**Research Report on Fake News Detection Using NLP**

**Introduction**

Fake news has become a critical issue in today's digital world. The project aims to develop a machine learning-based system to detect fake news articles using Natural Language Processing (NLP). Initially, Random Forest was considered due to its strong macro F1-score performance. However, after further evaluation, Logistic Regression was selected as the final model due to its better generalization and balanced classification capability.

**Objective**

To create a robust model that can accurately distinguish between real and fake news articles using machine learning and natural language processing techniques.

**Dataset Overview**

* Source: [Kaggle Fake and Real News Dataset]
* Columns: title, text, label
* Size: ~45,000 articles

**Data Preprocessing**

* Removed null values and duplicates
* Tokenization
* Stopword removal
* Lemmatization
* Label encoding (0 for fake, 1 for real)

**Exploratory Data Analysis (EDA)**

* Word clouds for both fake and real news to identify common words
* Bar plots of word frequency
* Distribution of news types (balanced dataset)

**Feature Extraction**

* Used **TF-IDF Vectorization**
* Captured important words while reducing common noise

**Model Evaluation (Initial Phase)**

We tested three models:

1. **Random Forest** - Strong F1-score, especially on training data, but generalization on test data was weak; tended to overfit.
2. **Multinomial Naive Bayes** - Lightweight and fast but less accurate.
3. **Logistic Regression** - Balanced performance, low overfitting, and excellent generalization.

**Final Model: Logistic Regression**

We selected Logistic Regression due to:

* Low number of false positives/negatives
* Better performance on unseen (test) data
* Simpler and interpretable model

**Confusion Matrix (Test Set)**

* **True Fake predicted as Fake:** 4600
* **Fake predicted as Real:** 69
* **True Real predicted as Real:** 4273
* **Real predicted as Fake:** 38

**Metrics**

* Accuracy: 0.9847
* Precision: 0.9839
* Recall: 0.9826
* F1 Score: 0.9832

**Conclusion**

While Random Forest initially appeared optimal, further testing revealed overfitting issues and instability in real/fake balance. Logistic Regression proved to be more stable and generalizable. The model successfully classifies news articles with high accuracy and minimal misclassification.

**Tools Used**

* Python, Pandas, NumPy
* Scikit-learn
* Matplotlib, Seaborn
* NLTK, WordCloud
* Jupyter Notebook

**Next Steps**

* Develop a web application for public use
* Continue improving the model with more data sources

**End of Report**