

ABSTRACT

The rapid growth of social media platforms such as X has generated a large volume of public opinion related to mangrove ecosystem rehabilitation in Indonesia, requiring analytical approaches that are both effective and transparent. Topic modeling using BERTopic is well-suited for short-text analysis; however, its black-box nature limits the interpretability of results. This study aims to optimize the objective function to balance topic coherence and diversity, while also implementing Explainable Artificial Intelligence (XAI) to enhance model transparency. A total of 59,460 Indonesian-language tweets were modeled using IndoBERTweet embeddings, with three multi-objective scalarization schemes explored: weighted sum, geometric mean, and weighted sum with penalti. For interpretation, a Logistic Regression-based surrogate model was employed, analyzed globally using SHAP and locally using LIME, and further validated through labeling by a Large Language Model (LLM) and human judgment. The results show that the penalti-based scenario ($\alpha = 0.7$) achieves the best algorithmic performance, with a surrogate accuracy of 0.85 and a controlled outlier proportion of 7.34%, while effectively preserving the integrity of ecological discourse from noise. However, human evaluation indicates a preference for the geometric mean scenario, as it produces sharper and more interpretable clusters despite exhibiting over-filtering behavior. These findings highlight a gap between mathematical performance and human interpretability, and demonstrate the important role of XAI in bridging this trade-off in public discourse analysis.

Keywords : Topic Modeling, BERTopic, Multi-objective Evaluation, Explainable AI, Mangrove