International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Conference Proceedings: Volume: 05 ACME – 2026 DOI: 10.55041/ISJEM.ACME041
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

A Novel Explainable AI Framework for Detecting Misinformation on Federated Social Media Data

P.Lokanadham[1] , Punuru Kavyashree[2] , S.Khuranun Mubeen[3] , N.Venu Madhav[4], K.Deepak Raj[5]

[1]Assistant Professor, Dept of Information Technology, SV College of Engineering, Tirupati, India

[2]B. Tech, Dept of Information Technology, SV college of Engineering, Tirupati, India.

[3]B. Tech, Dept of Information Technology, SV college of Engineering, Tirupati, India.

[4]B. Tech, Dept of Information Technology, SV college of Engineering, Tirupati, India.

[5]B. Tech, Dept of Information Technology, SV college of Engineering, Tirupati, India.

Abstract

Web Information Processing (W.I.P.) has enormously impacted modern society since a huge percentage of the population relies on the internet to acquire information. Social Media platforms provide a channel for disseminating information and a breeding ground for spreading misinformation, creating confusion and fear among the population. One of the techniques for the detection of misinformation is machine learning-based models. However, due to the availability of multiple social media platforms, developing and training AI-based models has become a tedious job. Despite multiple efforts to develop machine learning-based methods for identifying misinformation, more work must be done on developing an explainable generalized detector capable of robust detection and generating explanations beyond black-box outcomes. Knowing the reasoning behind the outcomes is essential to make the detector trustworthy. Hence employing explainable A.I. techniques is of utmost importance.

In this work, the integration of two machine learning approaches, namely domain adaptation and explainable A.I., is proposed to address these two issues of generalized detection and explainability. Firstly, the Domain Adversarial Neural Network (DANN) develops a generalized misinformation detector across multiple social media platforms. DANN generates the classification results for test domains with relevant but unseen data. The DANN-based traditional black-box model cannot justify and explain its outcome, i.e., the labels for the target domain. Hence a Local Interpretable Model-Agnostic Explanations (LIME) explainable A.I. model is applied to explain the outcome of the DANN model.

To demonstrate these two approaches and their integration for effective explainable generalized detection, COVID-19 misinformation is considered as a case study. We experimented with two datasets and compared results with and without DANN implementation. It is observed that using DANN significantly improves the F1 score of classification and increases the accuracy by 3% and AUC by 9%. The results show that the proposed framework outperforms well in the case of domain shift and can learn domain-invariant features while explaining the target labels with LIME implementation. This can enable trustworthy information processing and extraction to combat misinformation effectively.

Keywords: Covid19, DANN, LIME, Misinformation Detection, Social Media, Text Processing, Web Information Processing, XAI.

