

## Explainable Artificial Intelligence Approaches in NLP-Based Text Classification: A Systematic Literature Review

Gunasekara S.A.G.K.<sup>1\*</sup>, Kaushalya P.K.D.K.<sup>1</sup>

<sup>1</sup>Department of Computing and Information Systems, Faculty of Computing,  
Sabaragamuwa University of Sri Lanka, Sri Lanka

\*sagkgunasekara@std.appsc.sab.ac.lk

Artificial Intelligence plays a crucial role in this digital era by enabling the processing of large amount of textual information. However, tools built based on large language models, such as ChatGPT, can produce incorrect results when certain conditions are present, including factual hallucination, mismatched domains, or a lack of contextual grounding. Despite their impressive accuracy, Natural Language Processing models built using Deep Learning models lack transparency and interpretability because they usually operate as “black boxes” without providing explicit reasoning. Explainable Artificial Intelligence addresses these challenges by providing human understandable explanations for model predictions. This Systematic Literature Review examines peer-reviewed articles published between 2020–2025 and retrieved from major academic databases, chosen according to the predefined inclusion and exclusion criteria under the PRISMA framework. A narrative synthesis method is used to examine and classify XAI methods used in NLP-based text classification. The findings show that the most popular XAI techniques are LIME, SHAP, Attention Visualization, and Natural Language Explanations. Model-agnostic approaches such as LIME and SHAP are suitable for general interpretability and regulatory analysis, whereas attention-based and natural language explanation methods are more suitable to transformer-based models and user-focused applications. Furthermore, the recent studies point to the significance of Human-Centered and domain-specific explainability, especially in the context of healthcare, misinformation recognition, and emotion recognition. The review recognizes that the lack of standardized evaluation measures, limited scalability to large language models, and that there was a weak user centered validation as the main research gaps and offers suggestions on future research.

**Keywords:** *Explainable AI (XAI); Natural Language Processing (NLP); Text Classification; Transparency; Human-Centered AI*