**Team Name - Data Wizards**

**Fake News Detection**

**Project Report**

## Introduction

The proliferation of digital media in recent years has drastically changed people’s patterns of news consumption. With the help of social media platforms and Websites, news now circulates just with the "speed of light," with real news and fake news being difficult to tell apart. The phenomenon of “fake news” the deliberate spread of misinformation or false news presented as legitimate news — is deeply toxic to public discourse, democratic processes, and public safety.

The goal of this work is to build a traditional machine learning-based FND system. This system reads the actual text of news articles and determines which are "real" and which are "fake." Through the use of NLP, we are using models like Logistic Regression, Naive Bayes, SVM, Decision Tree, and Random Forest to distill patterns that differentiate truth telling from misinformation.

This work also investigates the hyper-parameter tuning such that these classifiers perform better. The objective is not only to obtain high accuracy but also to set a basic framework which can be extended by the use of more advanced NLP techniques (i.e. transformers or deep learning).

## Related Work

Fake news detection is an extensively studied area, ranging from rule-based systems to sophisticated deep learning models. Initial approaches based on manual curation and rule-based heuristics tried to identify sensational keywords, and non-credible source domains. Although these approaches offered some levels of understanding, they were not scalable and were not capable of evolving along with fake news.

The statistical and machine learning models came the next stage of advancement. Classic methods like Naive Bayes or Logistic Regression have been the standard solutions for spam filtering and sentiment analysis, and can serve the purpose of a preliminary task-oriented classifier for fake news. Such models are able to utilize word frequency and document-level statistics to train decision boundaries between fake and real articles.

SVMs and DTs are more flexible by letting nonlinear patterns to be handled while also offering interpretability. Methods such as Random Forests improve robustness through performing aggregation of predictions of many weak learners to mitigate overfitting. However, state-of-the-art neural models e.g., long short-term memory (LSTM) networks, bidirectional encoder representations from transformers (BERT) and graph neural networks (GNNs) further improved performances which are computationally expensive and depend on large amounts of data.

In spite of the success of deep learning, classic machine learning models are still important for their simplicity, interpretability and ease of use, particularly on small to moderate sized data. Our approach takes this pragmatic direction: we try to find this balance between effectiveness and interpretability.

## Methodology

The methodology followed in this project consists of several clearly defined stages:

1. Dataset Acquisition:

* As the label of the dataset indicates there are new articles with ‘REAL’ and ‘FAKE’ label.
* It comes with functionalities, like title, text and label columns.

1. Data Cleaning:

* Null values and duplicates are dropped.
* Removed unnecessary spaces and unwanted columns.
* Removed empty titles or content after cleaning up.

1. Text Preprocessing:

* All text was normalized to lowercase to minimize case-dependency.
* Punctuation, special characters except @ and #-, and numeric were stripped using regular expressions.
* Tokenizing was performed using the TreebankWordTokenizer from NLTK.
* Common but uninformative words were removed using stopwords.
* Stemming was done with the use of Porter Stemmer to reduce the words to their root form.

1. Feature Engineering:

* A description of the text features was vectorized through a process of TF-IDF (term frequency-inverse document frequency).
* This method highlights terms that are common in a document, but rare in the corpus.

1. Label Encoding:

* The 'label' column was coded numerically to use in the model.

1. Model Training:

* Five classifiers including Logistic Regression, Naive Bayes, Support Vector Machine (Linear SVC), Decision Tree and Random Forest were trained.
* A train-test split (80:20 is standard) was used.

1. Hyperparameter Tuning:

* Fine tuning of each model was done through its performance in terms of accuracy via GridSearchCV.
* Parameters including regularization strength, tree depth, and number of estimators were tuned.

1. Evaluation Metrics:

* The accuracy, precision, recall and F1 score were calculated.
* Visualization of classification rendered confusion matrices.
* Model comparison before and after tuning was represented by a bar chart.

## Results and Discussion

The findings reveal that classical machine learning techniques can discriminate fake news quite well when receiving well-preprocessed text. Below are key observations:

* Logistic Regression gave the highest general accuracy (93.25%) since it was tuned, affirming its stability to linearly separable data and fair performance.
* Support Vector Machine exhibited high performance (93.17%), which is indicative that SVM is effective on high-dimensional text data.
* Random Forest obtained comparable accuracy (92.46%) and it was more interpretable and resistant to overfitting.
* Naive Bayes, proved little accuracy of 88.02%. Its simplicity makes it quick to compute, but less versatile to model complex word relationships.
* Decision Tree obtained the worst accuracy (79.68%), due to the overfitting problem on the training data unless you deal with the tree-pruning.

There were significant benefits in Decision Trees and Logistic Regression after tuning hyperparameters. Based on confusion matrices and classification reports for each model, the majority of misclassified predictions were between fake articles with complex language and real news that does not adhere to a sound structure.  
  
The bar graph in before-after consistency comparison makes a strong case in favor of calibration for that task. Logistic Regression and Random Forest showed good stability indicating their robustness.

## Conclusion and Future Work

* The project has accomplished a production-ready Fake News Detection system utilizing several machine learning models and NLP pre-processing techniques. Logistic Regression and SVM were the two best performing models, achieving above 93% accuracy on the test set. The Random Forest has also been good and is chosen because it is understandable and robust.
* Effective text cleaning, TF-IDF vectorization, and hyperparameter tuning considerably improved the power of the models to differentiate between real and fake news. These findings once again prove the importance of classical methods, properly applied to structured problems.

**Future improvements can include:**

* Utilizing state-of-the-art NLP models such as BERT to improve semantic comprehension.
* Stacking by adding an ensemble, the power of the different classifiers is then pooled.
* To improve the generalization, train and label on a diversified and updated dataset.
* Pairing your own live-scraping tools and APIs to tracking the news as it breaks.
* Investigating explainable AI (XAI) methods to make model predictions clear and understandable to users.

In general, the work of this project demonstrates that classical machine learning can be relevant and have a real impact on real-world applications such as fake news detection.