International Journal of Advanced Science and Technology
Vol. 29, No. 9s, (2020), pp. 620-630

Real Time News Classification Using Machine Learning

Akanksha Patro, Mahima Patel, Richa Shukla and Dr. Jagurti Save
Dept. of Information Technology, Fr.Conceicao Rodrigues college of Engineering, Mumbai, India

Abstract

With the existence of a number of sources on the internet generating immense amount of daily news, there is a necessity to classify the news articles to make the information available to users quickly and effectively. So the task of news classification starts by collecting real time news articles from news websites through web scraping and then automatically classifying it using various classification algorithms. Thus news classification is a way to identify topics of untracked news as well as make Individual suggestions based on the user's prior interest. This paper discusses various steps of news classification and implements a few algorithmic approaches including Naïve Bayes, Logistic Regression, SVM, Decision tree and Random forest for automatic classification of news articles into topics using the BBC News dataset that contains articles belonging to five different categories (Business, Entertainment, Politics, Sport, Technology). The paper reviews the results of different classification algorithms and compares them as per various performance measures. It also shows the performance of our news classifier on real time news articles crawled from news websites.

Keywords: News classification, Natural language processing, Feature selection, Classification model, Multinomial Logistic Regression

1. INTRODUCTION

With the rapid growth of online unstructured textual data, it has become a need to classify the text into categories so as to analyse and interpret relevant insights that contribute to decision making. Thus text categorization or classification is the process of assigning tags or classes to unstructured data according to its semantic content [1]. This not only eases the procedure of indexing the rapidly growing data but also helps in retrieval of desired content from a large information space.

A challenging problem in natural language processing, information retrieval and machine learning is the classification of the semantic content. News articles provide a particularly great example of such classification owing to the fact that the content of news articles is generally precise, incisive and consistent. Most news groups generate a large number of news stories on a daily basis which should be available to individual users in an organized or classified way. The task of manually labelling news articles is not only tedious but also time consuming. It makes it difficult for an application to manually label the latest news articles and feed then to the readers in real time. This demands the use of a tool that would automatically classify the news articles in real time so that the users can access the latest labelled news stories [2].

Classification of news can be automated with the help of machine learning. The process is related to natural language processing where in the classifier tries to obtain the relationship between the text features and the text categories as per the labelled training dataset and then uses this classifier to label the latest news articles. The classifier tries to discover the most probable words for each category and takes them into consideration while classifying new articles. The classifier discards all such words that appear frequently for all categories.

The main objective of news classification [3] is to analyse news data technically to uncover patterns in news production and content. Topic classification of news articles is found to be useful for specific applications and researches. Given the social and political impact of news and media, such textual interpretation and modelling not only helps to discover unidentified biases and rationale behind news stories but also enables automatic tagging of online news reports, aggregation of news sources by topic as well as provides the basis for news recommendation systems. [4]

2. Related Work

Zach Chase [5] proposed a multi-label topic classification model which includes labels such as business, arts, technology, style, books, home, sports, science and health. Multi-label topic classification made use of a binary classifier for getting optimized and efficient results. This binary classifier summarized the results into three feature sets: full article text, lead paragraph and article headline. The derivative of the term frequency-inverse document frequency model was implemented for information retrieval. The calculated final score per category needed to be checked whether it belongs to each class or not and for this two different machine learning techniques were used namely threshold with k-means and threshold with logistic regression.

Online news classification using deep learning techniques aimed to increase the accuracy in the prediction of news categories by implementing Neural Networks. The first step was the creation of a training database. The next step was the selection of training sides which was done using the Neural Network arc selection process. The weights were initialized, interpreted and computed. It was found that Random Forest provides highest discrimination power of 73% accuracy. The whole simulation was done in MATLAB 2010a that includes parameters like precision rate, recall rate and accuracy. The average values for all the three parameters as calculated were 0.76125, 0.05375 and 99.285 respectively [6].

Chen-Wei Tsai [7] proposed a real-time news classifier that collects news from U.S news websites and classifies them into seven categories namely US, world, opinion, business, tech, entertain and sport. Naïve Bayes Multinomial(NBM) model and linear kernel Support Vector Machine(SVM) models were implemented. A python library Scikit-learn was used where encoded training data was fed and the prediction model was saved. Top 30 Term frequency-inverse document frequency(TF-IDF) scores were calculated for each of the categories. The result was given in terms of testing error rates and implied that the classifier performs well for categories of tech and entertainment with the minimum error rates of 0.1176 and 0.1107 respectively.

Jake E. Sembodo [8] proposed an automatic tweet classification based on news category in Indonesian language. The classification process involved three parts namely data collection, data labelling, machine learning and experiment. The authors implemented algorithms such as ZeroR, Naïve Bayes Multinomial(NBM), Support Vector Machine(SVM), Random Forest and Sequential Minimal Optimization(SMO). WEKA was used as a classification tool for machine learning whereas Ms. Excel was used for preprocessing and labelling. Ten-fold cross validation technique was used for performance evaluation. Information was retrieved using term frequency-inverse document frequency(TF-IDF) which reflects how important a word is in the article. Multinomial Naïve Bayes gave the highest performance accuracy of 77,47% whereas ZeroR algorithm resulted in lowest performance accuracy i.e. 19,17%.

Olga Fuks [9] proposed a model that takes news headlines and a short description as input and then finally outputs the category. This model used traditional machine learning algorithms as well as deep learning algorithms. The various machine learning algorithms used were Naïve Bayes, Multinomial Logistic Regression, kernel Support Vector Machine and Random Forest. Deep learning involved Convolutional Neural Network(CNN) and Recurrent Neural Network(RNN). With observations the authors concluded that several models often get confused between the true label and the label of the most frequent class. This leads to the prediction of top three labels by each model. The dev set achieved 68.85% accuracy when top 1 label was taken into consideration whereas it achieved 88.72% accuracy when top 3 labels were predicted by the model.

3. Steps of News Classification

The process of news classification typically starts by gathering news articles from different sources. The unstructured data so collected needs to be preprocessed in order to convert it into a format suitable for applying machine learning algorithms. The performance of the news classifier can then be checked using various measures. Thus the steps involved are news collection, preprocessing, feature

extraction, application of classification algorithms and evaluation of performance measures [10] which have been described as follows:

3.1. News Collection

Collection of news articles can be done from many sources such as newspapers, magazines, press, radio and many more. But due to web expansion, the internet has become a major source for obtaining news. The data may be available in any format such as .csv, .json , .html , .doc etc.

3.2. News Preprocessing

The data that has been gathered from multiple sources needs to be cleaned so as to free it from inconsistent and futile data making it suitable for further process. Preprocessing [11] is thus considered necessary for news classification as it converts the unstructured data into structured data. Some preprocessing steps include:

3.2.1. News Tokenization

It is the process of splitting or fragmenting huge text into a list of segments or tokens. Each word in the news article can be thought of as a token. The output of this step is given as input for the further steps of preprocessing

3.2.2. Stop word Removal

Stop words are words such as conjunctions, prepositions, articles etc. present in the news which do not carry much information relevant for classification. They are treated as words of low worth and must be removed prior to applying any classification algorithm so as to improve the accuracy. Removal of stop words can be done in three different ways.

- On the basis of Concepts: The words which provide very less information helpful for classification are removed
- Using a predefined list: A list of approximately 545 English stop words has been provided by the Journal of Machine Learning Research. Stop words can be removed by removing the words present in the above list. It can be done using the NLTK library in python.
- Frequency Count: In this method word frequency is computed and the weights are assigned to these words. On the basis of these weights the stop words are removed.

3.2.3. Stemming and Lemmatization

Word stemming is a fundamental rule based process that strips common suffixes and prefixes from the word to normalize it to its root. Lemmatization on the other hand is a step by step process of converting a word to its root by considering the vocabulary and morphological analysis.

Lemmatization is found to be more powerful and organized as compared to stemming as it uses dictionaries which are formed using in-depth linguistic knowledge. It helps to obtain better features as compared to stemming.

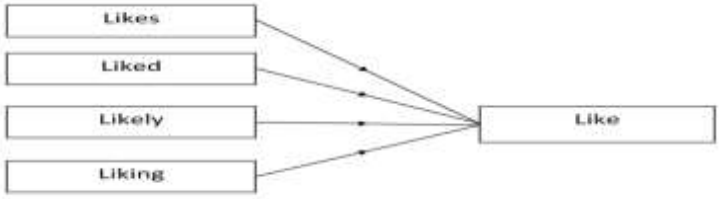

Figure 1. Example of Lemmatization

Figure 1 demonstrates an example of lemmatization where different words 'Likes', 'Liked ', 'Likely', 'Liking' are all converted into its base word 'Like' by slicing off the suffixes. Similarly, the words 'Playing', 'Played', 'Plays' get transformed into its root form 'Play'.

3.3. Feature Selection

It is the process of selecting relevant and highly effective features which contribute the most in classification so as to reduce training time, avoid overfitting and improve accuracy. Some of the feature selection methods are

3.3.1. Count Vectors as features

Count Vector is a matrix notation of the dataset in which every row represents a document from the corpus, every column represents a word from the corpus, and every cell represents the frequency count of a particular token in a particular document. For instance, consider a dataset consisting of a document with text "One Cent, Two Cents, Old Cent, New Cent: All About Money". This text can be converted into a matrix as shown in Table-1.

Table-1 Count Vector

	About	All	Cent	Cents	money	new	Old	one	two
doc	1	1	3	1	1	1	1	1	1

Table-1 represents the count vector for the text which consists of 9 columns each representing a unique word from the text and a single row representing the only document from the dataset. The values in each cell give the word count.

3.3.2. Term frequency-inverse document frequency (TF-IDF) as features

TF-IDF score represents the relative importance of a word in the document as well as the corpus. The score consists of two parts. The term frequency (TF) score can be calculated as the number of times that token 'w' occurs in an article 'a', summed across all the articles in a particular class.

$$\forall w, tf(w) = \sum_{articles\ a} \frac{\sum_{j=1}^{n} 1\left\{w = x_j^{(a)}\right\}}{\sum_{j=1}^{n} x_j^{(a)}} \tag{1}$$

The inverse document frequency (IDF) score can be calculated by dividing the total number of words in the corpus by the count of the instances of the particular word in the data and then taking the logarithm of that quotient.

$$\forall w, idf(w) = \log\left(\frac{\sum_{articles\ a,\ \ words\ j} x_j^{(a)}}{\sum_{articles\ a} x_w^{(a)}}\right) \tag{2}$$

The TF-IDF score can be obtained as:

$$tf\ idf(w) = tf(w) \times idf(w) \tag{3}$$

3.4. Classification models

The process of feature selection is succeeded by the classification of unseen news into their respective categories. Classification algorithms most commonly used for news classification are Naïve Bayes, Logistic regression, SVM, Decision tree and Random forest. [12]

3.4.1. Multinomial Naïve Bayes

Multinomial Naïve Bayes [13], [14] is a probabilistic algorithm used in natural language processing problems. In order to predict the category of a given sample, it applies Bayes theorem to calculate the probability of each category. It works on the strong assumption that each feature is independent of the other. News classification using Naïve Bayes algorithm is easy to implement and performs satisfactorily. But it shows poor performance when the assumption is violated i.e. when the features are highly correlated.

3.4.2. Multinomial Logistic regression

It is a supervised classification algorithm commonly used in machine learning when the target variable is discrete. It estimates probabilities by making use of the logistic/sigmoid function to compute the relationship between the categorical dependent variable and one or more independent variables. [15]

News classification using logistic regression is very efficient and outputs well calibrated predicted probabilities. Feature engineering plays a role in determining its performance as it works better when the model has little or no multicollinearity.

3.4.3. Support Vector Machine

SVM is a supervised algorithm used for both classification as well as regression [16]. The algorithm plots each data item as a point in n-dimensional space, where each data point represents the value of each feature as a co-ordinate. The model then extracts a best possible hyper-plane that segregates the classes.

SVM [17], [18] is a widely implemented algorithm for news classification. It supports high dimensional datasets and avoids overfitting of the data by minimizing not just the error but both error and complexity. Also SVM has the ability to find the global minimum.

3.4.4. Decision Tree

Decision tree is a tree based learning algorithm in which each internal node represents a test on the attribute, each branch determines a result of the test and each leaf node represents a class label. The root of the tree is the top-most node. The classification rules are given by the path from root to the leaf.

Decision tree for news classification[19] is easy to comprehend as the rules can be easily produced. They work the best to solve intricate problems. However, in classification problems with many classes and small number of training examples, decision tree is prone to produce errors. Training a decision tree is computationally expensive.

3.4.5 Random Forest

Random forest is an ensemble based learning method which can be used for both classification and regression problems. In this algorithm a number of decision trees work as an ensemble. Each individual decision tree in the random forest gives out a class prediction and the class with the maximum number of votes becomes our model's prediction. [20]

News classification using random forest[21] presents appreciable results as compared to any individual tree as it uses bagging and feature randomness when building each individual free in order to create forest of trees that are uncorrelated to each other. The performance of Random forest degrades when the trees of the forests and their prediction are correlated to each other.

3.5. Performance Measure

Evaluation is the last stage of news classification wherein different performance metrics are used to assess the overall performance of the classification models. Confusion matrix [22] is a matrix based performance measure used for machine learning classification problems with two or more classes. It summarizes the four elements: true positive(TP), false positive(FP), true negative(TN) and false negative(FN). The summation of TP, FP, TN and FN gives the total number of test cases used for classification. Researchers have made use of the confusion matrix in order to compute numerous performance measures to evaluate the classifiers such as Precision, Recall, Accuracy, F1 score and Error rate for evaluation.

Precision: It measures patterns that were correctly predicted from the total predicted patterns in the positive class.

$$Precision\ p = tp/(tp + fp) \quad (4)$$

Recall: It measures the portion of patterns belonging to positive class that were correctly classified.

$$Recall\ r = tp/(tp + fn) \quad (5)$$

Accuracy: It is calculated as the ratio of the total number of correct predictions to the total number of test instances.

$$Accuracy\ a = (tp + tn)/n \quad (6)$$

F1-Score: It is the harmonic mean of precision and recall.

$$F1 = 2 \times (p \times r)/(r + p) \quad (7)$$

Error rate: It measures patterns that were incorrectly predicted from the total number of test instances in the dataset.

$$Error\ rate = (fp + fn)/(tp + tn + fp + fn) \quad (8)$$

Some other metrics which can be used for evaluating classifiers are AUC-ROC Curve, Log loss, F-Beta score etc.

4. Implementation

4.1. Experimental Data

The dataset used for news classification is BBC News dataset consisting of news articles collected from its website. The dataset consists of 2225 documents corresponding to stories in five topical areas namely business, entertainment, politics, sport and tech from 2004 to 2005. The dataset is further divided into training and testing data. The number of articles in the training dataset is 1891 which constitutes 85% of the entire dataset. The remaining 15% of the dataset forms the testing dataset comprising 334 articles.

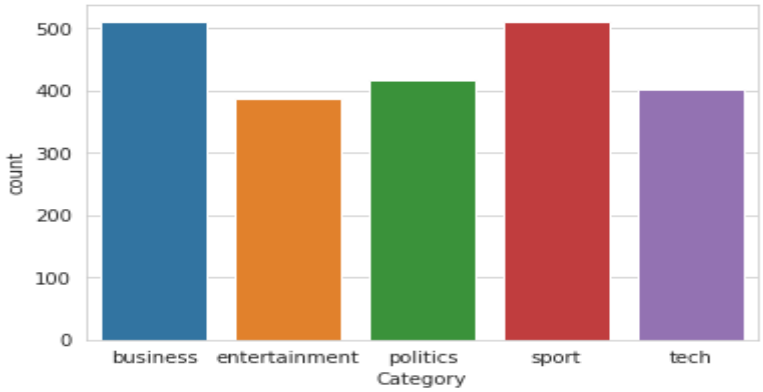

Figure 2. Graphical representation of categories

Figure 2 represents the proportion of news articles that belong to each of the five categories.The maximum proportion belongs to the sports category with 511 articles followed by business with 510 articles.On the other hand, the entertainment category has the minimum number of articles of about 386.

Figure 3. Data Instance

Figure 3 represents a data instance belonging to each of the five categories. It constitutes a news headline along with the article content.

4.2. Dataset Preprocessing and Feature Selection

Before the selection of features, the dataset is cleaned to ensure no distortions are introduced in the model later. Cleaning of the dataset involved certain steps like removal of special characters such as @,"/n"etc. ,removal of punctuation marks such as quotes, double quotes, full stop etc, downcasing the words in the articles. All these steps are necessary as special characters,punctuation marks and the case of the words won't contribute much to the prediction of the category. The dataset is then preprocessed through the process of news tokenization,stop word removal, stemming and lemmatization. Selection of features is done using TF-IDF vectorizer which combines count vectorizer and TF-IDF.

4.3. Algorithms Implemented

After performing the required preprocessing techniques on the dataset, classic machine learning algorithms such as Multinomial Naïve Bayes, Multinomial Logistic Regression, SVM, Decision tree and Random forest have been implemented to figure out which one better fits the data and correctly classifies the articles into their respective categories

5. Results and Discussion

The machine learning classification algorithms have been implemented in python using a laptop with 8gb RAM and i3 core processor.In order to measure the performance of all these algorithms, precision ,recall, accuracy rate and f1 score have been considered.

Table -2 Accuracy Rate of classifiers

Classification Models	Training Accuracy(%)	Testing Accuracy(%)
Multinomial Naïve Bayes	95.50	92.51
Multinomial Logistic Regression	**97.35**	**95.50**
SVM	96.40	93.41
Decision Forest	100	83.52

Random Forest	100	93.11

Note: Bold data shows best result based on accuracy of an algorithm

Table-2 is a summary of different models and their accuracy rates. Overall the accuracy comes out to be good for every model. However Multinomial Logistic regression provides the best accuracy among all.

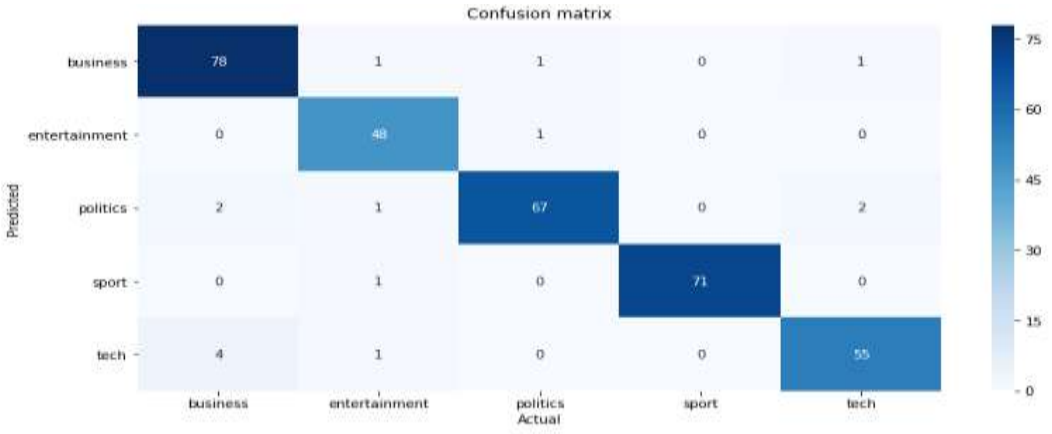

Figure 4. Confusion Matrix for BBC dataset

Figure 4 shows the confusion matrix produced by the logistic regression classifier on the given data. It highlights the fact that the Business category accounts for maximum correct classification followed by the Sports Category. However, entertainment category comes out to be the last owing to the fact that the dataset had minimum number of articles belonging to this category

Table-3 Performance metrics for different categories

Categories	Precision	Recall	F1-Score
Business	0.93	0.96	0.95
Entertainment	0.92	0.98	0.95
Politics	0.97	0.93	0.95
Sport	**1.00**	**0.99**	**0.99**
Tech	0.95	0.92	0.93

Note: Bold data shows best result based on accuracy of an algorithm

Table 3 shows the precision, recall and F1 score for each of the five categories. It can be observed that the precision, recall and F1 score for Sports category is higher as compared to the other categories.

In order to understand how the model interprets the category, some of the misclassified articles have been analyzed to understand that the model fails to identify the category of only those articles which clearly do not belong to any unique class. However,since Multinomial Logistic Regression performs better as compared to other algorithms it is further used for real time news classification.

6. Real Time News Classification

Web scraping or Web harvesting is a method of extracting unstructured data from online sources and transforming it into comprehensible format.When web scraping code is run, a request is sent to the URL mentioned in the code. The server then allows the HTML or XML pages to be read and responds to this request by sending the required data. The code then parses the HTML or XML page to get the

targeted data.Web Scraping in python can be implemented using two libraries: Requests.and BeautifulSoup. Requests in python is a simple HTTP library that allows to send HTTP/1.1 requests and BeautifulSoup works well with a parser in order to get data from HTML or XML files.

In order to fetch the latest news articles from news websites in real time, web scraping is used.The code for web scraping gathers news articles from the web page of a newspaper by sending a request to the targeted URL using the Requests library and then goes to the particular link to scrape the required data using the BeautifulSoup library.

For real time news classification ,a total of 100 news articles are scraped from the news website of India today with around 20 articles belonging to each of the five categories: business, entertainment, politics, sport and tech.This data is then subjected to methods of preprocessing and feature selection so that it is converted into a structured format that can be given as an input to the classifier.Multinomial logistic regression is then applied to the data as it showed the best results on the BBC dataset. The accuracy rate obtained by this algorithm is around 87%.

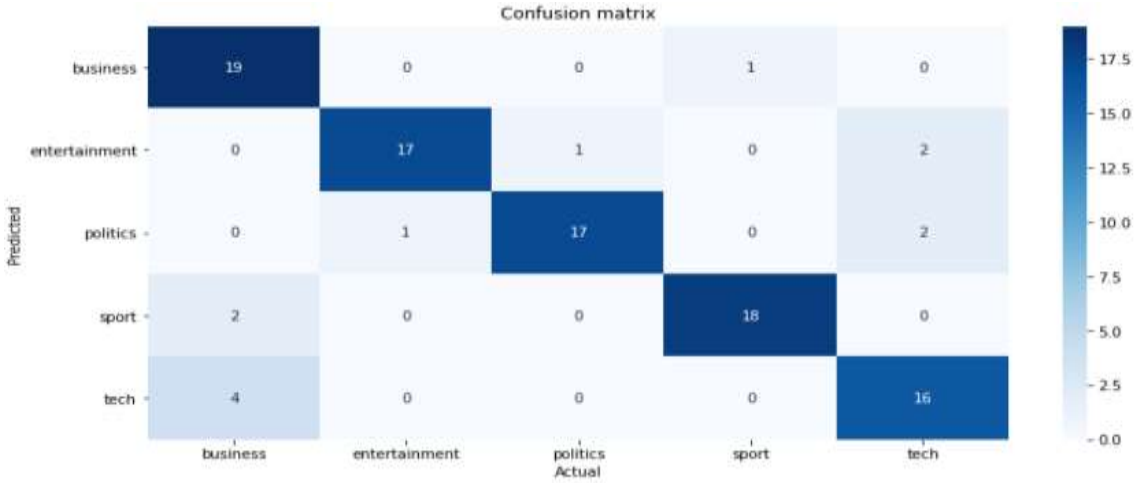

Figure 5. Confusion matrix for real time data

Figure 5 shows the confusion matrix produced by the logistic regression classifier on the real time data.Similar to the results obtained after classification of BBC testing dataset, Business category accounts for the maximum correct classification followed by the Sports category.

Table-4 Results for real-time data

Categories	Precision	Recall	F1-Score
Business	0.76	0.95	0.84
Entertainment	0.94	0.85	0.89
Politics	0.94	0.85	0.89
Sport	**0.95**	0.90	**0.92**
Tech	0.80	0.80	0.80

Note: Bold data shows best result based on accuracy of an algorithm

Table-4 shows the precision,recall and the F1 score for each of the five categories. It can be observed the precision and f1 score comes out to be the highest for the Sports Category similar to the results obtained using the BBC testing dataset.

Conclusions

All the steps for news classification have been examined in this paper. After studying various techniques about data mining concepts, the methods such as text tokenization, stop words removal, word stemming and word lemmatization have been used for preprocessing of the collected news. For

selection of the features, TF-IDF vectorizer has been used. Various algorithmic approaches such as Multinomial Naïve Bayes, Multinomial Logistic Regression, SVM, Decision tree and Random forest have been implemented on the BBC dataset in this paper. We have considered accuracy rate, precision, recall and F1 score as our performance measures. In comparison, Multinomial Logistic Regression is found to provide the best testing accuracy of 95.5% on the BBC dataset. Among the five categories, Business category has the most correct classification followed by Sports. The precision, recall and F1 Score values for Sports category are near to perfect. Multinomial logistic regression when applied on real time data scraped from the news website of India Today produces an accuracy rate of 87%. Similar to the results obtained in case of BBC dataset, the maximum correct classification goes to business category. When misclassified articles were analyzed, we came to the conclusion that the model fails only for those articles that clearly do not belong to any particular class.

Certain improvements can be made in the classification process so as to avoid the misclassification of the news articles. One such improvement could be the use of word synonyms in the classification phase. Also to deal with articles not belonging to a single unique class, multiple relevant labels for the news articles can be predicted. Moreover, Improvement can be made in terms of the time required to predict the category of an article by considering only the news headline. Consideration of only the news headline would eliminate tough statistical calculations and thus would decrease the overall time.

For future work, more categories of news can be added. The task of news classification can be extended for languages other than English. Also some other term weighting schemes can be used to check their performance against classifiers. A custom made stemmer can be built to amplify the classifiers performance. The algorithms can be further tested in combination on a large corpus.

References

[1] Kowsari, Kamran, et al. "Text classification algorithms: A survey." Information 10.4 (2019): 150

[2] Chan, Chee-Hong, Aixin Sun, and Ee Peng LIM. "Automated online news classification with personalization." (2001).

[3] Kaur, Gurmeet, and Karan Bajaj. "News Classification and Its Techniques: A Review." IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN (2016): 2278-0661.

[4] Carreira, Ricardo, et al. "Evaluating adaptive user profiles for news classification." Proceedings of the 9th international conference on Intelligent user interfaces. ACM, 2004.

[5] CHASE, Zach, Nicolas Genain, and Orren Karniol-Tambour. "Learning Multi-Label Topic Classification of News Articles." (2014).

[6] Kaur, Sandeep, and Navdeep Kaur Khiva. "Online news classification using Deep Learning Technique." International Research Journal of Engineering and Technology (IRJET) 3.10 (2016).

[7] Tsai, Chen-Wei, and Xuanang Chen. "Real-time News Classifier.",2017

[8] Sembodo, Jaka E., Erwin B. Setiawan, and Moch Arif Bijaksana. "Automatic Tweet Classification Based on News Category in Indonesian Language." 2018 6th International Conference on Information and Communication Technology (ICoICT). IEEE, 2018.

[9] Fuks, Olga. "Classification of News Dataset." Standford University (2018).

[10] Rana, Mazhar Iqbal, Shehzad Khalid, and Muhammad Usman Akbar. "News classification based on their headlines: A review." 17th IEEE International Multi Topic Conference 2014. IEEE, 2014.

[11] Srividhya, V., and R. Anitha. "Evaluating preprocessing techniques in text categorization." International journal of computer science and application 47.11 (2010): 49-51.

[12] Pradhan, Lilima, et al. "Comparison of text classifiers on news articles." Int. Res. J. Eng. Technol 4.3 (2017): 2513-2517.

[13] Kumar, R. R., Reddy, M. B., & Praveen, P. (2019). Text classification performance analysis on machine learning. International Journal of Advanced Science and Technology, 28(20), 691-697.
[14] Frank, Eibe, and Remco R. Bouckaert. "Naive bayes for text classification with unbalanced classes." European Conference on Principles of Data Mining and Knowledge Discovery. Springer, Berlin, Heidelberg, 2006.
[15] Tang, Bo, et al. "A Bayesian classification approach using class-specific features for text categorization." IEEE Transactions on Knowledge and Data Engineering 28.6 (2016): 1602-1606.
[16] Indra, S. T., Liza Wikarsa, and Rinaldo Turang. "Using logistic regression method to classify tweets into the selected topics." 2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS). IEEE, 2016.
[17] Yang, Yujun, Jianping Li, and Yimei Yang. "The research of the fast SVM classifier method." 2015 12th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP). IEEE, 2015
[18] Dilrukshi, Inoshika, Kasun De Zoysa, and Amitha Caldera. "Twitter news classification using SVM." 2013 8th International Conference on Computer Science & Education. IEEE, 2013
[19] Sawarkar, C. H., & Mulkalwar, P. N. (2019). Exploring the use of machine learning for highly accurate text-based information retrieval system. Test Engineering and Management, 81(11-12), 6592-6599.
[20] Shahi, Tej Bahadur, and Ashok Kumar Pant. "Nepali news classification using Naïve Bayes, Support Vector Machines and Neural Networks." 2018 International Conference on Communication information and Computing Technology (ICCICT). IEEE, 2018.
[21] Kaur, Komal Arunjeet, and D. Bhutani. "A review on classification using decision tree." IJCAT-International Journal of Computing and Technology (2015).
[22] Aljuaid, L., Wei, K. T., & Sharif, K. (2019). Machine learning: Tasks, modern day applications and challenges. International Journal of Advanced Science and Technology, 28(2), 329-340.
[23] Biau, GÃŠrard. "Analysis of a random forests model." Journal of Machine Learning Research 13. Apr (2012): 1063-1095.
[24] Liparas, Dimitris, et al. "News articles classification using Random Forests and weighted multimodal features." Information Retrieval Facility Conference. Springer, Cham, 2014.
[25] Patil, Tina R., and S. S. Sherekar. "Performance analysis of Naive Bayes and J48 classification algorithm for data classification." International journal of computer science and applications 6.2 (2013): 256-261.