|  |
| --- |
| **Fake News Detection Using NLP** |
|  |
|  |
|  |



**Abstract**

The spreading of fake news has given rise to many problems in society. It is due to its ability to cause a lot of social and national damage with destructive impacts. Sometimes it gets very difficult to know if the news is genuine or fake. Therefore it is very important to detect if the news is fake or not. Fake News is a term used to represent fabricated news or propaganda comprising misinformation communicated through traditional media channels like print, and television as well as non-traditional media channels like social media. Techniques of NLP and Machine learning can be used to create models which can help to detect fake news.

**Problem Statement**

Since a lot of time is spent by users on social media and people prefer online means of information it has become difficult to know about the authenticity of the news. People acquire most of the information by these means as it is free and can be accessed from anywhere irrespective of place and time. Since this data can be put out by anyone there is lack of accountability in it which makes it less trustable unlike the traditional methods of gaining information like newspapers or some trusted source. Fake news is dangerous as it can deceive people easily and create a state of confusion among a community. This can further affect the society badly .The spread of fake news creates rumours circulating around and the victims could be badly impacted.

The main aim of the project is to obtain a model which will help in detecting if a news article is fake or not. The problem of detecting fake news is a very difficult task and many researchers are trying to obtain a solution to it.

**Design Thinking**

Firstly, the datasets are collected .The datasets are then merged to obtain a master dataset. This dataset is then preprocessed. Preprocessing of the datasets include lowering of the data, stop word removal, stemming, tokenization and padding is also performed in order to obtain the same length.

1. ***Tokenization****:* Tokenization is the process of breaking down a stream of text into tokens, which can be words, phrases, symbols, or any other significant items. This step's major purpose is to extract individual words in a sentence. The tokenization is done on each text in the dataset.
2. ***Stop Words****:* Stop words are the commonly used words and are removed from the text as they do not add any value to the analysis. These phrases have little or no meaning. A list of terms that are regarded as stop words in the English language is included in the NLTK library. All the stop words from the texts are removed.
3. ***Capitalization****:* Sentences can have a combination of capital and lowercase letters. A written document is made up of multiple sentences. One of the method for reducing the issue space is to convert everything to lower case. This aligns all of the words in a document in the same location. Using the python function, all the words are converted to lower case.
4. ***Stemming****:* Stemming is the process of reducing the words to its root form by eliminating extraneous characters. Porter Stemmer is one of the stemming model which is used here to convert the words into its root form.
5. ***Lemmatization:*** Text lemmatization is the process of removing a word's superfluous prefix or suffix and extracting the basic word. All the suffixes and prefixes from the words are removed to reduce space.

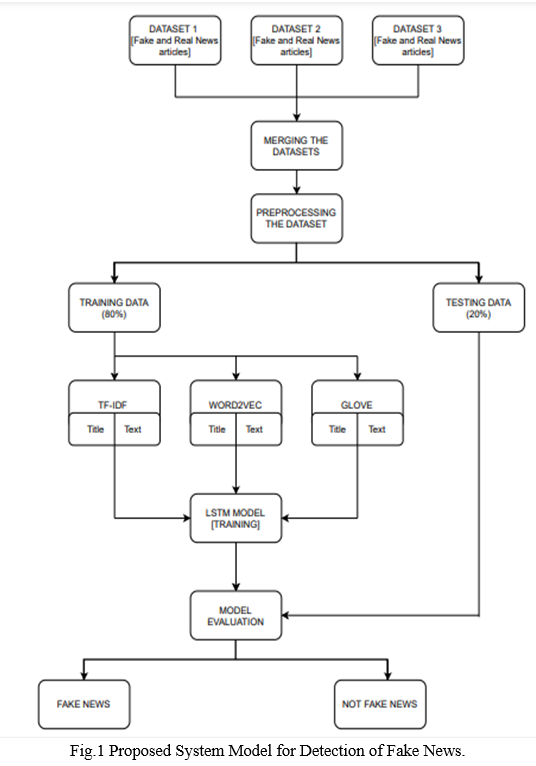
The dataset is then split into training data and testing data. To overcome the problem of detecting fake news this project proposes 6 similar LSTM models which are to be trained and each model will be fed with the different text vectors of news headline and news content. This will help in obtaining a good model which will tell if the news is true or it is fake. In this project we have used six similar LSTM models.

Three text vectorization techniques are used which are GloVe, Word2vec and TF-IDF. The first LSTM model will be fed with the vectors of the title of the news using GloVe. The second model will be fed with the vectors of the content of the news using GloVe. Similarly, two models will be built using the Word2vec technique each for the title of the news and the content of the news respectively.

Lastly, the LSTM model will be fed with the text vectors of the title of the news using TF-IDF and another model will be fed with the text vectors of the content of the news using TF-IDF. By doing so we can identify which technique gives better results and identify which model performs well. Lastly, the performance is measured using the performance metrics accuracy, precision and recall.

The libraries used are

* Pandas: For importing the dataset.
* Seaborn /Matplotlib : For data visualization.
* NLTK(Natural language toolkit): For preprocessing



**Dataset Link:** [**https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset**](https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset)

# Conclusion

Fake news have increased in recent years and it has caused a lot of harm to the society. This project aimed to develop a model using the techniques of NLP and ML to detect if a news article/headline is fake or not and identify which methods give better output.