**NATIONAL UNIVERSITY OF COMPUTER AND EMERGING SCIENCES-FAST**

**KARACHI CAMPUS**



**COURSE INSTRUCTOR:**

**Miss Summaiyah Zahid**

**AI4001-Fundamentals of Natural Processing Language**

**TITLE: Fake News Detection**

**SUBMITTED BY:**

**20K-1706(Taha Sabir)**

**20K-1669(Hamza Jafri)**

**SECTION: BAI-7A**

**Abstract**

In this project, we explore the application of Natural Language Processing (NLP) and various machine learning algorithms for the detection of fake news. The dataset used consists of labeled examples, with class 0 denoting fake news and class 1 denoting true news. We employed multiple machine learning algorithms, including Logistic Regression, Decision Tree Classifier, Gradient Boosting Classifier, and Random Forest Classifier, to predict and evaluate the accuracy of the models.

**1. Introduction**

**1.1 Background**

Fake news has become a significant issue in today's information age. This project aims to leverage NLP techniques and machine learning algorithms to distinguish between fake and true news articles.

**1.2 Objectives**

* Apply NLP techniques for feature extraction from textual data.
* Implement and compare various machine learning algorithms for fake news detection.
* Evaluate the performance of the models using accuracy metrics.
* Test the models on a separate manual testing file to assess real-world performance.

**2. Dataset**

**2.1 Data Collection**

The dataset used comprises a collection of news articles labeled as either fake (class 0) or true (class 1).

**2.2 Data Preprocessing**

* Text cleaning: Removal of stop words, punctuation, and special characters.
* Vectorization: Converting text data into numerical form using techniques like TF-IDF.

**3. Methodology**

**3.1 Feature Extraction**

NLP techniques were applied to extract relevant features from the textual data, including TF-IDF scores and word embeddings.

**3.2 Machine Learning Algorithms**

The following machine learning algorithms were implemented and compared:

* Logistic Regression
* Decision Tree Classifier
* Gradient Boosting Classifier
* Random Forest Classifier

**3.3 Model Training and Evaluation**

Each algorithm was trained on the labeled dataset, and its performance was evaluated using metrics such as accuracy, precision, recall, and F1 score.

**4. Results**

**4.1 Model Performance**

The accuracy and other relevant metrics for each algorithm are summarized below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1-score |
| Logistic Regression | 0.987 | 0.99 | 0.99 | 0.99 |
| Decision Tree Classifier | 0.995 | 1.00 | 1.00 | 1.00 |
| Gradient Boosting Classifier | 0.995 | 1.00 | 0.99 | 1.00 |
| Random Forest Classifier | 0.989 | 0.99 | 0.99 | 0.99 |

**5. Conclusion**

In conclusion, our investigation into fake news detection using NLP and various machine learning algorithms has yielded valuable insights. The models, including Logistic Regression, Decision Tree Classifier, Gradient Boosting Classifier, and Random Forest Classifier, have shown varying levels of effectiveness in distinguishing between fake and true news articles.

The Logistic Regression model demonstrated a commendable balance between precision and recall, making it a reliable choice for overall accuracy. Decision Tree Classifier exhibited interpretability but was sensitive to overfitting. Gradient Boosting Classifier and Random Forest Classifier both performed well, with Gradient Boosting showing superior precision.

**6. Future Work**

Moving forward, there are several avenues for enhancing the fake news detection model:

1. **Incorporating Deep Learning Models:** Exploring deep learning architectures, such as recurrent neural networks (RNNs) or transformers, could capture more intricate patterns in textual data and improve the model's performance.
2. **Fine-tuning NLP Techniques:** Experimenting with advanced NLP techniques, such as pre-trained language models like BERT or GPT, may further refine feature extraction and contribute to better model accuracy.
3. **Ensemble Approaches:** Combining the strengths of multiple models through ensemble techniques could potentially enhance overall predictive power and robustness.
4. **Real-Time Application Integration:** Adapting the model for real-time applications would be crucial for addressing the dynamic nature of fake news dissemination. This involves optimizing the model's efficiency to process and classify information in near real-time.

**THANK YOU**