Text Classification for News Articles

Jobeda Khanam Ria, Sadman Majumder, MD. Reaz Uddin,
MD. Mustakin Alam, Md Sabbir Hossain, and Annajiat Alim Rasel
Department of Computer Science and Engineering (CSE)
Brac University
{jobeda.khanam.ria, sadman.majumder, md.reaz.uddin,
md.mustakin.alam, md.sabbir.hossain1}@g.bracu.ac.bd, annajiat@gmail.com

***Abstract*—Text classification is an essential process in NLP(natural language processing) that include classifying or labeling text data according to predefined categories. However, there is a large amount of text information available on the internet text classification is becoming an important tool for many applications, including sentiment analysis, recommendation systems, and information retrieval. In our research we are focusing on text classification for news articles using NLP(natural language processing) methods. The objective of our research is to use different feature extraction and machine learning techniques to increase the accuracy and effectiveness of text classification for news articles. We will try to compare the result of several machine learning algorithms such as TF-IDF and vectorize method. We used Random forest, logistic regression and Naive bayes algorithms. We will analyse the result we get from these algorithm and try to find the best performance among the different feature extraction methods.For further research we will use the dataset of containing various type of article. We will work on the dataset that contains the news of various genres so that we could able to judge the efficiency.Text data is pre-processed, features are extracted using various methods, and classification models are trained using various machine learning algorithms. After attaining the result and accuracy we will analyze the performance of these models. The results of this study can be applied to increase the accuracy and effectiveness of text classification for news articles and other text-based applications. The results can be applied to develop reliable text classification algorithms that will improve data efficiency and accuracy.**

I. Introduction

Text classification is an important process in natural language processing that involves categorising or labelling text inputs according to predetermined categories. Text classification is becoming a crucial tool for many applications, including sentiment analysis, recommendation systems, and information retrieval, because to the vast amount of textual data that is available on the internet. We concentrate on text classification for news items using NLP approaches in this research work. The major intension of this research is to use various feature extraction and machine learning techniques to increase the precision and effectiveness of text classification for news articles. We want to compare the performance of several machine learning algorithms, including random forest, naive bayes, and logistic regression, tf-idf and investigate the efficacy of various feature extraction methods. A dataset of news items are labelled with predetermined categories is gathered for the research. Text data is preprocessed, features are extracted using a variety of methods, and classification models are trained using a variety of machine learning algorithms. We analyze the performance of these models using various criteria including accuracy and F1-score. The results of this study can be applied to increase the text classification accuracy and effectiveness for news articles and other text-based applications. The results can be applied to create trustworthy text categorization algorithms that will improve the efficiency and standard of information retrieval in the digital era.

II. Literature Review

Some online papers are chosen for this evaluation of the literature centre on the subject of text classification using various NLP methods. We find various relatable works on this topic in different sectors. The first study, "Automated Text Classification of News Articles: A Practical Guide" by Barbera et al. (2020), provides a thorough overview of feature extraction and machine learning methods for text classification of news items.

Barbera et al.'s study from 2020 offers a useful manual for automatic text classification of news stories. The study focuses on the application of several feature extraction techniques and machine learning algorithms that improve text classification's precision and effectiveness. The study focuses on employing feature extraction and machine learning methods including decision trees, support vector machines, and deep neural networks to categorise news items. According to the study, deep neural networks with pre-trained word embeddings produced the classification of news articles with the highest accuracy, 94%. The study's findings can be used to increase the text classification's effectiveness and efficiency for news articles and other text-based applications. Researchers looking to advance text categorization methods for news stories and other text-based applications can learn a lot from this study.

Rini Wongso's paper classes Indonesian news articles. Preprocessing, feature selection, and classification were used to classify. Pre-processing cleaned text data by removing stop words, stemming, and tokenizing. TF-IDF was used for feature selection and SVM and NB for classification. 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 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.

Extensive use of machine learning and natural language processing methods for text classification has been observed. Using a combination of preprocessing techniques, feature extraction methods, and machine learning algorithms, Hui Li and Zeming L provide a methodology for text classification. Their work contributes to the existing body of knowledge in the field of text classification by comparing the performance of different machine learning algorithms and feature extraction techniques. In comparison to 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 feature extraction method outperformed the bag-of-words model. Their investigation into and application of numerous machine learning algorithms and feature extraction strategies to the problem of text categorization yields 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 methods. Further research is needed to determine the efficacy of deep learning algorithms for text classification.

In this paper, it shows the probability to use KNN algorithm along with TF-IDF method and framework for text classifications. The framework is designed so that it can enable the classification and measurement of similarity of documents based on the required text sample. Finally, the use of both KNN algorithm and TF-IDF method has been discussed as a good choice with minor modifications in their implementation.

In this paper, Naive Bayes classifier which is used for text classification in ML and based on conditional probability of features belonging to a class by an auxiliary feature method is proposed. 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.

In this paper, Naive Bayes classifier which is used for text classification in ML and based on conditional probability of features belonging to a class by an auxiliary feature method is proposed. 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.

This paper discusses the variety of Machine Learning and Deep Learning algorithms used in text classification with their advantages and shortcomings. 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 a variety of text feature extraction, dimensionality reduction methods,existing algorithms and techniques and evaluation methods to classify the text correctly and accurately more 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 news paper article we need a dataset that contains news of various topic. One of the dataset we are using is The BBC News Archive dataset which is comprises 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.

Category	1623
File Name	1535
Title	1517
content	1698

TABLE I
BBC NEWS ARCHIVE DATASET

Business, entertainment, politics, sports, and technology are only 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 title of the collected contents and the news article is present under the content section.

We also used AG's News Topic Classification Dataset. This file contains 120,000 different examples of training for writing three-column news stories. The column containing the Class ID comes first, then the column containing the title, and finally the column containing the description.

IV. Methodology

In this research, we collected some dataset in 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 step words.

Method Selection:

TF-IDF : Term frequency or TF calculate how many times a term occurs in a document. 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.

Inverse document frequency(IDF) evaluate the how much unique the term or the word is in the document collection.

It basically divides the total number of documents in the collection by the number of documents that contains the term then the result will be logarithmically taken.

Multinomial Naive bayes: As this algorithm learns from a labelled training dataset we will try to apply this to automatically categorize articles into different topics or classes. The topic or the 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.

Logistic Regression is efficient for computing purposes and also can handle 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.

Random Forest's randomization helps prevent overfitting, and it 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 on our dataset to determine the precision, recall and the F1-score.

V. Result and Discussion

Precision means the 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 true positives and false negatives. F-1 score is a numerical average which calculates the precision and recall and the higher value will show better performance.

	Model	Test Accuracy	Precision	Recall	F1	process
0	Random Forest	25.48	0.25	0.25	0.25	tfidf
1	Random Forest	97.86	0.98	0.98	0.98	vectorize
2	Logistic Regression	20.95	0.21	0.21	0.21	tfidf
3	Logistic Regression	97.86	0.98	0.98	0.98	vectorize
4	Multinomial Naive Bayes	19.76	0.20	0.20	0.20	tfidf
5	Multinomial Naive Bayes	96.43	0.96	0.96	0.96	vectorize

Performance of the model

Here, from the result table we can see that when we apply tf-idf and vectorize process both for Random Forest method, Naive Bayes and logistic regression. Random forest and Logistic Regression give the highest value is 97.86%, F-1, recall and precision score are 0.98 for both methods. Another highest result is 96.43% which we get after applying Naive Bayes. F-1, recall and precision score are 0.96 for the method. The results are very low in the tf-idf process. Test accuracy for Random forest is 25.48%. Test accuracy for LR is 20.95% and for multinomial naive bayes is 19.76%.

After applying all the methods, we can say that naive bayes and random forest methods give us the highest accuracy in the Vectorize process. So it will fare if we use Random Forest and Logistic regression in Vectorize method for further researches.

VI. Conclusion

Our proposed study, titled "Text Classification for News Article using NLP Techniques," will expand upon the application of Natural Language Processing to text classifications. In this situation, we got varying conclusions by using several datasets with various methods. We started by reading through some online copies of previous researches works. We compared the methods Multinomial Naive Bayes, Logistic Regression and Random forest with two processes; Vectorize and TF-IDF. The results were not helpful from 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.

VII. Reference

1. Li, H., Li, Z. (2022c). Text Classification Based on Machine Learning and Natural Language Processing Algorithms. Wireless Communications and Mobile Computing, 2022, 1–12. https://doi.org/10.1155/2022/3915491

2. Barberá, P., Boydstun, A. E., Linn, S., McMahon, R., Nagler, J. (2021). Automated Text Classification of News Articles: A Practical Guide. Political Analysis, 29(1), 19–42. https://doi.org/10.1017/pan.2020.8

3. Wongso, R., Luwinda, F. A., Trisnajaya, B. C., Rusli, O. (2017). News Article Text Classification in Indonesian Language. Procedia Computer Science, 116, 137–143. https://doi.org/10.1016/j.procs.2017.10.039

4. Bahassine, Said Madani, Abdellah Kissi, Mohamed. (2017). Arabic text classification using new stemmer for feature selection and decision trees. Journal of Engineering Science and Technology. 12. 1475-1487

5. Fan, H., Qin, Y. (2018). Research on Text Classification Based on Improved TF-IDF Algorithm. In Proceedings of the 2018 International Conference on Network, Communication, Computer Engineering (NCCE 2018). https://doi.org/10.2991/ncce-18.2018.79

6.Goudjil, M. B., Koudil, M., Bedda, M., Ghoggali, N. (2018). A Novel Active Learning Method Using SVM for Text Classification. International Journal of Automation and Computing, 15(3), 290–298. https://doi.org/10.1007/s11633-015-0912-z

7. Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, 34(1), 1–47. https://doi.org/10.1145/505282.505283

8. Lewis, D. A. (1995). A sequential algorithm for training text classifiers. Sigir Forum, 29(2), 13–19. https://doi.org/10.1145/219587.219592 ACM Digital Library

9. Kowsari, K., Meimandi, K. J., Heidarysafa, M.,

Mendu, S., Barnes, L. E., Brown, D. E. (2019). Text Classification Algorithms: A Survey. Information, 10(4), 150. https://doi.org/10.3390/info10040150
10. Aggarwal, C. C., Zhai, C. (2012). A Survey of Text Classification Algorithms. In Springer eBooks (pp. 163–222). https://doi.org/10.1007/978-1-4614-3223-4_6
11. Some Effective Techniques for Naive Bayes Text Classification. (2006, November 1). IEEE Journals Magazine — IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/1704799
12. Raschka, S. (2014, October 16). Naive Bayes and Text Classification I - Introduction and Theory. arXiv.org. https://arxiv.org/abs/1410.5329
13. Zhang, W., Gao, F. (2011). An Improvement to Naive Bayes for Text Classification. Procedia Engineering, 15, 2160–2164. https://doi.org/10.1016/j.proeng.2011.08.404
14. Wang, S., Jiang, L., Li, C. (2015). Adapting naive Bayes tree for text classification. Knowledge and Information Systems, 44(1), 77–89. https://doi.org/10.1007/s10115-014-0746-y
15. Zhang, W., Yoshida, T., Tang, X. (2008). Text classification based on multi-word with support vector machine. Knowledge Based Systems, 21(8), 879–886. https://doi.org/10.1016/j.knosys.2008.03.044
16. Mitra, V., Wang, C., Banerjee, S. (2007). Text classification: A least square support vector machine approach. Applied Soft Computing, 7(3), 908–914. https://doi.org/10.1016/j.asoc.2006.04.002
17. Wan, C. F., Lee, L. H., Rajkumar, R. K., Isa, D. (2012). A hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbor and support vector machine. Expert Systems With Applications, 39(15), 11880–11888. https://doi.org/10.1016/j.eswa.2012.02.068
18. Gharib, T. F., Habib, M. B., Fayed, Z. T. (2016). Arabic Text Categorization Using Support vector machine, Naïve Bayes and Neural Network. GSTF International Journal on Computing, 5(1). https://doi.org/10.7603/s40601-016-0016-9
19. Yun-Tao, Z., Ling, G., Yong-Cheng, W. (2005). An improved TF-IDF approach for text classification. Journal of Zhejiang University Science, 6(1), 49–55. https://doi.org/10.1007/bf02842477
20. Trstenjak, B., Mikac, S., Donko, D. (2014). KNN with TF-IDF based Framework for Text Categorization. Procedia Engineering, 69, 1356–1364. https://doi.org/10.1016/j.proeng.2014.03.129
21. Liu, H. (2018, November 1). Towards Explainable NLP: A Generative Explanation Framework for Text Classification. arXiv.org. https://arxiv.org/abs/1811.00196
22. Dogra, V., Verma, S., Kavita, N., Chatterjee, P., Shafi, J., Choi, J., Ijaz, M. (2022). A Complete Process of Text Classification System Using State-of-the-Art NLP Models. Computational Intelligence and Neuroscience, 2022, 1–26. https://doi.org/10.1155/2022/1883698
23. Trstenjak, B., Mikac, S., Donko, D. (2014b). KNN with TF-IDF based Framework for Text Categorization. Procedia Engineering, 69, 1356–1364. https://doi.org/10.1016/j.proeng.2014.03.129
24. Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L. E., Brown, D. E. (2019b). Text Classification Algorithms: A Survey. Information, 10(4), 150. https://doi.org/10.3390/info10040150
25. Speech and Language Processing. (n.d.). https://web.stanford.edu/ jurafsky/slp3/
26. NLTK Book. (n.d.). https://www.nltk.org/book/