**Fake News Detection using NLP**

**Objective and goal of the project:**

The objectives are to develop and implement automated systems or models that can accurately identify and distinguish between genuine, reliable information and false or misleading information in textual content, especially in the context of news and social media. The main aim of the project is to obtain a model which will help in detecting if a news article is fake or not.

**1.Data Collection:**

A dataset of news articles, both real and fake was collected which are from the Kaggle dataset.

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

**2. Data Preprocessing:**

1. The text data was cleaned by removing HTML tags, special characters, and irrelevant information.
2. The text is tokenized into words or subword tokens.
3. Stemming or lemmatization is performed to reduce words to their root forms.

**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.

**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.

**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.

**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.

**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.

**3.Data Splitting:**

The dataset is split into training data and testing data.

**4.Model Selection:**

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.

**5. Model Evaluation:**

The performance is measured using the performance metrics accuracy, precision and recall.

**6. Monitoring and Maintenance:**

By continuously monitor the model's performance in a production environment and retrain it as needed to adapt to evolving fake news tactics.

**7. Scalability and Deployment:**

The selected model is deployed in intended environment, whether it’s a web application , mobile app or other platforms.

**8. User Interface:**

A user-friendly interface for users is developed to input news articles and receive classification results for public use.

**9. User Education:**

We should educate users on how to critically evaluate news sources which include information about how your fake news detection system works.

**10.Documentation:**

Clear and detailed documentation is maintained which includes data sources, model architecture, and instructions for using the system.