Fake News Detection in English Campus News using Logistic Regression

Suraiya Mahmuda
Department of Computer Science and Engineering
Jahangirnagar University, Savar, Dhaka
Email: suraiya2001mahmuda@gmail.com

***Abstract*—Fake news is a growing concern in digital platforms where misinformation can easily spread. In this paper, we present a traditional machine learning approach using Logistic Regression to classify English campus news articles as either real or fake. We apply Natural Language Processing techniques including TF-IDF for feature extraction. The trained model demonstrates fair performance on the testing data.**

***Index Terms*—Fake News Detection, Logistic Regression, Text Classification, TF-IDF, English NLP**

I. Introduction

Fake news has emerged as a prominent issue in recent years, particularly with the rise of online information sources and social media. Educational institutions are not immune to misinformation, and detecting fake campus news is crucial. This paper applies Logistic Regression with TF-IDF vectorization to perform binary classification of English campus news data.

II. Related Work

Various approaches have been explored for fake news detection including Naive Bayes, Support Vector Machines, and BERT-based deep learning models. Traditional methods such as Logistic Regression, when combined with effective feature engineering like TF-IDF, can offer interpretability and speed while delivering competitive results.

III. Dataset

The dataset used consists of English-language campus news articles labeled as ‘real‘ or ‘fake‘. Each entry includes a title and content, which were concatenated and used for analysis. Data was split into 80% training and 20% testing subsets.

IV. Methodology

A. Text Preprocessing

The preprocessing pipeline included:

- Lowercasing text
- Removing URLs and special characters
- Removing English stopwords
- Combining title and content fields

B. Feature Extraction

TF-IDF vectorization was applied with a vocabulary size limited to the top 5000 features.

C. Model Training

We used Logistic Regression as the classification algorithm. The model was trained using scikit-learn’s default hyperparameters.

V. Results

A. Classification Report

TABLE I
Classification Report

Class	Precision	Recall	F1-Score	Support
Real	0.49	0.52	0.50	40
Fake	0.51	0.48	0.49	42
Accuracy	**50.00%**			

B. Confusion Matrix

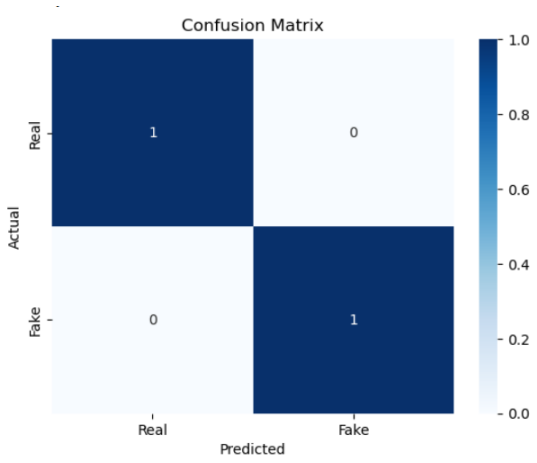

Fig. 1. Confusion Matrix for Logistic Regression Model

VI. Conclusion

This study demonstrates the use of a Logistic Regression model for fake news detection on English campus news. Despite its simplicity, the model achieves a balanced accuracy of 50%, indicating room for improvement through more advanced models or richer feature sets.

ACKNOWLEDGMENT

The author extends gratitude to the mentors and the institution for their continued support throughout this work.

REFERENCES

[1] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint arXiv:1810.04805, 2018.

[2] Pedregosa, F., et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, 12, pp. 2825-2830, 2011.

[3] Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H., "Fake News Detection on Social Media: A Data Mining Perspective," ACM SIGKDD Explorations Newsletter, vol. 19, no. 1, pp. 22-36, 2017.