INFORMATICS INSTITUTE OF TECHNOLOGY

INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with

UNIVERSITY OF WESTMINSTER

Explainable Disinformation Detection for News Article

A Dissertation by

Mr. Udith Liyanage

Supervised by

Mr. Vaseekaran Varatharajah

Submitted in partial fulfilment of the requirements for the BEng in Software Engineering degree at the University of Westminster.

April 2025

Abstract

In the digital age, the rapid proliferation of fake news presents a significant challenge to both media outlets and the general public. Disinformation refers to false information disseminated to manipulate public opinion, with its negative impact on society evident in various areas, such as shaping political narratives and influencing economic markets. Recognizing and stopping the dissemination of false information is essential for preserving trust and reliability in news outlets. This initiative tackles the requirement for a stronger and more efficient system for detecting fake news, emphasizing real-time evaluation and understandable comparisons of news articles.

To solve this problem, it is suggested to investigate a novel approach for disinformation detection with an explainable factor that assists end-users in identifying news credibility. The proposed system processes the article by building an input article to mean pooled embedding using the BERT transformer model. A Multilayer perceptron model used for fake news detection. For the news classification explanation, input embedding words are masked and converted to a weighted pool embedding and calculates the misleading degree of classification for each word. Using BERT weighted and mean pool embedding techniques shows better results than in previous work.

The Proposed system is trained and evaluated on the ISOT fake news dataset, with most articles focusing on political and world news topics. To assess the accuracy, precision, recall and F1- score were used. After hyperparameter tuning, an accuracy of 98.71%, precision of 98.73%, recall of 98.70%, and an F1 score of 98.71% were achieved.

Subject Descriptors:

- Information systems -> Information retrieval -> Retrieval tasks and goals -> Clustering and classification
- Computing methodologies -> Machine learning -> Machine learning approaches -> Neural networks

Keywords: Disinformation Detection, Explainable, Neural Network, Embedding Techniques, Fake News