

S. No	*Labels*	*Encoded*
01	Business	0
02	Entertainment	1
03	Politics	2
04	Sport	3
05	Tech	4

MACHINE LEARNING ALGORITHMS

A classifier is a machine learning model that maps input data to a proper category. In this study, Naïve Bayes (NB), Random Forest (RF), K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) algorithms are used to train a model that can classify news articles into the right categories — business, sports, technology, politics, and entertainment.

Naïve Bayes

Naïve Bayes is a probabilistic (i.e., conditional probability) classification algorithm based on Bayes' Theorem. It is simple yet quite useful in a model, especially in text classification. In Bayes Theorem, the probability of any specific event is estimated by calculating its frequency in the past. The fundamental Naive Bayes assumption is that each feature is independent and unrelated to any other class feature. The Bayes theorem is

$$p\,(C \mid f\,) = \frac{p\,(f \mid C)*p\,(C\,)}{p\,(f\,)} \quad (1)$$

That is the probability of occurrence of C given that event f has already occurred. The event f is termed as evidence, p(C) is the prior probability of class, $p\,(f \mid C)$ is termed as likelihood and $p\,(C \mid f\,)$ is the posterior probability. In text classification features can be numerous such as $f\,(f1,\ f2\,,f3\,,f4\,\ldots\ldots fn\,)$ so by substituting f and expanding using the chain rule we get

$$p\,(C\mid f_1\,,f_2\,n,\ldots\ldots fn\,) = \frac{p\,(f_1\mid C\,)*p\,(f_2\mid C\,)*\ldots.p\,(f_n\mid C\,)*p\,(C\,)}{p\,(f_1\,)*p\,(f_2\,)*\ldots p\,(f_n\,)} \quad (2)$$

Thus, we can find out the category by finding the class with maximum probability.

Random Forest

Random Forest is a machine learning algorithm based on a set of trees classifiers. The RF is an ensemble method used for classification that constructs several decision trees at training time and makes a final decision on majority voting. It uses bootstrap sampling in which data samples are sampled independently and with the same distribution for all trees in the forest. The graphical depiction of the Random Forest Algorithm is given below

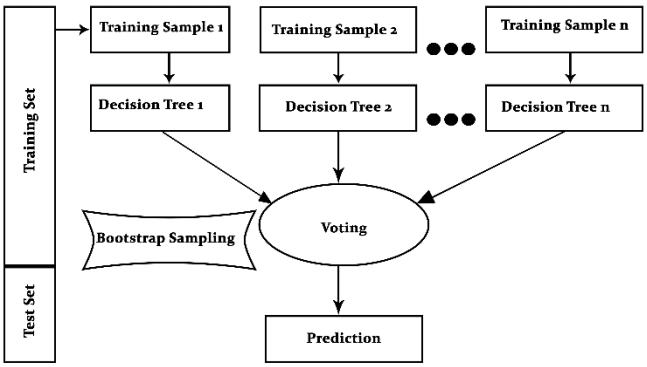

Fig. 1 Random Forest Classifier

K-Nearest Neighbours

KNN is an intuitive supervised learning algorithm and an easy method to implement. It is used to classify objects based on their nearest examples in training sets space. The procedure to identify an object is classified by a majority vote of its neighbors like an object is assigned to a common class among its closest neighbors. The new vector classification is found by classes of its k-nearest neighbors where k is a positive integer. This algorithm is implemented using Euclidean distance metrics to detect the nearest neighbor [23]. The main challenge in KNN is to determine the optimal value of k; higher the k value will increase the rise of over-learning so, it is necessary to take a valid value of k that reduces the over-learning. The Euclidean distance metrics d (x, y) between two points is computed as Eq. 1.

$$d\,(x,y) = \sum_{i=1}^{N}\sqrt{(x_i^2) - (y_i^2)} \quad (3)$$

where N is the number of features like $x = \{\,x_1\,,\ x_2\,,\ x_3\,,\ldots\ldots,\ x_N\,\}$ and $y = \{\,y_1\,,\ y_2\,,\ y_3\,,\ldots\ldots,\ y_N\,\}$. The number of k-neighbours used to test new vector varied from 1 to 10.

Support Vector Machines (SVM)

The SVM is the kernel-based machine learning algorithm that can categorize input data input into specific classes or categories. SVM constructs a classifier that makes the decision boundary for every class and defines the hyper-plane to linearly or non-linearly separate them. The accuracy of categorization can be increased by increasing the hyper-plane margin that also enlarges the distance among classes. Hence, the farthest hyper-plane provides more immunity against noise. SVM is a kernel-based classifier that defines the process of mapping the training data set to develop its

similarities to a linearly independent data set. The main reason to use mapping is to enhance the depth of the data set done by kernel function like some commonly used kernel are linear, RBF, and quadratic, etc.

PERFOMANCE EVALUATION METRICS

To precisely gauge the performance of the category predictor (i.e., categorizing the news articles), there are different performance evaluation techniques and metrics such as Confusion Matrix, Accuracy, Precision, Recall or sensitivity, and F1-Score. In this study, the confusion matrix is evaluated first, then accuracy, precision, and recall are analyzed to get a true insight into prediction performance.

RESULTS AND DISCUSSION

The performance results of multi-class category predictors (i.e., categorizing news articles into specific categories) based on different supervised learning models are evaluated and compared in this section. This study's learning models are Naïve Bayes, Random Forest, K-Nearest Neighbor (KNN), and Support Vector Machine (SVM). The evaluation was done by observing each category predictor's prediction results by analyzing the confusion matrix and quantifying the Precision, recall, and overall accuracy. This analysis was made on a test dataset containing 298 samples of news. In general, every category predictor has achieved good accuracy; however, SVM based category predictor achieved the highest accuracy for the data with five categories. In the NB model, the most misclassified category was Technology with four incorrect predictions, while the most accurately classified category was sport having no any wrong prediction. The detailed category/class wise analysis of Precision and recall for SVM and NB model is given in Table II and Table III, respectively. If we analyze the category predictor model based on random forest algorithm, the predictionperformance is satisfactory.

However, there more incorrect predictions as compared to SVM and NB models. The most misclassified category was business with seven inaccurate predictions, while the most accurately classified category was sport. RF model achieved an accuracy of 94.9\%, with 283 correct predictions out of 298 test samples. The detailed category/class wise analysis of Precision and recall for the NB model is given in Table IV. KNN based category predictor achieved the lowest accuracy. It achieved an accuracy of 94.2\% with 17 wrong predictions, as shown in Fig.4 (c). The detailed category/class wise analysis of Precision and recall for the KNN model is given in Table V. Although KNN has achieved the lowest accuracy grades as compared to SVM, NB, and RF, it is still satisfactory performance.

TABLE I. PERFORMANCE EVALUTION OF CATEGORY PREDICTOR USING SVM ALGORITHM

Category	***Precision***	***Recall***
Business	0.92	0.91
Entertainment	0.94	0.95
Politics	0.96	0.92
Sport	0.97	0.95
Tech	0.96	0.91

TABLE II. PERFORMANCE EVALUTION OF CATEGORY PREDICTOR USING NAÏVE BAYES ALGORITHM

Category	***Precision***	***Recall***
Business	0.92	0.92
Entertainment	0.93	0.96
Politics	0.94	0.92
Sport	0.99	0.97
Tech	0.93	0.93

TABLE I. PERFORMANCE EVALUTION OF CATEGORY PREDICTOR USING RANDOM FOREST ALGORITHM

Category	**Precision**	**Recall**
Business	0.95	0.90
Entertainment	0.95	0.95
Politics	0.96	0.95
Sport	0.94	0.98
Tech	0.92	0.91

TABLE II. PERFORMANCE EVALUTION OF CATEGORY PREDICTOR USING KNN ALGORITHM

Category	**Precision**	**Recall**
Business	0.91	0.89
Entertainment	0.92	0.92
Politics	0.92	0.93
Sport	0.97	0.92
Tech	0.92	0.95

CONCLUSION AND FUTURE WORK

This paper presents a comparative analysis of the multi-class category predictor's prediction performance. At first, the NEWS category predictor is developed by deploying/training well-known machine learning algorithms (i.e., Naïve Bayes, RandomForest, K-Nearest Neighbor (KNN), and Support Vector Machine (SVM)) on *Self-made* news dataset having five categories —business, sports, technology, politics, and entertainment. Later, using performance evaluation metrics, we analyzed the confusion matrix and quantified the test dataset's precision, recall, and overall accuracy. As a result, the SVM model proved the best among four supervised learning models in correctly categorizing the NEWS articles with 98.3% accuracy. In contrast, the lowest accuracy was obtained KNN model with (K=5). However, the KNN model's performance can be enhanced by investigating the optimal number of neighbors (K) value. Furthermore, deep learning schemes will be introduced to improve further said dataset's performance as to future work.

REFERENCES

[1] A. Hakim, et al., "Automated document classification for news article in bahasa indonesia based on term frequency inverse document frequency (tf-idf) approach," in 2014 6th International Conference on Information Technology and Electrical Engineering (ICITEE), Oct 2014, pp. 1–4.

[2] G. Mujtaba, et al., "Prediction of cause of death from forensic autopsy reports using text classification techniques: A comparative study," Journal of Forensic and Legal Medicine, vol. 57, pp. 41 – 50, 2018.

[3] V. S. Padala, et al., "Machine learning: The new language for applications," IAES International Journal of Artificial Intelligence, vol. 8, no. 4, p. 411, 2019.

[4] F. Miao, et al., "Chinese news text classification based on machine learning algorithm," in 2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), vol. 02, Aug 2018, pp. 48–51.

[5] S. M. H. Dadgar, et al., "A novel text mining approach based on tf-idf and support vector machine for news classification," in 2016 IEEE International Conference on Engineering and Technology (ICETECH), March 2016, pp. 112–116.

[6] G. L. Yovellia Londo, et al., "A study of text classification for indonesian news article," in 2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT), March 2019, pp. 205–208.

[7] R. Wongso, et al., "News article text classification in indonesian language," Procedia Computer Science, vol. 116, pp. 137 – 143, 2017, discovery and innovation of computer science technology in artificial intelligence era: The 2nd International Conference on Computer Science and Computational Intelligence (ICCSCI 2017). [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050917320872

[8] N. Chy, et al., "Bangla news classification using naive bayes classifier," in 16th Int'l Conf. Computer and Information Technology. IEEE, 2014, pp. 366–371.

[9] Dilrukshi, et al., "Twitter news classification using svm," in 2013 8th International Conference on Computer Science & Education. IEEE, 2013, pp. 287–291.

[10] H. Sawaf, et al., "Statistical classification methods for arabic news articles," Natural Language Processing in ACL2001, Toulouse, France, 2001.

[11] Mrema, et al., "A Survey of Road Accident Reporting and Driver's Behavior Awareness Systems: The Case of Tanzania," Engineering, Technology & Applied Science Research, vol. 10, no. 4,p. 6009, 6015.

[12] Kiruthika, et al., "A Survey of Road Accident Reporting and Driver's Behavior Awareness Systems: The Case of Tanzania," Engineering, Technology & Applied Science Research, vol. 10, no. 3,p. 5759, 5768.

[13] N. M. N. Mathivanan, et al., "Performance analysis of supervised learning models for product title classification," IAES International Journal of Artificial Intelligence, vol. 8, no. 3,p. 228, 2019.

[14] A. Khan, et al., "A review of machine learning algorithms for text-documents classification," Journal of advances in information technology, vol. 1, no. 1, pp. 4–20, 2010.

[15] V, "Twitter tweet classifier," IAES International Journal of Artificial Intelligence, vol. 5, no. 1, pp.41–44, 2016.

[16] M. A. Al-Hagery, "Extracting hidden pattern from dates' product data using a machine learning technique," IAES International Journal of Artificial Intelligence (IJ-AI), vol. 8, no. 3, pp. 9–18, 2019.

[17] Ferrario and M. Naegelin, "The art of natural language processing: Classical, modern and contemporary approaches to text document classification," Modern and Contemporary Approaches to Text Document Classification (March 1, 2020), 2020.

[18] E. Haddi, et al., "The role of text pre-processing in sentiment analysis," Procedia Computer Science, vol. 17, pp. 26 – 32, 2013, first International Conference on Information Technology and Quantitative Management. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050913001385

[19] Guyon and A. Elisseeff, "An introduction to variable and feature selection," J. Mach. Learn. Res., vol. 3,pp. 1157–1182, 2003.

[20] M. Farhoodi, A. Yari, and A. Sayah, "N-gram based text classification for persian newspaper corpus," in The 7th International Conference on Digital Content, Multimedia Technology and its Applications, Aug 2011.

[21] F.Sebastani, "Machine Learning in Automated Text Categorization", ACM Computing Surveys, Vol.: 34, Issue: 1, pp.1-47, 2002.

[22] S. R Basha, et al. "Impact of feature selection techniques in Text Classification: an experimental study", J. Mech. Cont.& Math. Sci., Special Issue, No. 3, pp. 39-51, 2019.

[23] J. Keziya Rani "A Comparative Approach of Dimensionality Reduction Techniques in Text Classification" Engineering, Technology & Applied Science Research, ISSN -1792-8036, Vol. 9, No. 6, Dec 2019, PP:4974-4979.