Fake News Detection using Statistical Machine Learning Models

Pranab Kumar Paul

November 2023 - December 2023

Abstract

In this project, we developed and implemented machine learning models, including BERT, Random Forest, and SVM, to classify fake news. The models achieved a detection accuracy of up to 97% and an F1-score of 98.4%. Data preprocessing techniques such as text cleaning and tokenization were employed to enhance model performance.

1 Introduction

Fake news has become a pervasive issue in the digital age, leading to misinformation and public distrust. Detecting fake news accurately is crucial for maintaining the integrity of information dissemination. This project aimed to develop robust machine learning models that can classify news articles as either fake or genuine with high accuracy.

2 Methodology

1. **Data Collection:** A dataset containing labeled news articles was used for training and evaluation.
2. **Data Preprocessing:**
   - *Text Cleaning:* Removed stop words, punctuation, and special characters to reduce noise in the data.
   - *Tokenization:* Converted text into tokens for further analysis.
3. **Model Implementation:**
   - *BERT:* Leveraged Bidirectional Encoder Representations from Transformers (BERT) for contextual understanding of text.
   - *Random Forest:* Implemented an ensemble learning method for classification.
   - *SVM:* Employed Support Vector Machine (SVM) for high-dimensional data classification.
4. **Training and Evaluation:** Models were trained on the preprocessed data and evaluated using accuracy and F1-score metrics.

3 Results

The BERT model achieved the highest performance with an accuracy of 97% and an F1-score of 98.4%. The Random Forest and SVM models also performed well, demonstrating

the effectiveness of statistical machine learning in detecting fake news. The preprocessing steps significantly contributed to the models' performance by improving the quality of the input data.

4 Conclusion

This project successfully demonstrated the application of machine learning models in detecting fake news with high accuracy. The use of BERT, Random Forest, and SVM models, combined with thorough data preprocessing, resulted in a robust detection system. Future work could involve exploring deep learning techniques and larger datasets to further enhance detection capabilities.