Approach

Note: detailly mentioned in the Markdowns

Our project on fake news detection involves several key steps and methodologies to achieve accurate and reliable results. The following is a brief overview of the approach we adopted:

Dataset Selection:

We selected the ISOT Fake News dataset, which consists of a collection of satirical and legitimate news articles. The dataset was obtained from authentic resources, including verified information gathered through a Reuters.com article crawl. Hoaxes were sourced from questionable outlets previously identified by fact-checking organizations like Politifact and Wikipedia. This diverse dataset covers a wide range of article types and subjects, with a majority focusing on politics and international affairs.

Data Cleaning and Preprocessing:

To prepare the dataset for analysis, we conducted thorough data cleaning and preprocessing. This involved normalizing the text by replacing links, numbers, and emails with placeholders. We removed HTML tags, eliminated punctuation, and converted the text to lowercase. Additionally, stemming techniques were applied to reduce words to their root forms, and common stopwords were removed. While the data was processed and sanitized, the original punctuation and errors specific to false news were retained.

Parallelization for Efficiency:

To expedite the preprocessing steps, we implemented parallelization techniques. The dataset was divided into smaller chunks, and parallel processing was performed on these chunks. By leveraging the multiprocessing capabilities, we achieved faster and more efficient data cleaning and preprocessing.

Feature Extraction:

Feature extraction plays a crucial role in fake news detection. We applied Natural Language Processing (NLP) techniques to transform the preprocessed text into numerical features. This involved generating uni-, bi-, and tri-grams, which capture different patterns and relationships within the text. These features serve as input for our machine learning models.

Machine Learning Models:

We trained and evaluated several machine learning models for fake news detection. The models we employed include Naive Bayes, Support Vector Machine (SVM), Logistic Regression, and Multilayer Perceptron (MLP). Each model was trained on the preprocessed dataset and evaluated using performance metrics such as accuracy, precision, recall, and F1 score.

Performance Comparison and Selection:

To identify the most effective model for fake news detection, we compared the performance of the different machine learning models. We analyzed their accuracy, precision, recall, and F1 score on the evaluation dataset. Based on these metrics, we determined the model that achieved the highest performance and selected it as our primary fake news detection model.

System Implementation:

Once the model was selected, we implemented it as a fake news detection system. The system takes a news article as input and processes it using the preprocessing techniques and feature extraction methods discussed earlier. The selected machine learning model then classifies the article as fake or genuine based on its learned patterns and features.

Evaluation and Fine-tuning:

We performed extensive evaluation and fine-tuning of our fake news detection system. We assessed its performance on a separate validation dataset, analysing metrics such as accuracy, precision, recall, and F1 score. We made adjustments and optimizations as needed to improve the system's accuracy and effectiveness.