Detection of Fake Accounts on Social Networks

Patrik Žáček · Bachelor's Thesis, 2026 · Supervisor: **Ing. Anton Firc, Ph.D.** · FIT VUT Brno

Excel@FIT 2026

Faculty of Information Technology
Brno University of Technology

MOTIVATION & PROBLEM

Social media bots have become a dominant force shaping online discourse. Generating **over half of global internet traffic** (2023), they manufacture fake consensus, spread misinformation, and manipulate financial markets. The rise of **Large Language Models (LLMs)** has rendered classical syntax-based detectors obsolete - modern bots write grammatically perfect text indistinguishable from humans.

This work addresses the detection of **social bots** on Twitter/X using a hybrid multi-modal approach that avoids the graph-data bottleneck of state-of-the-art GNN-based systems, enabling **real-time inference on isolated accounts**.

Bot Detection · NLP · Ensemble Learning · Social Media Security · TwiBot-22 · RoBERTa · Random Forest

KEY RESULT

54.86% F1-Score	**0.418** MCC
77.55% Accuracy	**67.22%** Precision

Evaluated on **TwiBot-22** - the largest Twitter bot benchmark (1 M users). Outperforms Botometer (F1 42.8%) and all text-only baselines, rivals BotRGCN (F1 57.5%) without requiring graph data.

FIGURE 1 - TAXONOMY OF MALICIOUS SOCIAL BOTS

Spam Bots
Disseminate malicious URLs, engage in cyberbullying, trolling. High volume, low quality content.

Political Bots
State/party-operated. Amplify polarization, astroturf elections. 2016 US: ~400 k active bots.

Sybil Accounts
Multiple fake identities to infiltrate communities or manipulate reputation systems.

Stegobots
Hide C&C commands in images via steganography to coordinate botnet operations covertly.

Cyborgs
Hybrid human + automation. Hardest to detect - manual oversight mixed with scripted actions.

LLM-Driven Bots
GPT-powered agents producing diverse, grammatically perfect, contextually aware text. Hardest to detect linguistically.

FIGURE 2 - DUAL-MODAL STACKING ENSEMBLE ARCHITECTURE

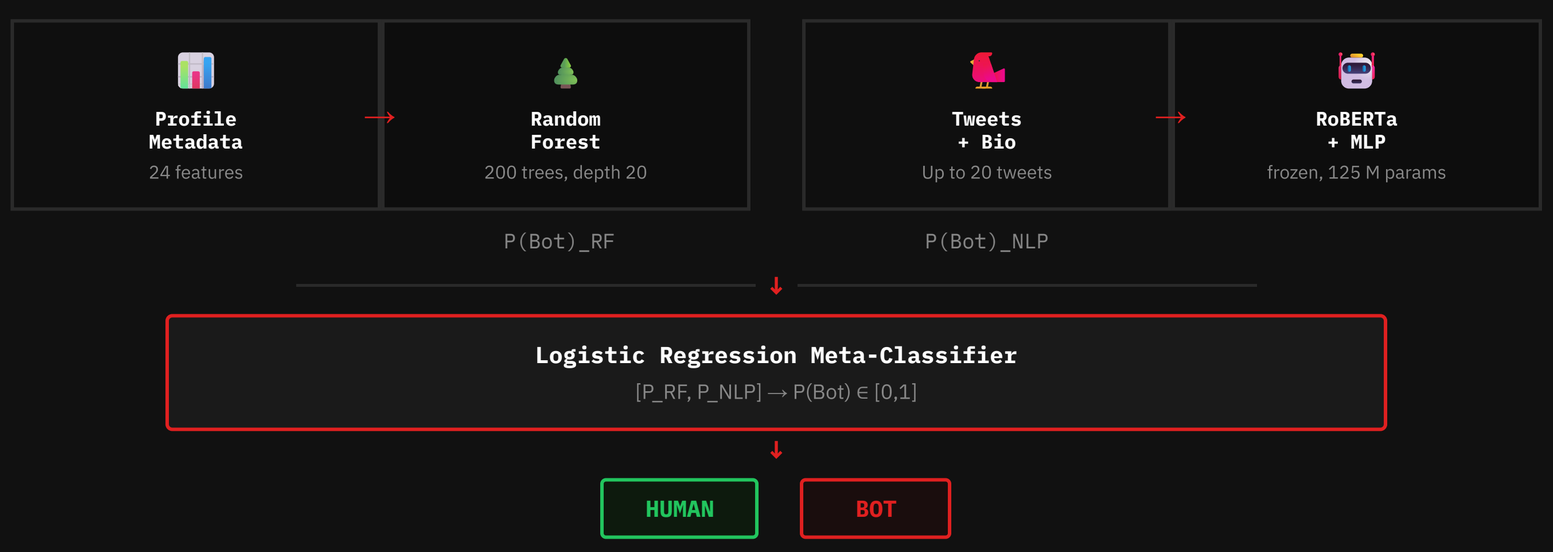

Figure 2. *Two independent branches produce probability scores that are fused by a trained linear meta-classifier. The temporal LSTM branch was excluded after ablation showed negligible contribution (LR weight ≈ 0).*

FIGURE 3 - TOP METADATA FEATURES (GINI IMPORTANCE)

Figure 3. *Follower count and activity volume are dominant discriminative features. Ratio and age signals expose artificial profiles.*

TABLE 1 - PERFORMANCE ON TWIBOT-22 BENCHMARK

Method	Paradigm	Accuracy	F1-Score
Botometer	Feature-based	49.9%	42.8%
BotHunter	Feature-based	72.8%	23.5%
SGBot	Feature-based	75.1%	36.6%
RoBERTa (baseline)	Text-based	72.1%	20.5%
T5	Text-based	72.1%	20.2%
RGT	Graph-based	76.5%	42.9%
BotRGCN	Graph-based	79.7%	57.5%
Dual-Modal Ensemble	**Hybrid (ours)**	**77.55%**	**54.86%**

Table 1. *Our system rivals BotRGCN while requiring no graph data at inference time. Text-only baselines collapse on TwiBot-22 (Accuracy Paradox: high Acc, near-zero F1).*

FIGURE 4 - ENSEMBLE PREDICTION PROBABILITY DISTRIBUTION

Figure 4. *Clear bimodal separation. Overlap in the 0.3–0.5 range corresponds to sophisticated "cyborg" accounts and legitimate high-volume publishers (false positives).*

KEY FINDINGS & CONTRIBUTIONS

1. **Multi-modality beats single-view.** Fusing semantic embeddings (RoBERTa) with structural metadata (Random Forest) resolves each model's blind spots. McNemar's test confirmed: $p < 0.0001$, $\chi^2=2984.77$.
2. **Temporal features fail on modern bots.** BiLSTM on inter-arrival times collapsed to majority-class prediction without weighting. The meta-classifier assigned near-zero weight to the LSTM branch, prompting its exclusion.
3. **Graph data is not required.** The Dual-Modal system reaches 95% of BotRGCN's F1 without ego-network data, making **real-time inference on isolated accounts** practical under current API restrictions.
4. **Live deployment reveals new evasion tactics.** Malicious bots build follower-based "structural shields" to neutralize metadata classifiers. Only multi-modal evidence catches these accounts.
5. **Dataset temporal decay.** TwiBot-22 predates LLM-bots and purchasable verification (Twitter Blue). Real-world accuracy is lower than benchmark metrics suggest - continuous data pipelines are essential.

TABLE 2 - LIVE DETECTION SAMPLES (APRIL 2026)

Account	Conf.	Pred.	Real
@BBCBreaking	0.068	HUMAN	✓
@EarthquakesSF	0.242	HUMAN	BOT*
@DormantUser	0.564	BOT	HUMAN†
@ashonicee	0.281	HUMAN	BOT‡

* Utility bot with large organic following (Structural Shield).
† Dormant human mimics sleeper bot.
‡ Evasive spam bot with purchased followers.

Table 2. *Failure modes: evasion via social capital farming and legitimate dormancy patterns.*