ABSTRACT

Currently, especially on social media, hate speech almost always involves offensive language. Twitter (X) is a social media platform like others, but what sets it apart is that posts are called tweets, and there is a retweet feature to share them with your followers. Despite its popularity, Twitter (X) is increasingly used to spread hatred and misinformation due to its viral nature and anonymity. In one sentence, hate speech can have several labels that refer to several topics.This study explore the effectiveness of various BERT-based models, including BERT, BERT-CNN, BERT-LSTM, and BERT-BiLSTM, for multi-label hate speech detection across different text lengths. The results reveal that model performance varies with sentence length. For longer texts, the BERT-BiLSTM model achieved the highest accuracy of 83.20%, along with superior recall and F1 scores, demonstrating its ability to capture complex and nuanced context. BERT-CNN also performed well on long texts, showing good accuracy and precision, albeit with a slightly lower F1 score than BERT-BiLSTM. On the other hand, BERT and BERT-LSTM provided moderate results but were less effective in managing detailed context in extended passages. For short texts, BERT-CNN excelled, achieving the highest accuracy (79.8%) and F1 score (79.10%), indicating the efficacy of convolutional layers in extracting key features from brief content. BERT-LSTM also demonstrated balanced precision and recall, while BERT-BiLSTM showed strong recall but slightly lower accuracy, suggesting its strength lies in processing richer contextual information. These findings highlight the importance of aligning model architecture with text characteristics: BERT-BiLSTM is optimal for deep contextual understanding in longer texts, while BERT-CNN effectively identifies critical features in shorter, concise samples.

Keywords:Deep Learning, Classification Task, Hate Speech Detection, Sentence Length, Social Media.