Fake News Detection Report

Introduction

The rapid spread of misinformation in the digital age poses a significant challenge to public trust and information integrity. With the vast volume of news articles published daily, distinguishing between credible and misleading content has become increasingly difficult. This report details the development of a semantic classification model to address this issue, utilizing the Word2Vec method to capture semantic relationships in text and classify news articles as true or fake. The objective was to build a robust system using supervised learning techniques, leveraging datasets containing true and fake news articles to identify patterns and improve automated detection of misinformation.

Data Preparation and Preprocessing

Data Overview

The analysis utilized two datasets: True.csv (21,417 true news articles) and Fake.csv (23,502 fake news articles). Each dataset contained three columns: title, text, and date of publication. The datasets were combined into a single DataFrame, with a news_label column added to indicate whether an article was true (1) or fake (0). The combined dataset was cleaned to handle missing values, duplicates, and inconsistencies, ensuring data quality for subsequent analysis.

Text Preprocessing

Text preprocessing was critical to prepare the data for semantic analysis. The following steps were applied:

- **Lowercasing**: Converted all text to lowercase to ensure consistency.
- **Removing Special Characters and Punctuation**: Used regular expressions to eliminate non-alphanumeric characters and punctuation.
- **Tokenization and Stopword Removal**: Employed NLTK for tokenization and removed common stopwords to focus on meaningful words.
- **Lemmatization**: Applied spaCy's en_core_web_sm model to reduce words to their base forms, preserving semantic meaning.
  The processed text was stored in a lemmatized_text column, combining cleaned titles and article bodies to capture comprehensive semantic content.

Train-Validation Split

The dataset was split into training (70%) and validation (30%) sets using stratified sampling to maintain the proportion of true and fake news. This resulted in 31,391 training samples and

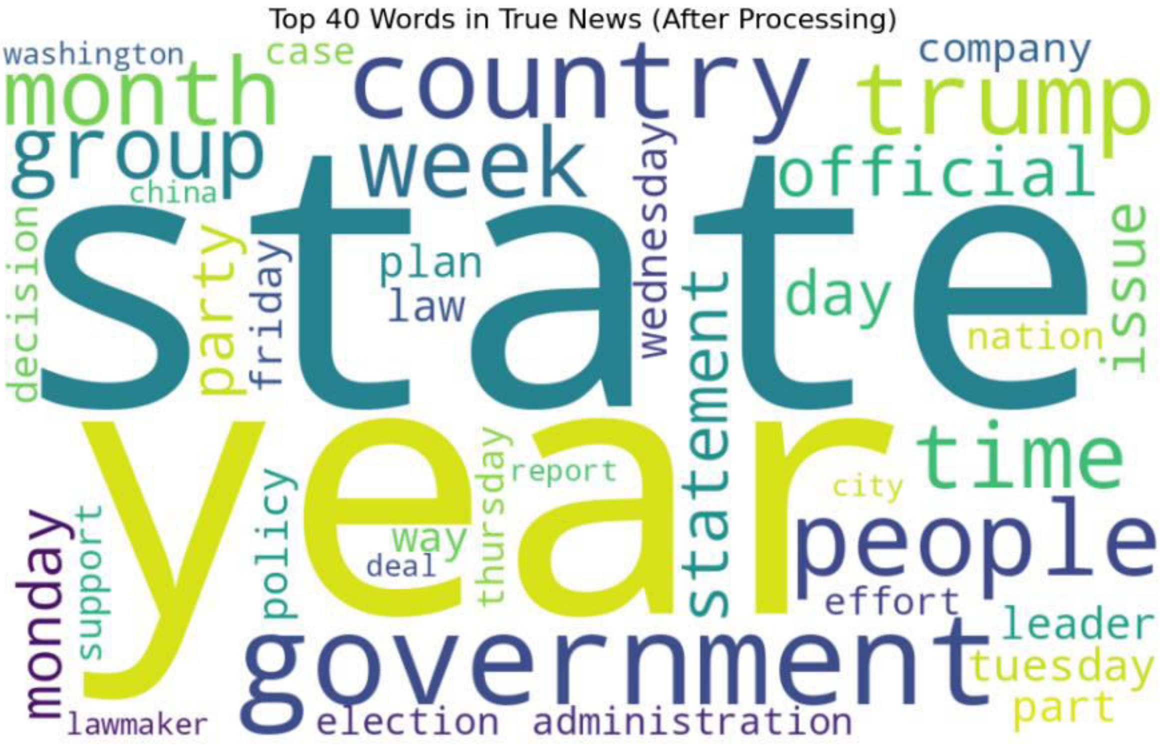
13,454 validation samples, ensuring a balanced representation for model training and evaluation.

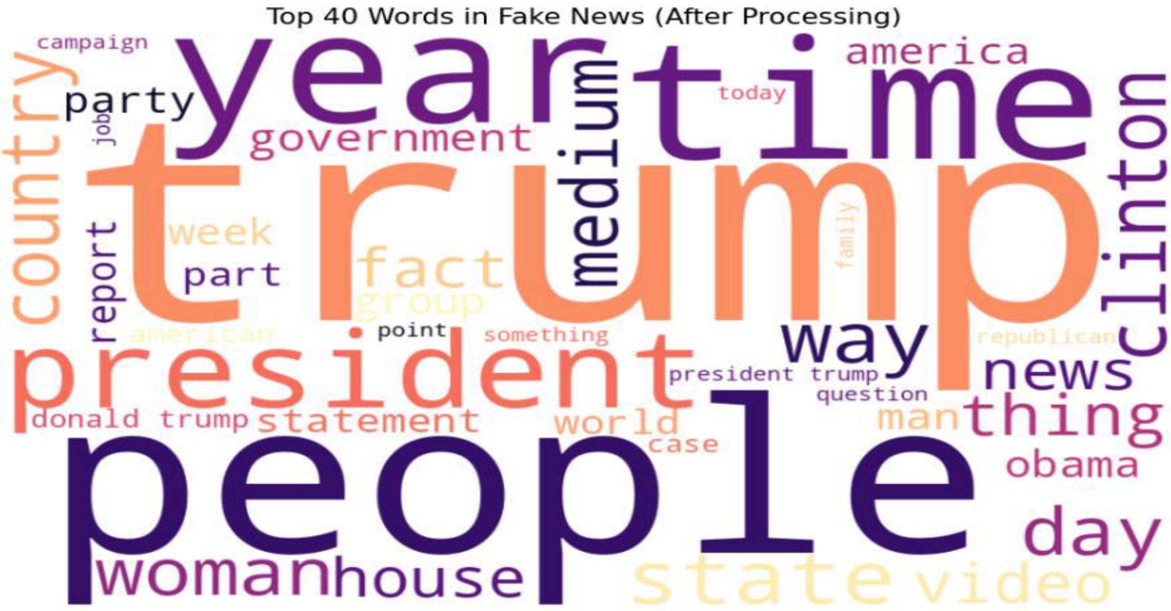
Exploratory Data Analysis (EDA)

Training Data Insights

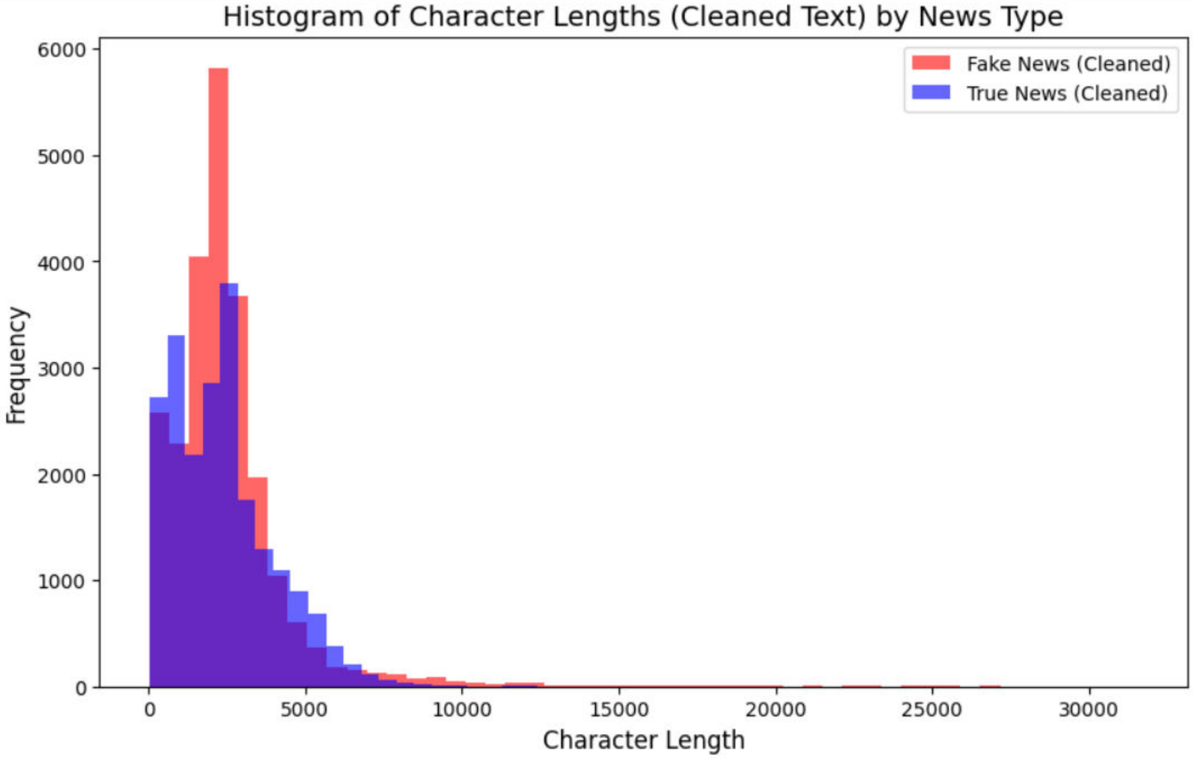
EDA revealed distinct patterns in true and fake news:

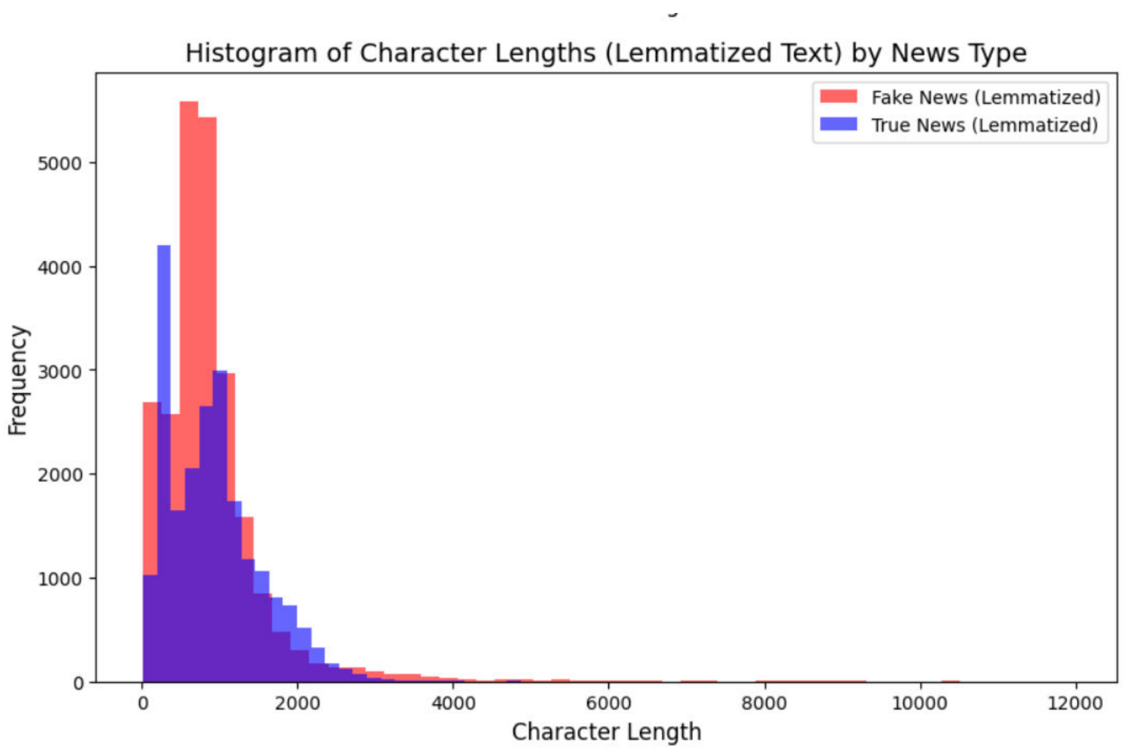
- **True News**: Characterized by structured, factual language with frequent mentions of entities like locations (e.g., "United States"), organizations (e.g., "Reuters"), and official terms (e.g., "government"). Bigrams like "White House" and trigrams like "United States government" highlighted a geopolitical focus.
- **Fake News**: Exhibited sensational and emotive language, with terms like "shocking," "exposed," and "conspiracy" appearing frequently. Bigrams such as "deep state" and trigrams like "fake news media" indicated alarmist and subjective content.
- **Word Clouds**: Visualizations showed true news emphasizing verifiable entities, while fake news leaned toward hyperbolic and persuasive terms.


Top 40 Words in True News (After Processing)



- **Article Length**: False news articles were generally longer and more provocative, while True news tended to be shorter and more detailed.





Feature Extraction

Feature extraction was performed using the pre-trained word2vec-google-news-300 model, which converted lemmatized text into 300-dimensional vectors by averaging word embeddings for each article. This approach captured semantic relationships, enabling the model to understand contextual nuances. The resulting training and validation vector shapes were (31,391, 300) and (13,454, 300), respectively, with corresponding labels extracted for supervised learning.

Model Training and Evaluation

Three supervised models were trained and evaluated: Logistic Regression, Decision Tree, and Random Forest. The F1-score was prioritized as the primary evaluation metric due to its balance of precision and recall, critical for minimizing both false positives (misclassifying true news) and false negatives (missing fake news) in the context of misinformation detection.

Logistic Regression

- **Training**: A Logistic Regression model with the liblinear solver was trained on the training vectors.
- **Performance**:
  - Accuracy: 0.9337

    - Precision: 0.9255
    - Recall: 0.9365
    - F1-Score: 0.9310
- **Analysis**: Logistic Regression achieved the highest performance, demonstrating robust classification capabilities due to its simplicity and effectiveness on high-dimensional data.

Decision Tree

- **Training**: A Decision Tree model with the gini criterion was trained without a maximum depth constraint.
- **Performance**:
    - Accuracy: 0.8482
    - Precision: 0.8578
    - Recall: 0.8177
    - F1-Score: 0.8373
- **Analysis**: The Decision Tree underperformed, likely due to overfitting on the high-dimensional semantic vectors, leading to reduced generalization on the validation set.

Random Forest

- **Training**: A Random Forest model with 100 estimators and the gini criterion was trained, utilizing all available cores for efficiency.
- **Performance**:
    - Accuracy: 0.9262
    - Precision: 0.9288
    - Recall: 0.9156
    - F1-Score: 0.9222
- **Analysis**: Random Forest performed strongly but was slightly outperformed by Logistic Regression, possibly due to higher computational complexity without proportional gains on the balanced dataset.

Model Comparison

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Logistic Regression	0.9337	0.9255	0.9365	0.9310
Decision Tree	0.8482	0.8578	0.8177	0.8373
Random Forest	0.9262	0.9288	0.9156	0.9222

Logistic Regression was selected as the best model due to its superior F1-score, indicating a strong balance between precision and recall, essential for reliable fake news detection.

Conclusion

The semantic classification approach using Word2Vec embeddings proved highly effective for fake news detection, capturing contextual and semantic patterns that distinguished true from fake news. The Logistic Regression model's high F1-score of 0.9310 highlights its suitability for this task, offering a scalable and accurate solution for real-world applications like social media moderation and journalism verification. The model's ability to reduce human bias and curb misinformation demonstrates its potential impact. However, limitations such as the Decision Tree's overfitting, Random Forest's computational cost, and Word2Vec's limited contextual depth suggest areas for improvement, including hyperparameter tuning or exploring advanced embeddings like BERT.

Pros

- **High Accuracy**: Logistic Regression's F1-score of 0.9310 ensures reliable classification.
- **Semantic Understanding**: Word2Vec captured contextual relationships, enhancing model performance.
- **Scalability**: The approach can handle large text volumes, suitable for real-time applications.
- **Balanced Metrics**: The F1-score focus minimized both false positives and false negatives.

Cons

- **Overfitting Risk**: Decision Tree's lower F1-score (0.8373) indicates overfitting on semantic vectors.
- **Computational Cost**: Random Forest required more resources without significant performance gains.
- **Limited Context**: Word2Vec may miss long-range dependencies compared to advanced models like BERT.

- **Data Dependency**: Performance relies on training data quality, which may not capture evolving fake news patterns.

Recommendations

- **Hyperparameter Tuning**: Optimize model parameters to improve generalization.
- **Advanced Embeddings**: Explore BERT or other transformer-based models for deeper contextual understanding.
- **Class Imbalance Handling**: Address potential imbalances in future datasets to enhance robustness.
- **Real-Time Integration**: Deploy the model in real-world systems with continuous retraining to adapt to new misinformation trends.

This report demonstrates the effectiveness of semantic classification for fake news detection and provides a foundation for further advancements in combating misinformation.