ABSTRACT

Twitter, a social media platform with 586 million active users, has become a public space where information can spread rapidly. However, the growth of fake accounts, especially social spambots on Twitter that can imitate real user behavior, creates serious challenges in maintaining interaction quality and information credibility. Previous studies have shown that single-modality approaches, either based on account metadata or tweet text, can achieve high classification performance. Nevertheless, relying on only one type of information is not sufficient to fully represent the characteristics of increasingly sophisticated fake accounts. This study proposes a multimodal deep learning model that combines account metadata and tweet textual content using a late fusion approach to classify fake and real users on X. Account metadata are processed using a Multi-Layer Perceptron (MLP), while tweet text is processed using a Long Short-Term Memory (LSTM) with GloVe Twitter word embeddings. This study uses Cresci-2017 dataset and class imbalance is handled through random undersampling. The results showed that in the metadata-based approach, XGBoost achieved an F1-score of 0,9950 and MLP achieved an F1-score of 0,9840, while in the text-based approach, SVM achieved an F1-score 0,9210 and LSTM reached an F1-score of 0,9392. The multimodal approach, which combines both modalities through late fusion, produced the best performance with an accuracy of 0,9920 and a weighted F1-score of 0,9921, for both multimodal deep learning and machine learning models. The findings indicate that improvement of evaluation results in multimodal modeling over unimodal modeling shows that the model is able to provide better representation by combining information from tweet text and account metadata, and able to maintain a balance of performance between two user classes.

Keywords: Fake Accounts, Social Spambots, Social Media X, Multimodal Deep Learning, LSTM, MLP, Late Fusion, Cresci-2017