Trust Oriented Explainable AI for Fake News Detection

1st Krzysztof Siwek
Warsaw University of Technology
Warsaw, Poland
ORCID: 0000-0003-2642-2319
krzysztof.siwek@pw.edu.pl

2nd Daniel Stankowski
Student of
Warsaw University of Technology
Warsaw, Poland

3rd Maciej Stodolski
Warsaw University of Technology
Warsaw, Poland
ORCID: 0009-0003-0644-8105
maciej.stodolski@pw.edu.pl

***Abstract*—This article examines the application of Explainable Artificial Intelligence (XAI) in NLP based fake news detection and compares three widely used interpretability methods, SHAP, LIME, and Integrated Gradients. Two neural classifiers, a recurrent LSTM model and a convolutional model, were trained on the ISOT Fake News Dataset and interpreted with each method. Explanation quality was assessed with four faithfulness measures, completeness, sufficiency, area over the perturbation curve, and the number of tokens needed to flip a decision, together with runtime. The results show that the most suitable explanation method depends on the architecture, SHAP gives the most faithful explanations for the recurrent model, while Integrated Gradients is the most faithful and by far the fastest choice for the convolutional model. The study also discusses practical limitations, including computational cost, sensitivity to parameterization, and the out of distribution effect of token masking used during evaluation. Overall, the findings indicate that integrating XAI with NLP based fake news detection is a practical way to improve the transparency and trustworthiness of automated moderation systems.**

***Index Terms*—Fake News Detection, Explainable Artificial Intelligence, Natural Language Processing, SHAP, LIME, Integrated Gradients**

I. Introduction

The spread of digital misinformation has pushed research toward high capacity deep learning models for automated veracity assessment. Architectures such as Transformers and graph neural networks reach strong empirical accuracy, yet they behave as black boxes, which is a real obstacle in sensitive domains such as public health, elections, and finance [1]. A prediction without a justification is hard to audit and does not give human fact checkers the evidence they need. Explainable Artificial Intelligence has emerged as a corrective, shifting attention from raw predictive performance toward human understandable reasons for a decision.

This work compares three post hoc XAI methods, SHAP, LIME, and Integrated Gradients, on a complete fake news detection pipeline that covers data preparation, classifier training, and explanation, and it evaluates them on two architectures with contrasting inductive properties, a recurrent network and a convolutional network. The experiments confirm that XAI improves transparency while the classifiers retain strong performance. Although the quantitative differences between methods within a single model are moderate, each technique has a distinct profile, SHAP gives detailed and faithful local attributions, LIME gives accessible surrogate explanations, and Integrated Gradients is efficient and works particularly well with the convolutional model. The study also surfaces practical limitations, including computational overhead, sensitivity to parameterization, and the risk of misinterpretation by end users, which together argue for a comparative, model specific evaluation rather than a single universal ranking.

II. Related Work

Research on explainable fake news classification splits broadly into model agnostic post hoc interpretation and intrinsically interpretable architectures. The dominant paradigm interprets pretrained classifiers after the fact. LIME and SHAP remain the foundational tools for token level saliency, since they quantify how individual words shift a model toward one label or another [2], [3]. SHAP draws on cooperative game theory and assigns each feature a Shapley value relative to a baseline, which gives a fair and mathematically grounded attribution [4]. These methods are effective locally, but they give feature attribution rather than a causal narrative, so they often miss the linguistic nuance of deceptive content.

To move past this limitation, intrinsically interpretable models such as the Tsetlin Machine express a decision as explicit logical clauses rather than opaque weights [5], and hybrid architectures combine FastText style representations with interpretable modules to balance accuracy and transparency [6]. A further shift moves from feature attribution to generative justification, where large language models synthesize natural language rationales that compare a claim with evidence [7], turning XAI into a communicative tool for end users at the cost of a risk of hallucinated reasoning. Detection has also moved beyond isolated text, with systems that use social propagation patterns and source or account reputation as part of the explanation [8], [9]. A persistent open problem is that features a model finds important do not always match what a person would consider relevant [10]. Our contribution is a controlled, model specific comparison of three established post hoc methods on one shared pipeline, which isolates the effect of classifier architecture on explanation faithfulness.