



Leveraging Natural Language Processing for Real-Time Detection of Fake News in Whatsapp Messages

Hammed Mudasiru & Jiboku Folahan Joseph*

Department of Computer Science, The Federal Polytechnic, Ilaro, Nigeria.

Corresponding author email: folahan.jiboku@federalpolyilaro.edu.ng

Introduction

The rampant spread of false and misleading communication on messaging platforms like WhatsApp has raised significant concerns about public misinformation, which can lead to confusion, fear, and even social unrest. Traditional detection methods often struggle to interpret the subtleties of natural language and operate effectively in real time. This study introduces an innovative system that leverages Natural Language Processing (NLP) to detect and mitigate the dissemination of disinformation within WhatsApp. The approach employs machine learning models trained to identify key characteristics of misinformation, such as exaggerated language, lack of credible references, and patterns consistent with known fraudulent content. The system flags questionable messages by evaluating both content and context and provides users with real-time alerts or fact-checking prompts. A feedback mechanism allows users to indicate messages that require closer scrutiny, enhancing the model's adaptability and accuracy over time. The overall objective is to demonstrate that NLP-driven techniques can significantly improve safe information sharing and foster a more trustworthy digital communication environment.

Methodology

The methodology adopts a framework incorporating NLP to detect and control the spread of disinformation on WhatsApp. A diverse dataset is collected, consisting of various message types and user behaviours. Key user-level features, such as account credibility and engagement patterns, are analyzed alongside message content. Preprocessing steps including text cleaning, tokenization, sentiment analysis, and normalization ensure that data is properly structured for model training. Feature extraction techniques are used to convert textual and behavioural data into meaningful representations for classification.

Results and discussion

The system features a user-friendly interface with secure login and real-time WhatsApp API integration. Messages go through capture, preprocessing, analysis, and validation. The homepage includes tools like a message input field, a send button, and an alert area to provide users with instant feedback on message credibility. Secure login ensures data protection through authentication and session control. One of the system's key innovations is its seamless integration with the WhatsApp API, which enables it to intercept and validate messages in real time before transmission. During preprocessing, irrelevant elements such as URLs or special characters are removed, and the message is normalized for NLP analysis. The core NLP module utilizes algorithms for text classification, sentiment analysis, and named entity recognition to determine message authenticity. Based on this evaluation, users receive alerts containing both the outcome and reasoning behind the classification. If the message is verified, it is forwarded to the recipient through WhatsApp.

Conclusion

In conclusion, this study illustrates the effectiveness of NLP in reducing the spread of fake messages on WhatsApp by enhancing content credibility, user awareness, and platform trust. Despite challenges such as false positives and evolving tactics of disinformation, the findings affirm the potential of intelligent, adaptable systems in safeguarding digital communication. Future work should prioritize continuous model updates, real-time fact-checking, and stronger NLP infrastructure to sustain and expand these benefits.

Keywords: Misinformation Detection, Natural Language Processing (NLP), Machine Learning, WhatsApp Security, Fact-Checking System.