News Article Classification Using NLP

A Project Report in partial fulfillment of the degree
Bachelor of Technology

in

Computer Science & Engineering

By

19K41A0543	**K. SRIDHAR SAI**
19K41A0544	**K. SRIKAR SAI**
19K41A0451	**P. SAI GANESH**

Under the Guidance of
Mr. Ramesh

SR
Engineering
College
Innovation . Creativity . Entrepreneurship

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

CERTIFICATE

This is to certify that the Project Report entitled **"News Article Classification Using NLP"** is a record of bonafide work carried out by the student(s) **K. SRIDHAR SAI, K. SRIKAR SAI, P. SAI GANESH** bearing Roll No(s) **19K41A0543, 19K41A0544, 19K41A0451** during the academic year 2022-2023 in partial fulfillment of the award of the degree of **Bachelor of Technology** in **Computer Science and Engineering**.

Supervisor **Head of the Department**

External Examiner

ABSTRACT

Text classification is the classical application of Natural Language Processing. Nowadays, there are many sources produce a significant amount of news every day on the Internet. Additionally, user demand for news has been steadily increasing, thus it is necessary that the news to be classified to enable users to quickly and effectively find the news based on the interests of user. User's previous interests could be utilized to analyze categories of news and to generate customised suggestions using the AI algorithms for automatic news classification. This main objective is to implement a ML model that takes news and a brief description as inputs and classify it into a specific news category.

Table of Contents

1. INTRODUCTION

Reading news is a regular hobby for almost all the people in this world. Different people likes different categories of news. Some may like sports news some other may like political news etc. and it differs from person to person. Before releasing an item, every news website organises it into categories so that users may quickly select the categories of news that interest them. For instance, every time we visit a news website, we click on the technology section because we prefer to read about the most recent technological developments. Others, though, might choose to read about politics, business, entertainment, or even sports rather than technology. The content administrators of news websites currently classify the news stories by hand. However, in order to save time, they can also incorporate a machine learning model on their websites that reads the news headline or the news's content and categorizes it.

Due to the large volume of news that these sources produce and receive, it is impossible for a human to manually label every item that arrives, let alone old news that may have been improperly labelled in the past. As a result, categorizing the news that these sources typically produce and receive can create a significant problem. The goal is to develop a machine learning algorithm that can automatically classify brief news items into labels. To do this, the algorithm will take in new information and identify which label (category) it belongs to. In a news article, recent or current events are discussed. They can be of general interest, as in daily newspapers, or they can be focused on a particular subject, as in trade or political news magazines, club newsletters, or technology news websites. Reports from people who witnessed an occurrence can be included in news articles. When we visit a news website, we must have seen that the news is separated into categories. Tech, entertainment, sports, and other popular categories can be found on practically all news websites. This post is for you if you want to learn how to categorise news categories using machine learning.

Natural language documents can be categorised based on their content using text classification datasets. Consider categorising news stories by topic or categorising book reviews according to whether they received positive or bad feedback. The organisation of consumer reviews, the identification of fraud, and language detection all benefit from text classification. While manual completion of this process takes time, machine learning models can automate it. A multi-label text classification problem, category classification for news is. An article of news is to be given one or more categories. Using a collection of binary classifiers is a common strategy for multi-label text classification. In Natural Language Processing (NLP), text classification also known as text categorization is a classical problem which aims assigning labels or tags to textual units like phrases, queries, paragraphs, and documents. It can be used for a variety of things, such as content summarization, question answering, spam detection, sentiment analysis, news categorization, and user intent classification. Text information can be found in a variety of places, including websites, emails, chat rooms, social media, tickets, insurance claims, user reviews, and questions and answers from customer care representatives, to mention a few.

2. LITERATURE REVIEW

[1] The many news outlets are what a social network gets its news from. Depending on user choices, news recommendations are frequently made. However, it does not include how news is categorised or how individuals feel about things. The program will examine which groups the students are interested in if news is organised in a social network. The news is updated often. The categorization of news is a significant subject. It was unnecessary to journal the news for reclassification. A media company is interested in learning what kinds of news are of interest to its audience. The media sector has always developed a system that completely accounts for the quantity of proposals for each news category. This made it easier for the media outlet to understand the situation.

[2] A lot of information is available and is kept in electronic form. The need for tools that could evaluate, analyse, and extract information that could aid in decision-making has arisen as a result of such data. It takes a lot of time to extract hidden information from huge databases using data mining. Its precision would be the main disadvantage. In today's expanding globe, it is crucial to give information without any faults. The Zero Frequency Problem in the naïve Bayes algorithm is the second problem. We are therefore employing the Multinomial Naive Bayes Algorithm to get over these disadvantages.

[3] Sentiment analysis, news categorization, question answering, and natural language inference are just a few of the text classification tasks where deep learning based models have outperformed traditional machine learning based methods. In this study, we present a thorough analysis of more than 150 deep learning-based text classification models created in recent years and talk about their technical merits, commonalities, and weaknesses. We also give an overview of more than 40 well-known datasets that are frequently used for text classification. Finally, we offer a quantitative evaluation of various deep learning models' performance on well-known benchmarks and highlight potential future research routes.

[4] Maintaining irregular data is a major difficulty in all applications where data plays a key role, such as universities, businesses, research institutes, technology-intensive corporations, and government funding agencies. Most of the data for an entity (an item, location, or thing) is in an erratic format. The entity connections in a dataset are currently examined by data analytics or text mining to find noteworthy patterns that reflect the data in the dataset. Decisions are made using this knowledge. Words analytics turns text into numbers and numbers, which helps to create the data and identify trends. The analysis will be better if the data is better organised, which will lead to better conclusions in the end.

It is challenging to manually process every piece of data, as well as to categorise the data. In order to examine linguistic and lexical patterns, this leads to the emergence of intelligent text processing technologies in the field of NLP. It is important to review and understand the nature of the data before mining. Automation of the text classification process is necessary due to the

growing amount of information and the need for accuracy or precision. Building elaborate "text data models employing Deep learning systems" that are capable of performing challenging NLP tasks with semantic constraints is another alluring research possibility.

Data analytics serves as the foundation for text classification and can power information exploration. These findings may be applied to emergent applications that aid in decision-making. These choices help people enhance resources and provide the majority of the advantages. Future study will focus on better parameter optimization techniques that reflect efficient knowledge discovery.

[5] Over the past few years, text mining has grown in importance considerably. Users can now access data from a variety of sources, including electronic and digital media. There are several ways to convert this data to structured form even though it is typically only available in the least structured form. It is highly desired to categorise the information in a suitable set of categories in many real-life situations. One of the most crucial elements that affects many parts is the news content. The issue of categorising news stories has been taken into consideration in this research. In addition to presenting algorithms for categorising news, this study analyses the drawbacks of various algorithmic techniques.

[6] This thesis explores the multi-label text classification of Swedish news items using pre-trained contextualised language models. On pre-trained BERT and ELECTRA models, various classifiers are constructed, exploring global and local classifier techniques. Additionally, the implications of model compression, using additional metadata characteristics, and domain specialisation are examined. To build unlabeled and labelled datasets for pre-training and fine-tuning, respectively, several hundred thousand news articles are acquired.

The results demonstrate that BERT performs noticeably better than ELECTRA and that a local classifier technique is superior to a global classifier strategy. Notably, a base classifier constructed using SVMs produces competitive results. Further in-domain pre-training has varying effects; performance for ELECTRA increases while performance for BERT remains essentially unaltered. It has been discovered that combining text representations with metadata characteristics enhances performance. The robustness of BERT and ELECTRA to quantization and pruning enables model sizes to be decreased in half without compromising performance.

[7] People are misled by the spread of fake news on the Internet, which causes them to recognise certain events in a false light and make poor decisions. This widespread threat poses a serious political, economic, and ethnic challenge to contemporary society. This paper discusses various strategies and methodologies that use various word-vector algorithms to preprocess text input. To help people determine the veracity of the news, two datasets are finally chosen to compare four of the most well-liked Natural Language Processing (NLP) models.

In this study, two datasets from Kaggle Datasets are selected, preprocessed, and evaluated on these two datasets using four popular models: Long Short Term Memory Network (LSTM), Multinomial Nave Bayes (MNB), Gaussian Nave Bayes (GNB), Random Forest (RF), and Logistic Regression (Log-Reg).

This comparison is important since a stronger model can typically increase such capability by a significant amount, helping to improve the computer's ability to detect bogus news without using up excessive amounts of human resources. Due to the dataset's extreme imbalance, things are a little different for the performance of models on dataset. Over-Sampling approach (SMOTE) was used to deal with the unbalanced dataset and prevent over-fitting; following sampling, the number of negative classes is equal to the number of positive classes. The Log-Reg model, whose recall score can reach 1.0, performs the best. LSTM is also excellent despite having a slightly longer training period than the other four types. Because there is still some noisy data after SMOTE and over-fitting, RF does not perform as well as on dataset. The outcomes for Bayes models are identical to dataset 1. Their efforts fall short of expectations. They take far less time, though, and are a wise choice when the primary focus of practical application is on training time.

3. DESIGN

Requirement Specifications (S/W & H/W)

Hardware Requirements

- ✓ **System** : Processor Intel(R) Core (TM) i5-8265U CPU @ 1.60GHz, 1800 MHz, 4 Cores, 8 Logical Processors
- ✓ **RAM** : 8 GB
- ✓ **Hard Disk** : 600 GB
- ✓ **Input** : Keyboard and Mouse
- ✓ **Output** : PC

Software Requirements

- ✓ **OS** : Windows 10
- ✓ **Platform** : Google Colab / Jupyter Notebook/ Visual Studio Code
- ✓ **Program Language** : Python

4. METHODOLOGY

4.1. Dataset

Dataset contains around 210 000 news headlines from HuffPost from 2012 to 2022. It can be used as a benchmark for a number of computational linguistic tasks. This is one of the largest news datasets. At some point of time, this dataset's first collection in 2018, HuffPost discontinued maintaining an enormous archive of news items, as it impossible to gather such a dataset today. There were around 200k headlines between 2012 and May 2018, and 10k headlines between May 2018 and 2022, as a result of website updates.

Each record in the dataset consists of the following attributes:
category: category in which the article was published.
headline: the headline of the news article.
authors: list of authors who contributed to the article.
link: link to the original news article.
short_description: Abstract of the news article.
date: publication date of the article.

There are a total of 42 news categories in the dataset. The top-15 categories and corresponding article counts are as follows:
POLITICS: 35602
WELLNESS: 17945
ENTERTAINMENT: 17362
TRAVEL: 9900
STYLE & BEAUTY: 9814
PARENTING: 8791
HEALTHY LIVING: 6694
QUEER VOICES: 6347
FOOD & DRINK: 6340
BUSINESS: 5992
COMEDY: 5400
SPORTS: 5077
BLACK VOICES: 4583
HOME & LIVING: 4320

The process of classifying news involves several steps. Classification is a challenging task since it needs to be pre-processed in order to convert text information from an unstructured form to a structured form. The primary steps in the text classification process for news articles are as follows. These include collecting data, pre-processing, feature selection, classification techniques, and assessing performance metrics.

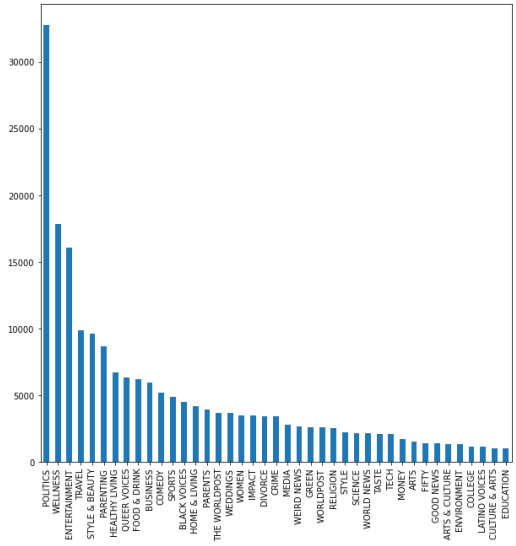

Figure 4.1. Category frequency graph

4.2. Data Collection

The gathering of news from multiple sources is the initial step in news classification. There are several places where you can find this information, including the World Wide Web, radio, television, newspapers, and magazines. However, with the expansion of the internet and information technology, it has become the primary source for news.

4.3. Data Pre-processing

Text pre-processing of the news text has to be done for futher steps. Since this data is compiled from a range of sources, it must be cleaned in order to be free of errors and useless information. In order to distinguish data from unrelated terms like semicolons, commas, double quotes, full stops, and special characters, etc., discrimination is required. Data is free from stop words, which are words that frequently exist in text.

4.3.1. Words Tokenization

News tokenization involves fragmenting the huge text into small tokens. Each word in the news is treated as a string. The output of this step is treated as input for the next steps involved in text mining.

4.3.2. Stop Words Removal

The stop words are linguistically distinctive and contain no significance. Conjunctions, pronouns, and prepositions are typically included. They are finally deleted since they are considered to be of negligible value. Before data is processed, these words must percolate. There are various ways to remove stop words from data. The removal will focus on the terms that contribute very little insight about classification, therefore it may be based on concepts.

Removing words from the list of English stop words is another method for getting rid of stop words. The list, which includes around 545 stop words, is made available by the Journal of Machine Learning Research. Depending on how frequently they are used, stop words may also be eliminated. With this procedure, word frequency is calculated before word weights are assigned. The stop words are then dropped based on these weights.

4.3.3. Word Stemming

Stemming is the next task that is carried out once stop words are eliminated. This technique is bringing a term back to its root. The purpose of stemming is to eliminate suffixes in order to reduce the amount of words. For instance, the term "USE" can be used to replace words like user, users, utilised, and using. As a result, less time and space will be needed. There are several stemmers available for stemming, including S-Stemmers, Lovins Stemmer, Porter Stemmer, and Paice/Husk Stemmer. M.F. is one of these stemmers which is most often utilised.

S-Stemmer : This stemmer is useful for words longer than three letters. This stemmer's goal is to understand about both singular and plural forms of news.

Lovins Stemmer : It was the initial stemmer that was proposed. The speed of Lovins' stemmer gives it an advantage over a number of other stemmers. Compared to other stemmers, it is speedier. There are 35 transformation rules, 294 endings, and 29 conditions in this stemmer. The 35 rules are applied on the ending after the longest endings that satisfy the condition are initially found and deleted.

Porter Stemmer : Due to its high precision and straightforward algorithm, Porter Stemmer is the most popular stemmer. It consists of five simple procedures that can be applied to each and every word.

Paice/Husk stemmer : It uses the same rules and suffixes for each loop of an algorithm that is based on iteration. Each rule consists of five steps, three of which must be followed without exception while the other two are optional.

4.4. MODEL

Bidirectional Encoder Representations from Transformers (BERT) :

Bidirectional Encoder Representations from Transformers (BERT) is a transformer-based machine learning approach for pre-training natural language processing (NLP) designed by Google. BERT is a free and open-source machine learning framework in dealing with natural language (NLP). BERT uses the surrounding text to provide context in order to help computers understand the meaning of ambiguous words in text. Jacob Devlin and his Google colleagues developed and released BERT in 2018. With the help of question and answer datasets, the BERT framework can be adjusted after being pre-trained on text from Wikipedia. Transformers is a deep learning model on which BERT is based. In Transformers, every input and output element is connected, and weightings between them are dynamically determined based on their connection. BERT primarily report results on two model sizes: BERTBASE (L=12, H=768, A=12, Total Parameters=110M) and BERTLARGE (L=24, H=1024, A=16, Total Parameters=340M). Input representation is shown in below fig. 4.2

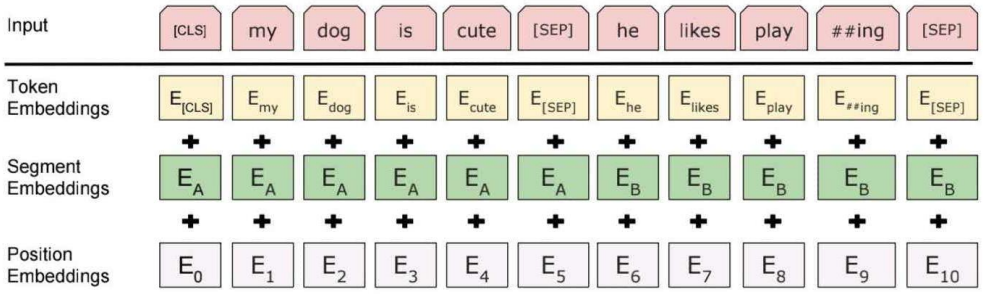

Figure 4.2. BERT input representation

This framework consists of two steps: pre-training and fine-tuning. The model is trained on unlabeled data over various pre-training tasks during pre-training. The pre-trained parameters are first used to initialise the BERT model, and then labelled data from the downstream jobs is used to fine-tune each parameter. Despite being initialised with the same pre-trained parameters, each downstream task has its own fine-tuned models.

In the past, language models could only interpret text input sequentially either from right to left or from left to right but not simultaneously. BERT is unique since it can simultaneously read in both directions. Bidirectionality is the name for this capacity, which the invention of Transformers made possible. Using this bidirectional capability, BERT is pre-trained on two different, but related, NLP tasks: Masked Language Modeling and Next Sentence Prediction. However, BERT was only trained for pre-use using an unlabeled plain text corpus (namely the English Wikipedia, and the Brown Corpus). Even while it is being utilised in practical applications, it still learns unsupervised from the unlabeled text and continues to advance (Google search). Its pre-training acts as a foundational layer of knowledge upon which to build.

In Masked Language Model (MLM) training, a word is hidden within a sentence, and the program is instructed to guess the word that has been maske based on the hidden word's context. The goal of Next Sentence Prediction training is to have the program determine if two provided sentences connect logically and sequentially or whether their relationship is just arbitrary.

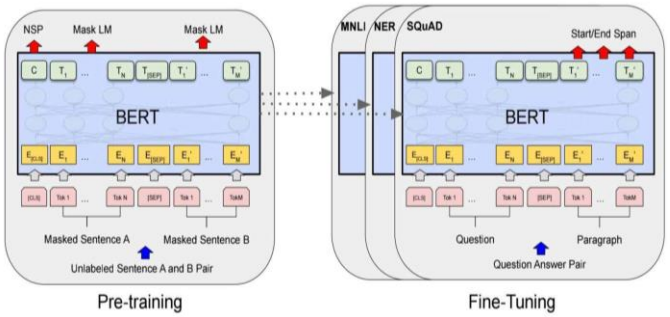

Figure 4.3. BERT Pre-training and Fine-tuning procedure

Masked LM (MLM)

Word sequences are changed with a [MASK] token for 15% of the words in each sequence before being fed into the BERT. Based on the context offered by the other, non-masked, words in the sequence, the model then makes an attempt to forecast the original value of the masked words. The prediction of the non-masked words is disregarded by the BERT loss function, which only considers the prediction of the masked values. The model's slower convergence rate compared to directional models is a result of this. Output prediction steps as follows :

1. The output of the encoder is added to a classification layer.
2. By dividing the output vectors by the embedding matrix, the vocabulary dimension is created.
3. Use softmax to determine the probability of each word in the lexicon.

Next Sentence Prediction (NSP)

In the BERT training phase, the model learns to predict whether the second sentence in a pair will come after another in the original document by receiving pairs of sentences as input. During training, 50% of the inputs are pairs in which the second sentence is the next one in the original text, and in the remaining 50%, the second sentence is a randomly selected sentence from the corpus. The underlying presumption is that the second phrase will not be connected to the first.

Before entering the model, the input is processed as follows to aid the model in differentiating between the two sentences during training:

1. The first sentence has a [CLS] token at the start, and each subsequent sentence has a [SEP] token at the end.
2. Each token has a sentence embedding that designates Sentence A or Sentence B. Token embeddings with a vocabulary of 2 and sentence embeddings share a similar notion.
3. Each token receives a positional embedding to denote its place in the sequence.

5. RESULTS

In this project millions of news articles and 40 categories were present in the dataset, out of which 70% were used for training purpose while the remaining 30% were used for model testing.Coding was carried out with Python in the Google Colab an open-source web application is employed as an Integrated Development Environment (IDE). Some library packages like Tensorflow, Keras, Matplotlib, Transformers, Sklearn, Numpy etc. were used extensively. The dataset is splitted into 70:30 ratio for training and testing. Adam with an initial learning rate LR = 3e-6 was used as an optimizer and softmax as activation function. A large amount of training was conducted to find the correct value of batch size. Initially, training was performed with parameters having different batch size and epochs. Finally it is found that a loss of 7% and 97% of model accuracy for training dataset and 70% accuracy for testing dataset. Below are few metrics, test cases, graphs for extensive analysis of results.

	description	true_category	predicted_category
75589	huffpost rise morning newsbrief, november 4wel...	POLITICS	POLITICS
21665	donald trump's lawyer claims president was nev...	POLITICS	POLITICS
51481	the feminist comic series for fans of 'strange...	ARTS & CULTURE	ARTS & CULTURE
20578	obamacare repeal moves ahead with key senate v...	POLITICS	POLITICS
58445	mount holyoke commencement speaker thanks acti...	COLLEGE	BLACK VOICES

5.1. Sample Test Cases

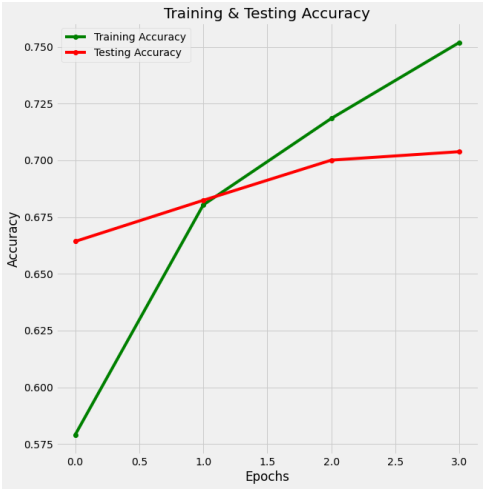

Figure 5.2. Training and Testing Accuracy

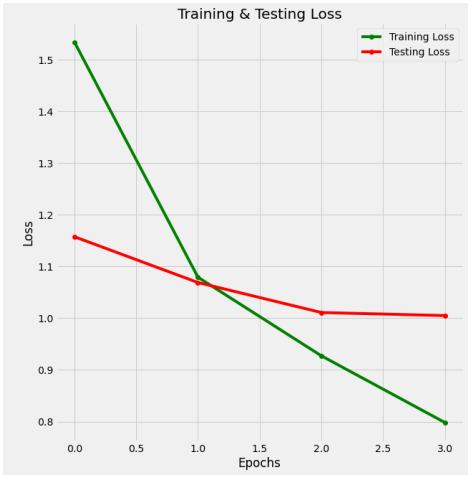

Figure 5.3. Training and Testing Loss

Model	Epoch	Batch size	Loss	Accuracy
BERT	10	32	7%	97%

Table 5.1. Training results

Model	Epoch	Batch size	Loss	Accuracy
BERT	10	32	10%	70%

Table 5.2. Testing results

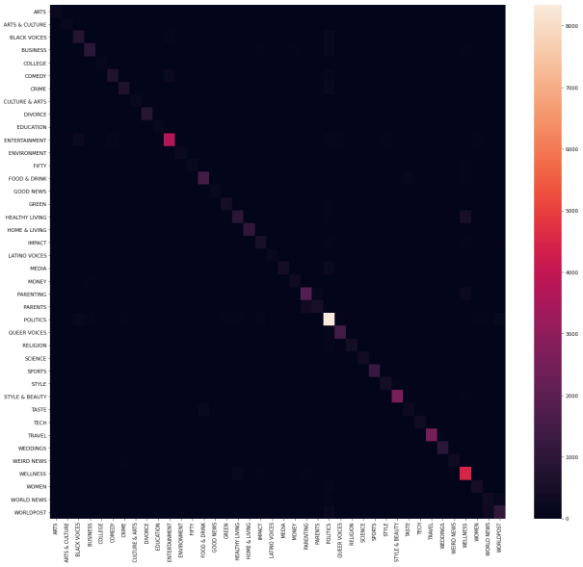

Figure 5.4. Confusion Matrix

6. CONCLUSION

In this project, we have implemented a way of classifying the news articles using NLP. The main objective is to classify the news article which enables the user to find the article based on the interests quickly. This helps the internet sources, news blogs to recommend and classify the news updates based on the historical data of user. The model was trained on a kaggle dataset which contains around two lakhs news article samples, their headlines, description and the category. We used BERT algorithm for word embeddings as well as classification and obtained 70% of model accuracy.

7. REFERENCES

[1] RitikPatil, Rohan Patil, Prathamesh Patil, News Classification usingNatural Language Processing

[2] Sundarababu, Mr & Chandramohan, Ch & Suthar, Mahendra & Harsha, Ch & Juveria, Lubna & Blessy, B & Mohammad, Sameer. (2020). NEWS CLASSIFICATION USING MACHINE LEARNING. SSRN Electronic Journal. 7. 657-660.

[3] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2020. Deep Learning Based Text Classification: A Comprehensive Review. 1, 1 (January 2020),

[4] Johnson Kolluri, Shaik Razia, Soumya Ranjan Nayak, Text classification using Machine Learning and Deep Learning Models

[5] Kaur, Gurmeet & Bajaj, Karan. (2016). News Classification using Neural Networks. Communications on Applied Electronics. 5. 42-45. 10.5120/cae2016652224.

[6] Lukas Borggren, Ali Basirat, Marco Kuhlmann, Automatic Categorization of News Articles With Contextualized Language Models

[7] Peiyang Yu et al 2021 J. Phys.: Conf. Ser. 1802 042010, Text Classification by using Natural Language Processing

[8] Gothane, Suwarna. (2021). Fake News Detection and Classification using Natural Language Processing. International Journal for Research in Applied Science and Engineering Technology. 9. 1837-1841. 10.22214/ijraset.2021.34700.