Text Classification for English News Articles

Jobeda Khanam Ria, MD. Reaz Uddin and Sadman Majumder
Department of Computer Science and Engineering (CSE)
School of Data and Sciences (SDS)
Brac University
66 Mohakhali, Dhaka - 1212, Bangladesh
{jobeda.khanam.ria, md.reaz.uddin, sadman.majumder, }@g.bracu.ac.bd

***Abstract*—In today's world Natural Language Processing (NLP) has become a productive method which is highly used in the artificial intelligence and machine learning sector. A chatbot like chatGPT to Blockchain, every method of taking the advantages of NLP. Text classification is an important process of Natural Language processing (NLP) which includes categorizing or labeling text data according to predefined categories. However, there is a lot of text information available about text classification that is becoming a tool for a lot of applications, including sentiment analysis, recommendation systems, and information retrieval. In our research, we aim on text classification for English news articles using NLP. To reach the objective of our research which is to use variations of feature extraction and machine learning (ML) algorithms to enhance the correction rate as well as the effectiveness of text classification for English news articles. We have analyzed the results we get from the algorithm and tried to find the best performance as we compared the results we get from ML such as the Term Frequency-Inverse Document Frequency (TF-IDF) and the Vectorize method. We used different algorithms such as Random Forest (RF), Logistic Regression (LE), and Naive Bayes (NB) algorithms. For the research, we used the dataset from BBC News containing different data and articles. We worked on that dataset which contains news of various genres and as a result, we could judge the efficiency. Text data are pre-processed, features are extracted using various methods and classification models are trained using different ML algorithms. After attaining the result and accuracy, we have analyzed the results of the models. The results of this study will be applied to enhance the accuracy of the word categorization for news articles in other text-based applications. The results can be applied to develop reliable text classification algorithms that will improve data efficiency and accuracy.**

***Index Terms*—NLP, TF-IDF, Naive Bayes, Text Classification, Random Forest.**

I. Introduction

Text classification is an indistinguishable part of natural language processing (NLP) which includes categorizing or labeling text inputs according to predetermined categories. Text classification is becoming an important and significant component for many applications including sentiment analysis, recommendation systems and information retrieval, because of the huge quantity of textual data that is present on the internet. Our concern here is using text classification for news items using NLP [1] techniques in this research. The actual purpose of this research work is to use different feature extraction [2] and ML algorithms to increase the precision and effectiveness of text classification for news articles. We want to calculate the performances of different ML [3] such as Random Forest (RE) [4], Naive Bayes (NB), and Logistic Regression (LE) along with the numerical statistics TF-IDF [5] to investigate the efficiency of various feature extraction methods. A dataset of news items labeled with predetermined categories is gathered for research. Text data is pre-processed, features are extracted using a variety of methods and classification models are trained using a variety of machine learning algorithms [6]. We have analyzed the performances of the models using different results for Recall, accuracy and F1-score. The findings from this research can be applied to enhance the accuracy and effectiveness of text classification for English news articles and other text-based applications. The results can be used to form trustworthy text categorization algorithms that will improve the efficiency and standard of information retrieval in the digital era.

II. Literature Review

We reviewed some paper related to our topic ML and NLP. One paper shows us the pre-processing, feature selection, and classification which were used to classify the words in Indonesian language [7]. The Pre-processing cleaned text data as it removed stop words, and tokens. For feature selection they used TF-IDF and Support Vector Machine (SVM) [8] and for classification they used naive bias. SVM with TF-IDF fared best in the experiment with 92.63% accuracy. TF-IDF was also shown to capture Indonesian language's distinctive traits. This research demonstrates the potential of the machine learning for Indonesian text classification. It shows how pre-processing and feature selection can improve text classification accuracy. It is useful for natural language processing researchers and practitioners who classify Indonesian texts.

We can see the use of the machine learning and NLP methods here. Using a set of pre-prosessing techniques, various extraction methods and ML algorithms, the authors provided a methodology for text classification. Their work contributed huge knowledge in the sector of text classification according to the performances of a different machine learning algorithms and feature extraction techniques [9] In comparison, the Naive Bayes method's 90% accuracy, the SVM method achieved a 92.5% success rate. With an accuracy of 92.5%, the TF-IDF method outperformed BoW (bag-of-words) [10]. Their investigation into and application of the numerous machine learning algorithms and word extraction strategies have some very useful insights. There are many possible applications for

the proposed method beyond sentiment analysis and spam detection. These findings may be helpful for researchers and practitioners developing text categorization [11] methods. Further research is needed to determine the efficacy of deep learning algorithms for text classification.

The probability to use KNN (k-Nearest Neighbors) [12] problems along with processes for text classifications [13]. The framework can check the similarity of different words. Finally, the use of both KNN [14] algorithm and TF-IDF method has been discussed as a good choice with minor modifications in their implementation.

Naive Bayes method used for text classification using conditions to a class in this [15]. After feature selection in text classification, Naive Bayes Classifier divides the text subspace composed of all documents and then again the auxiliary feature method proposed here partition the text subspace again so that it can show better results than the normal or traditional way which indicates that the proposed method indeed improves the performance of Naive Bayes Classifier [15].

The Naive Bayes classifier is a method of machine learning that is used for text categorization. This work suggests an auxiliary feature strategy that may be utilised in conjunction with the Naive Bayes (NB) classifier. It indicates that using this method will help us get a more accurate result.

In another paper, the variety of Machine Learning and Deep Learning algorithms used in text classification is discussed with their advantages and shortcomings [16]. Moreover, it includes the benefits and limitations of feature extraction, feature selection method and supervised and unsupervised machine and deep learning models used for a text classification task. Despite being expensive, this paper shows the hope to find the future that these deep neural networks will be applied efficiently in the automatic monitoring of web based text data and classifying unseen data into automated labels with the advancements of deep neural networks. This paper gives an overview of the varieties of text features in different sectors using different methods and algorithms accurately in the real world. The text classification algorithms use some manners and these metrics help to evaluate the algorithm. Finally, with these techniques, various algorithm for text classification is discussed here.

III. Dataset

Since our project focuses on the newspaper article, hence, we need a dataset that contains news of various topic. One of the dataset we are using is The BBC News Archive dataset [17], which is comprising a total of 2,225 news pieces, spanning a period from 2004 to 2005. Each article in the collection includes crucial information such as the headline, category, date and the full text of the news piece. The dataset comprises many different sorts of news stories, making it useful for testing and researching various text categorization systems.

Business, entertainment, politics, sports and technology are just a few of the news genres represented in the dataset. The names of the source files that included the news pieces that are relevant to this topic are provided in the filename. In the title section, there is a title of the collected contents and the news article is present under the content section.

TABLE I
DATASET OF BBC NEWS ARCHIVE

Category	1623
File	1535
Title	1517
content	1698

Stop word indicates to the commonly used words which occurs several times or frequently. Therefore, these words have a very little contribution in terms of understanding the content. As a pre-processing step ”stop word removal” is important because these words don’t carry any meaning that’s why the meanings are more relevant and concise without these. Moreover, the removal of stop words helps to reduce the memory requirements as well as improves the model performance as the model concentrates on the main content words.

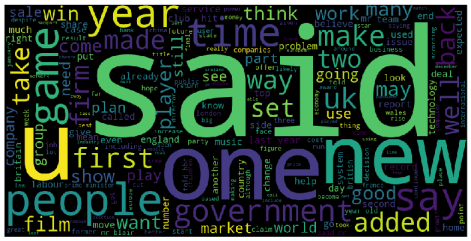

Fig. 1. Stopwords

IV. Methodology

In this research, we collected some dataset in a CSV format of news articles from different online sources. We pre-processed the data to get accurate results using different methods of NLP. The tokenization was done properly and removed the corresponding similar words and stop words. To evaluate the text classification, Confusion Matrix is a fundamental tool. Firstly, we convert the Confusion Matrix into Normalized Confusion Matrix. In a Confusion Matrix each class has the raw count of prediction whereas in a Normalized Confusion Matrix raw counts are transformed into percentages or proportions.

A. TF-IDF

Term frequency or TF calculate [18] the occurrence of the texts. It will be determined by the occurrences of the word in the documents. By doing so, we can find the word which is the important word and that will be used to fetch the article. The inverse document frequency (IDF) evaluate how much unique the term or the word is in the document collection. Finally, TF-IDF shows a simple and efficient algorithm to match words in a query to a document which returns documents that are precisely relevant to a particular query [19].

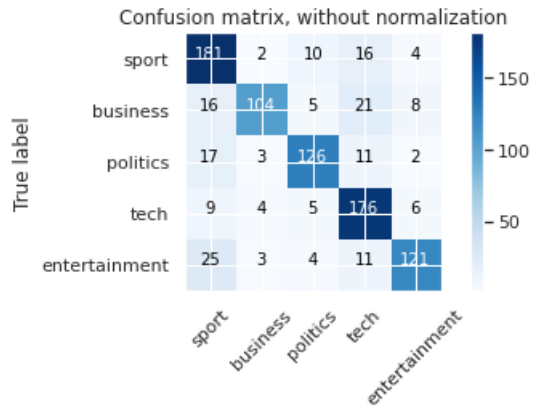

Fig. 2. Confusion Matrix

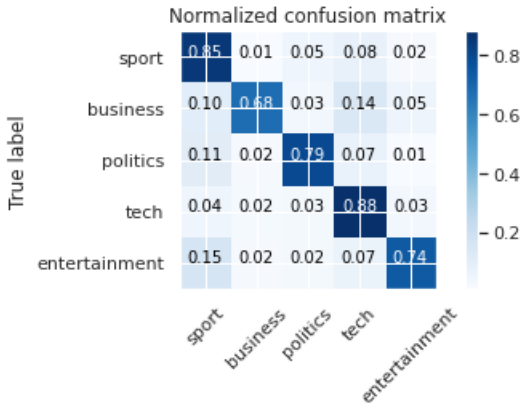

Fig. 3. Normalized Confusion Matrix

B. Multinomial Naive Bayes (MNB)

This algorithm [20] learns from a labeled training dataset, we will try to apply this to automatically categorize articles into different topics or classes. The topic or keyword will be from different genres. In contrast to our research, Naive Bayes have the capacity to deal with extensive vocabularies and high-dimensional feature spaces when working with a large number of words taken from a range of publications.

C. Logistic Regression (LR)

Logistic Regression [21] is efficient for computing purposes and can also handle a large dataset, as we are also working with the large dataset. It is also efficient in imbalance datasets.It does this by modifying the decision threshold or by employing class weights to compensate for the imbalance in class representation that exists. Therefore, within some extent Logistic Regression classification method with the given dataset has got the best classification accuracy in comparison to the analyzed classifiers [22].

D. Random Forest (RF)

Random Forest's randomization helps prevent overfitting and also makes the model more tolerant of erroneous information in text input. Together, these two features make the model more robust. It enhances the algorithm's capacity to generalise to new types of content that it has not previously encountered.

The above classifiers will be tested in our dataset to determine different rates such as F1 score.

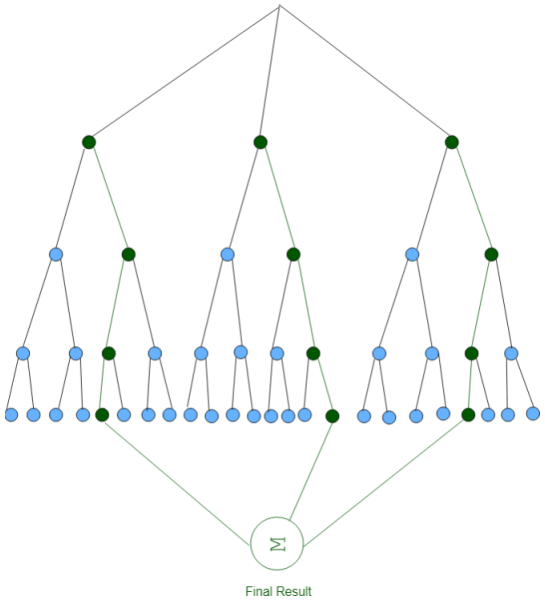

Fig. 4. Random Forest

V. Result and Discussion

Precision means predicted positive cases. Basically precision compares the number of positive cases which are compared with other cases if they are correctly predicted or not. Recall means how many positive predictions are correct. It calculates the ratio of accuracy. The F1 score is a numerical average which calculates the precision and recall and the higher value will show better performance.

TABLE II
Comparison of the Results

Model	Accuracy	Precision	Recall	F1	Process
RF	25.48%	25%	25%	25%	tf-idf
RF	97.86%	98%	98%	98%	vectorize
LR	20.95%	21%	21%	21%	tf-idf
LR	97.86%	98%	98%	98%	vectorize
MNB	19.76%	20%	20%	20%	tf-idf
MNB	96.43%	96%	96%	96%	vectorize

We look at the table of our results, where the applications of both TF-IDF and the vectorization process in both methods are verified. Among these, the Random Forest algorithm emerges as the pioneer, providing an impressive maximum value. 97.86% accuracy. The F1, recall, and precision scores

for both methods are at a respectable 98%. Another notable achievement is characterized by a remarkable 96.43% result achieved by applying the Naive Bayes technique. Here, the F1, recall, and precision scores are set at a solid 0.96 for this particular method.

Here the TF-IDF process registers relatively poor results. Apparently, the test accuracy of the random forest model is only 25.48%. The LR model and the MNB model have a test accuracy of 20.95% and 19.76% respectively.

After evaluating all of these different methods, it became clear that both the LR and RF methods provide the strongest accuracy in the vectorization process. Consequently, it is prudent to advocate the application of random forest and LR for future research efforts using vectorization methods.

In ML which is used in NLP, uses a lot of numerical data. The Vectorization methods convert the data, text and phrase into numerical data which helps the algorithm to access the data easily.

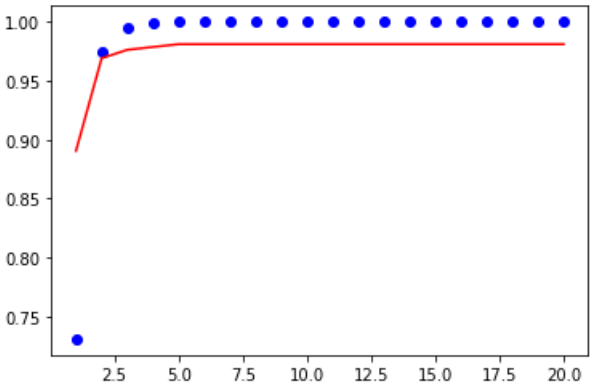

Fig. 5. Test and Training Accuracy

Here, from the Fig. 5 we can see after training 20 epochs and we check the validation after each epoch we got the training accuracy of 98% and the testing accuracy of 96% from the vectorized models at most from the dataset.

VI. Conclusion

The aim of this study is to explore the application of NLP techniques to text classification for news articles. We obtained a range of results by utilizing multiple datasets and different techniques. We reviewed previous literature on this topic and compared methods like Multinomial Naive Bayes, Logistic Regression and Random forest with two processes; Vectorize and the TF-IDF. The results were not helpful from the TF-IDF process as they gave very low results for all the methods. The information can be used for text searches in news articles. The text of news articles should be classified so that readers can find the relevant information they are looking for. We have researched on subject text classification but further research work will elaborate this sector more.

References

[1] E. R. Rhythm, R. A. Shuvo, M. S. Hossain, M. F. Islam, and A. A. Rasel, "Sentiment analysis of restaurant reviews from bangladeshi food delivery apps," in *2023 International Conference on Emerging Smart Computing and Informatics (ESCI)*, 2023, pp. 1–5.

[2] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," *Information*, vol. 10, no. 4, p. 150, 2019.

[3] M. Mhapsekar, P. Mhapsekar, A. Mhatre, and V. Sawant, "Advanced computing technologies and applications," 2020.

[4] S. Ahmed, M. H. K. Mehedi, M. A. Rahman, and J. B. Sayed, "Bangla music lyrics classification," in *Proceedings of the 2022 8th International Conference on Computer Technology Applications*, ser. ICCTA '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 142–147. [Online]. Available: https://doi.org/10.1145/3543712.3543752

[5] J. Chen and X. Tang, "Exploring societal risk classification of the posts of tianya club," *International Journal of Knowledge and Systems Science (IJKSS)*, vol. 5, no. 1, pp. 36–48, 2014.

[6] K. M. Hasib, M. R. Islam, S. Sakib, M. A. Akbar, I. Razzak, and M. S. Alam, "Depression detection from social networks data based on machine learning and deep learning techniques: An interrogative survey," *IEEE Transactions on Computational Social Systems*, pp. 1–19, 2023.

[7] R. Wongso, F. A. Luwinda, B. C. Trisnajaya, O. Rusli *et al.*, "News article text classification in indonesian language," *Procedia Computer Science*, vol. 116, pp. 137–143, 2017.

[8] M. Goudjil, M. Koudil, M. Bedda, and N. Ghoggali, "A novel active learning method using svm for text classification," *International Journal of Automation and Computing*, vol. 15, pp. 290–298, 2018.

[9] H. Li and Z. Li, "Text classification based on machine learning and natural language processing algorithms," *Wireless Communications and Mobile Computing*, vol. 2022, 2022.

[10] D. Guru, K. Swarnalatha, N. V. Kumar, and B. S. Anami, "Effective technique to reduce the dimension of text data," *International Journal of Computer Vision and Image Processing (IJCVIP)*, vol. 10, no. 1, pp. 67–85, 2020.

[11] F. Sebastiani, "Machine learning in automated text categorization," *ACM computing surveys (CSUR)*, vol. 34, no. 1, pp. 1–47, 2002.

[12] S. Rohwinasakti, B. Irawan, and C. Setianingsih, "Sentiment analysis on online transportation service products using k-nearest neighbor method," in *2021 International Conference on Computer, Information and Telecommunication Systems (CITS)*, 2021, pp. 1–6.

[13] S. E. Rad and A. R. Behjat, "Document classification base on ensemble classifiers support vector machine, multi-layer perceptron and k-nearest neighbors," *J. Biochem. Tech*, vol. 2, pp. 174–182, 2019.

[14] S. Bahassine, A. Madani, and M. Kissi, "Arabic text classification using new stemmer for feature selection and decision trees," *Journal of Engineering Science and Technology*, vol. 12, no. 6, pp. 1475–1487, 2017.

[15] W. Zhang and F. Gao, "An improvement to naive bayes for text classification," *Procedia Engineering*, vol. 15, pp. 2160–2164, 2011.

[16] V. Dogra, S. Verma, P. Chatterjee, J. Shafi, J. Choi, M. F. Ijaz *et al.*, "A complete process of text classification system using state-of-the-art nlp models," *Computational Intelligence and Neuroscience*, vol. 2022, 2022.

[17] "Bbc news archive. (n.d.). kaggle: Your machine learning and data science community. https://www.kaggle.com/datasets/hgultekin/bbcnewsarchive."

[18] N. Farhan, I. T. Awishi, M. H. K. Mehedi, M. M. Alam, and A. A. Rasel, "Ensemble of gated recurrent unit and convolutional neural network for sarcasm detection in bangla," in *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*, 2023, pp. 0624–0629.

[19] J. Ramos *et al.*, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, vol. 242, no. 1. Citeseer, 2003, pp. 29–48.

[20] A. Al Taawab, L. Tasnia, M. Dhar, and M. H. K. Mehedi, "Transliterated bengali comment classification from social media," in *2022 IEEE 10th Region 10 Humanitarian Technology Conference (R10-HTC)*, 2022, pp. 365–371.

[21] K. M. Hasib, N. A. Towhid, K. O. Faruk, J. Al Mahmud, and M. Mridha, "Strategies for enhancing the performance of news article classification in bangla: Handling imbalance and interpretation," *Engineering Applications of Artificial Intelligence*, vol. 125, p. 106688, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197623008722

[22] T. Pranckevičius and V. Marcinkevičius, "Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification," *Baltic Journal of Modern Computing*, vol. 5, no. 2, p. 221, 2017.