Text Classification for News Articles using NLP Techniques

JobedaKhanamRia, SadmanMajumder, MD.ReazUddin, MdSabbirHossain, MdMustakinAlam, AnnajiatAlim Rasel

***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, SVM and Naive Bayes. 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 decision trees, support vector machines, and deep neural networks, 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

The three papers chosen for this evaluation of the literature centre on the subject of text classification using various NLP methods. 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.

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.

$$t, f_{\text{t,d}} = \log_{10}(count(t, d) + 1)$$

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.

$$idf_{\text{t}} = \log_{10}(\frac{N}{df_{\text{t}}})$$

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.

$$P_{(\text{c})} = \frac{N_{\text{c}}}{N_{\text{doc}}}$$

Support Vector Machine (SVM): In our research SVM can be used to classify text documents into several groups or classes. To classify, first the text document is converted into numerical feature vector. Also, these feature vectors indicate whether a word is present or not.

By following the above phase classifiers will be tested on our dataset to determine the precision, recall and the time it requires.

V. Result and Discussion