ABSTRACT

Depression is a prevalent health condition that impacts individuals worldwide. Its defining characteristics include persistent feelings of sadness, hopelessness, and disinterest in previously enjoyable activities. On the other hand, Depression detection is the task of determining whether a human being is depressed or non-depressed by using some features or indicators. Linguistic indicator or detecting depression from text writing are used in this study. The text was taken from the X (Twitter) tweet data. Considering tweets are frequent use of abbreviations and slang, and lack of grammatical correctness, tweets often present a particular challenge. The utilization of particular slang phrases in the context of social media posts, specifically on the platform Twitter, has the potential to result in vocabulary mismatches. Machine learning frequently encounters vocabulary mismatches in capturing the meaning of the user's writing (in this case, it is used to detect depression) and produces less than-optimal accuracy. The implementation of feature enrichment, utilizing user-based, content-based, and LIWC features, along with feature extraction and expansion, will result in improved accuracy in depression detection systems using tweet data on X. The way feature enrichment works in classification is to combine User-based, content-based, and LIWC feature. The proposed methods have enhanced the performance of the accuracy on several experiments. Baseline combining with LIWC feature can improve accuracy by 25% and with content-based feature can improve accuracy by 18.75%. On the other hand, user-based feature cannot improve the system from the baseline. The best performance from several scenario tests was obtained with combination baseline (extracting tweet data using TFIDF) + LIWC feature. The combination LIWC feature can provide a performance of 93.75% the accuracy.

Keywords: depression, depression detection, vocabulary mismatches, feature enrichment, LIWC