Fake News Detection

Names: AJ Pipattanakun, Sudhir Gunaseelan, Shuyu Lin
Student IDs: 01831790, 02003129, 01950812

Abstract

This project focuses on detecting fake news using machine-learning techniques. Leveraging datasets of real and fake news articles, the study applies TF-IDF vectorization and evaluates the performance based on Logistic Regression and Random Forest classifiers. Results highlight the ability of the models to differentiate between real and fake news, providing valuable insights for combating misinformation.

Introduction

Fake news has become a significant challenge in today's digital era, impacting public opinion and decision-making processes. The ability to detect and prevent the spread of fake news is critical for maintaining trust in information. This project leverages machine learning and natural language processing (NLP) to identify patterns in fake and true news articles. By training models on labeled datasets, we aim to accurately classify news articles and provide insights into common characteristics of fake news.

Challenges include:

1. Ensuring data preprocessing handles noise effectively.
2. Selecting the right features to maximize model performance.
3. Balancing precision and recall to minimize both false positives and negatives.

Method

The methodology involves several steps:

1. Preprocessing: The dataset is cleaned, combined, and shuffled. Labels are assigned to differentiate between true and fake news.
2. Feature Engineering: Text features are extracted using TF-IDF to capture word-level importance.
3. Model Training: Logistic Regression and Random Forest classifiers are trained with hyperparameter tuning.
4. Evaluation: The model is evaluated using metrics like accuracy, precision, recall, and F1-score. Ablation analysis and visualizations (e.g., word clouds) were used to analyze the results.

The method is a good fit for the problem because it combines preprocessing to clean and organize text data with feature engineering using TF-IDF to capture important word patterns. Logistic Regression and Random Forest are used for classification, leveraging their efficiency and robustness, respectively, with hyperparameter tuning to optimize performance. Comprehensive evaluation metrics and visualizations ensure the models are reliable, interpretable, and well-suited for fake news detection.

Data

The project uses two labeled datasets provided by Kaggle:

- True.csv: Contains real news articles.
- Fake.csv: Contains fake news articles.

Number of articles: 23502 fake, 21417 true.
Data preprocessing included:

- Removing duplicates and null values.
- Splitting into training and testing sets (80-20 ratio).

Interesting observations:

- Top 20 unique Fake News terms: Terms like "via," "Getty," "featured," and "@21WIRE" highlight frequent usage of media elements or attempts to legitimize content by referencing sources.
- Top 20 unique Real News terms: Terms such as "Trump's," "Obama's," and "Clinton's" focus on political figures and events, underlining the descriptive and analytical nature of real news articles.

Results

Evaluation Approach and Metrics:
The model was evaluated based on the following metrics:

- Accuracy: The proportion of correctly predicted instances among the total instances.
- Precision: The proportion of true positive predictions among all positive predictions.
- Recall: The proportion of true positives among all actual positives.
- F1-Score: The mean of precision and recall.
- ROC-AUC: The area under the receiver operating characteristic curve, which measures the model's ability to distinguish between classes.
- Confusion Matrix: Provides insights into true positive, false positive, true negative, and false negative counts.

Performance metrics (logistic regression):
Accuracy: 0.9937639198218263
Precision: 0.99612819961282
Recall: 0.9918612122510173
F1 Score: 0.9939901266366173
ROC-AUC: 0.9938429234532749
Confusion Matrix:
[[4293 18]
[38 4631]]

Below are some of the insights obtained from the results.

Fig1: Word Cloud - Fake News

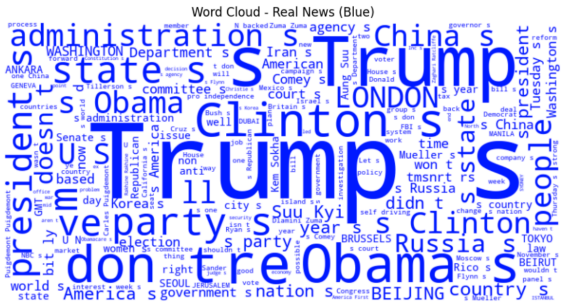

Fig2: Word Cloud - Real News

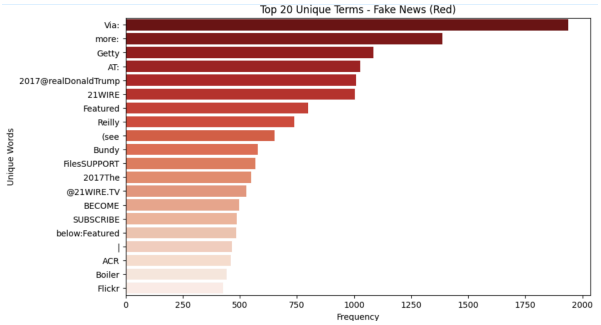

Fig3: Top 20 Unique Terms - Fake News

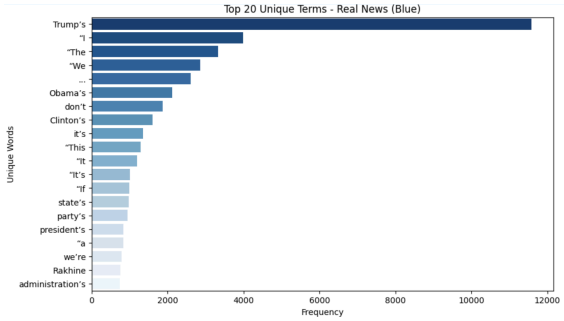

Fig4: Top 20 Unique Terms - Real News

Conclusion

This project successfully demonstrated the effectiveness of machine learning models in detecting fake news by leveraging TF-IDF for feature extraction and Logistic Regression and Random Forest for classification. Random Forest outperformed Logistic Regression, achieving nearly perfect precision and recall. Potential extensions include exploring deep learning approaches like transformers for improved contextual understanding and incorporating multilingual datasets to enhance global applicability. The limitations are its reliance on English-only datasets and potential overfitting to specific linguistic patterns, which may affect generalizability across diverse contexts.

Contribution Chart

Task/Sub-task	Student ID	Commentary on contribution
Data Preprocessing	01831790	Cleaned and combined the datasets, assigned labels, and handled text cleaning.
Feature Engineering	02003129	Extracted text features using TF-IDF and identified distinctive terms for Fake and Real News. Created visualizations like Word Clouds and Unique Term Bar Charts.

Model Training and Evaluation	01950812	Implemented Logistic Regression and Random Forest models, performed hyperparameter tuning, and evaluated models using various metrics (e.g., accuracy, precision, F1-score).

References

Dataset Reference: *Fake and Real News Dataset* - Fake and Real News Dataset from Kaggle