|  |  |  |  |
| --- | --- | --- | --- |
| Title | L-boost: Identifying offensive texts from social media post in bengali | | |
| Author(s) Name | Muhammad F Mridha, Md Anwar Hussen Wadud, Md Abdul Hamid, Muhammad Mostafa Monowar, Mohammad Abdullah-Al-Wadud, Atif Alamri | | |
| Contact Email(s) | firoz.mridha@aiub.edu | | |
| Published Journal Name | IEEE Access | | |
| Type of Publication | Journal | | |
| Volume | 9 | Issue |  |
| Publisher | IEEE | | |
| Publication Date | 2021/12/8 | | |
| ISSN | 2169-3536 | | |
| DOI | 10.1109/ACCESS.2021.3134154 | | |
| URL |  | | |
| Other Related Info. | https://ieeexplore.ieee.org/abstract/document/9642973 | | |
|  | | | |

|  |  |
| --- | --- |
| Abstract |  |
| Due to the significant increase in Internet activity since the COVID-19 epidemic, many informal, unstructured, offensive, and even misspelled textual content has been used for online communication through various social media. The Bengali and Banglish(Bengali words written in English format) offensive texts have recently been widely used to harass and criticize people on various social media. Our deep excavation reveals that limited work has been done to identify offensive Bengali texts. In this study, we have engineered a detection mechanism using natural language processing to identify Bengali and Banglish offensive messages in social media that could abuse other people. First, different classifiers have been employed to classify the offensive text as baseline classifiers from real-life datasets. Then, we applied boosting algorithms based on baseline classifiers. AdaBoost is the most effective ensemble method called adaptive boosting, which enhances the outcomes of the classifiers. The long short-term memory (LSTM) model is used to eliminate long-term dependency problems when classifying text, but overfitting problems occur. AdaBoost has strong forecasting ability and overfitting problem does not occur easily. By considering these two powerful and diverse models, we propose L-Boost, the modified AdaBoost algorithm using bidirectional encoder representations from transformers (BERT) with LSTM models. We tested the L-Boost model on three separate datasets, including the BERT pre-trained word-embedding vector model. We find our proposed L-Boost’s efficacy better than all the baseline classification algorithms reaching an accuracy of 95.11%. | |