Bot Detection on Twitter Using Machine Learning

Hamza Rashid
McGill University
Research Class

***Abstract*—This paper investigates bot detection on Twitter using machine learning techniques. We analyze tweet content and user metadata to identify patterns that distinguish bots from humans. Our approach involves topic categorization, feature extraction, and classification models to assess predictive performance. The findings provide insights into automated behavior and its implications for social media integrity.**

I. Introduction

The prevalence of bot accounts on social media platforms like X (formerly Twitter) is at the center of bot detection literature. Such research targets a particular class of bots, namely, ones deployed for malicious purposes, such as inciting arguments or spreading misinformation. We present and compare various methods for this task, including rule-based heuristics, feature engineering, and deep learning, while addressing the limitations of supervised techniques tested on static datasets.

II. Methodology

Our dataset consists of Twitter posts in JSON format, containing tweet text, timestamps, user metadata, and other features. Our pipeline includes data preprocessing, feature extraction, unsupervised topic categorization, and classification model training.

Given the small datasets and unbalanced classes (majority non-bots), initial methods struggled with low bot recall. Combining TF-IDF with LDA improved recall on a single batch but struggled with noisy bots.

When we expanded our dataset to four sessions, the accuracy of TF-IDF+LDA (treating all tweets from a user as a single document) dropped from the mid-90s to the mid-70s. Adding embeddings from all-MiniLM-v6 (on the concatenated tweets) improved accuracy to 89.4

To generalize better, we used individual tweets per user, extracting mean embeddings, topic distributions with BERTopic+HDBSCAN, and temporal features (mean time of posts and standard deviation). This latest approach achieved 97% accuracy on the expanded dataset.

III. Discussion

HDBSCAN, a density-based clustering method, is beneficial as it identifies tweet clusters without a predefined number, making it robust to noise and adaptable to varying data distributions.

IV. Conclusion and Future Work

Our findings highlight the importance of embeddings and dynamic topic modeling. Future work includes exploring faster embeddings, preventing data leakage, and refining HDBSCAN for evolving datasets.