BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2026
http://excel.fit.vutbr.cz

Detection of Fake Accounts on Social Networks: Poster Commentary

Patrik Žáček*

Abstract
Social media bots generate over half of global internet traffic, but modern Large Language Models (LLMs) have made traditional syntax-based detectors obsolete. This work addresses the detection of advanced, LLM-assisted social bots on Twitter/X using a hybrid dual-modal stacking ensemble that fuses structural profile metadata with semantic text embeddings. Evaluated on the massive TwiBot-22 benchmark, our system achieves a 54.86% F1-Score, outperforming feature-based baselines and rivaling state-of-the-art graph neural networks. By circumventing the graph-data bottleneck, this solution enables robust, real-time inference on isolated accounts in the wild.

*xzacek@stud.fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

[Motivation] Social media bots are no longer just simple scripts blasting identical spam links. As of 2023, automated agents generate over half of global internet traffic. Driven by Large Language Models, these bots now manufacture fake consensus, astroturf political elections, and manipulate financial markets by writing grammatically perfect, contextually aware text. Classical syntax-based detectors are entirely blind to them.

[Problem definition] The objective of this work is to accurately detect advanced, LLM-assisted social bots on the Twitter/X platform. Crucially, the solution must operate without relying on complex ego-network data. Due to strict modern API rate limits, gathering graph data is impractically slow, meaning a viable real-world detector must be able to classify an isolated account in real-time.

[Existing solutions] As seen in our results, traditional feature-based tools like Botometer are failing in the modern landscape. Conversely, state-of-the-art Graph Neural Networks (like BotRGCN) achieve high accuracy but require the extraction of an account's entire social graph, which is highly restricted. Text-only transformer baselines (RoBERTa, T5) collapse entirely when faced with the grammatical perfection of LLM bots.

[Our solution] We propose a Dual-Modal Stacking Ensemble architecture. Instead of relying on a single view, we fuse semantic text embeddings (processed by a frozen RoBERTa model) with structural profile metadata (processed by a Random Forest). A logistic regression meta-classifier then weighs these two independent viewpoints to make a final prediction.

[Contributions] Our system rivals heavy graph-based models while requiring zero graph data at inference time. Furthermore, we demonstrate that temporal features (tweet inter-arrival times) are now obsolete evasion vectors, and we identify new real-world evasion tactics, such as the farming of "structural shields" by malicious accounts.

2. The Dual-Modal Architecture

If you look at the taxonomy in Figure 1, you will see why this problem is so difficult. We are no longer just fighting "Spam Bots." We are dealing with Cyborgs (human-bot hybrids) and LLM-Driven Bots. Neither text alone nor metadata alone is sufficient to catch these sophisticated actors.

This brings us to the core of the methodology, illustrated in Figure 2. The system consists of two distinct branches that operate in parallel:

- **The Metadata Branch:** A Random Forest processes 24 structural features extracted from the user's profile (follower counts, ratios, account age).

- **The NLP Branch:** A frozen RoBERTa model generates embeddings from the user's biography and their 20 most recent tweets, which are then passed through a Multi-Layer Perceptron (MLP).

During the ablation phase, I originally included a third branch: a BiLSTM analyzing the temporal inter-arrival times of tweets. However, as noted in Key Finding 2, this branch collapsed. Modern bots randomize their execution timers so effectively that temporal data is essentially useless. The meta-classifier naturally assigned a weight near zero to the LSTM branch, so it was excluded from the final architecture for efficiency.

3. Discriminative Features and Performance

When we look at Figure 3, the Gini importance chart reveals exactly what exposes a modern bot. While they can fake text perfectly, they struggle to fake organic social growth. Bots often follow massive numbers of users hoping for follow-backs, creating an unnatural ratio that the Random Forest easily flags.

Table 1 presents our results on TwiBot-22 [1], the largest available benchmark (containing 1 million users). I want to draw your attention to the **Accuracy Paradox** present in the text-only baselines (RoBERTa and T5). They report over 72% Accuracy, but their F1-Scores sit at a dismal 20%. Because the dataset is highly imbalanced, text-only models simply learn to guess the majority class.

Our Dual-Modal Ensemble shatters this barrier. By fusing the metadata, we achieve an F1-Score of 54.86%. This vastly outperforms [2] (42.8%) and rivals the state-of-the-art BotRGCN [3] (57.5%) but with a critical advantage: our model does not need the social graph, meaning it can classify a user instantly.

If we examine Figure 4, you can see the probability distribution of our ensemble's predictions. There is a very clear bimodal separation between genuine humans (peaking near 0.1) and bots (peaking near 0.7). The interesting part is the overlap boundary between 0.3 and 0.5. This grey area represents highly active, legitimate publishers (like news organizations) and evasive "cyborgs," which explains our false positive rate.

4. Live Deployment and Evasion Tactics

Testing on static datasets is not enough. As noted in Key Finding 5, TwiBot-22 suffers from temporal decay – it predates the era of ChatGPT and purchased verification checkmarks.

To address this, we deployed the model against live accounts in April 2026. Table 2 illustrates our failure modes and discoveries. For example, *@EarthquakesSF* is technically an automated bot, but because it provides legitimate utility, humans organically follow it. It builds a massive "structural shield," causing our model to misclassify it as a human. Conversely, *@DormantUser* mimics the exact inactivity patterns of a "sleeper bot," resulting in a false positive.

These live samples prove that bots are actively evolving to game metadata classifiers by purchasing followers or hibernating, reinforcing the fact that multi-modal verification is the only path forward.

5. Conclusions

If I were to summarize the main takeaway from this poster, it is that **multi-modality beats the single-view approach**. As LLMs continue to perfectly mimic human syntax, analyzing text alone is a dead end. By combining semantic understanding with structural metadata, we can detect evasive bots without relying on heavily rate-limited graph data APIs.

If I had more time to expand this work, my immediate next step would be addressing dataset temporal decay. Social network behavior shifts every few months; moving forward, the community needs continuous data pipelines and online learning paradigms rather than static benchmark datasets.

Acknowledgements

I would like to thank my supervisor, Ing. Anton Firc, Ph.D., for his invaluable guidance, technical insights, and continuous support throughout the development of this thesis.

References

[1] Shangbin Feng, Zhaoxuan Tan, Herun Wan, and Wang. Twibot-22: Towards graph-based twitter bot detection. *arXiv preprint arXiv:2206.04564*, 2022. NeurIPS 2022 Datasets and Benchmarks track.

[2] Clayton A Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. Botornot: A system to evaluate social bots. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 273–274. International World Wide Web Conferences Steering Committee, 2016.

[3] Shangbin Feng, Herun Wan, Ningnan Wang, and Minnan Luo. Botrgcn: Twitter bot detection with relational graph convolutional networks. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 48–51. IEEE, 2021.