Abstract

The spread of hate speech on social media, particularly on platform X (formerly Twitter), poses serious risks such as social polarization, conflict, and psychological trauma. The complexity of the Indonesian language, ambiguous meanings, and multi-label data structures present major challenges in building an effective automatic detection system. This study proposes a multi-label classification approach combining Elman Recurrent Neural Network (ERNN) and Dolphin Echolocation Algorithm (DEA) to detect Indonesian hate speech. DEA is employed to optimize ERNN hyperparameters, while text representation is enhanced through the integration of Word2Vec and TF-IDF features. Four experimental scenarios were designed to evaluate the impact of both optimization and feature representation quality on model performance, using F1-score as the primary metric. Results show that combining TF-IDF and Word2Vec features yields a more substantial performance improvement than hyperparameter optimization alone. The best-performing scenario—DEA-ERNN with combined features—achieved a macro F1-score of 60.17%. These findings highlight the dominant role of feature representation quality in the success of multi-label hate speech classification and support the development of more adaptive and efficient content moderation systems for the Indonesian language context.

Keywords: hate speech, DEA, ERNN, multi-label classification, Word2Vec, TF-IDF