

Explainable Machine Learning-Based Fake News Detection Using NLP Techniques

N.A.R. Nawodi, U.S. Wijegunasinghe, W.K. Lakshan, and P. Vithuckshiha

Department of ICT, Faculty of Technological Studies, University of Vavuniya, Sri Lanka

Abstract

The rapid spread of misinformation on social media poses significant societal risks. This study develops an interpretable and accurate fake-news detection system that combines traditional supervised learning with Explainable AI (XAI). Two large, publicly available datasets were merged into a single corpus and preprocessed (cleaning, tokenization, stop-word removal, and lemmatization). Features were extracted with TF-IDF and n-gram vectorization (uni/bi-grams). We trained Support Vector Machine (SVM), Logistic Regression, and a Passive-Aggressive classifier, and evaluated them on a held-out test set. The SVM achieved the highest accuracy (93.5%), outperforming Logistic Regression (91.2%) and the Passive-Aggressive classifier (90.5%); precision, recall, and F1-score similarly supported this ranking. Confusion-matrix analysis indicated balanced detection across fake and real classes. To improve transparency, we applied SHAP and LIME to highlight influential tokens and n-grams for both global trends and individual predictions, enabling users to verify the cues driving classifications. These results show that pairing efficient linear text models with post-hoc explanations can deliver competitive accuracy while mitigating black-box concerns. Limitations include vulnerability to evolving adversarial writing styles, limited multilingual coverage, and the need for real-time inference. Future work will explore continual/online learning, cross-lingual transfer, and streaming deployment to enhance robustness and broader applicability.

Keywords: Fake News Detection, Machine Learning, Explainable AI, Natural Language Processing, Model Interpretability