IDSP Proposal

Title: Botnets in the Age of GenAI: A Multimodal Ensemble Framework for Detecting Synthetic Agents on X

Author

Luc Thiery, 51903906

Supervisor:

- Main Supervisor: Dr. Vito D. P. Servedio (Lecture Series 194.046) (CSH)
- Co-Supervisor: Dr. Jürgen Cito (TU Wien)

Abstract

This project investigates the detection of GenAI-powered botnets on X (formerly Twitter) through a multimodal, ensemble-based approach. Generative AI systems now produce not only text but also images and videos that mimic authentic human communication, making automated accounts nearly indistinguishable from real users. Traditional detection models, designed for rule-based or spam-style automation, fail to capture this new level of sophistication. The project aims to develop and evaluate a stacked ensemble detection framework that integrates predictions from multiple modality-specific models—metadata, text, behavioral, and visual—each trained on complementary datasets. Rather than building a new gold-labeled dataset from scratch, model outputs will be combined using majority voting and meta-learning to form a unified detection signal. (If time allows, a small human-labeled holdout validation set (400 users) will be created for calibration and benchmarking.) The outcome will be a reproducible, multimodal benchmark for identifying GenAI-driven influence networks.

Motivation / Problem Statement

The rapid evolution of Generative AI (GenAI) has enabled the creation of synthetic online identities that can produce convincing text, realistic images, and deepfake-style videos. These technologies are increasingly exploited for manipulation, disinformation, and coordinated influence operations on social media platforms. Traditional bot detection systems—focused on metadata or simple keyword heuristics—are no longer sufficient to identify these multimodal, adaptive synthetic agents. Concrete evidence of this development has already been documented across multiple independent investigations.

The OpenAI Threat Intelligence Report revealed that at least five covert influence operations—originating from Russia, China, Iran, and Israel—used GenAI tools to automatically generate multilingual social media comments, create fake personas, and even debug malicious automation code (OpenAI, 2024). Similarly, Meta's Adversarial Threat Report uncovered the world's largest known influence operation, Spamouflage, linked to Chinese law enforcement, which has evolved to incorporate GenAI-generated content to amplify political messaging across multiple platforms (Meta Platforms, 2024). The Alan Turing Institute's CETaS report further warns that GenAI tools capable of producing text, images, and videos now pose a systemic risk to democratic discourse, especially in the context of elections and public information integrity (CETaS, 2024).

Complementing these industry findings, Choi et al. conducted a large-scale academic study analyzing over six million tweets related to the war in Ukraine. Their mixed-method network analysis revealed a hybrid ecosystem of human users, automated bots, and state-linked influencers systematically reinforcing Russian propaganda narratives through coordinated posting and language adaptation (Choi et al., 2024).

From my own experience as a regular Twitter user and through a study I conducted as part of the Critical Thinking and Social Media course at WU—where I examined how well people can tell apart real and AI-generated images in social-media news posts—I've seen firsthand how blurred the line between authentic and synthetic content has become. What used to be clearly recognizable as spam or fake now often appears completely real. These changes show that GenAI-driven bots and influence campaigns are already shaping everyday online discussions, which makes it all the more important to develop a reliable, data-driven, and multimodal framework that can detect synthetic activity across text, visuals, and behavioral patterns.

Core Research Questions

- How can multimodal signals (text, metadata, behavioral, visual) be effectively combined through ensemble learning to improve detection of GenAI-powered bots?
- Which model modalities contribute most to generalization across datasets and time?
- How reliable are ensemble predictions without human labeling, and what improvement can a small gold-labeled validation set provide?

Methodology and Process

The project follows the CRISP-DM process, emphasizing reproducibility, multimodal fusion, and ensemble learning

ID	Dataset	What it contains	Input features (examples)	Labels used
A	**BotArtist** (2022–23, ~10.9M users)	Tabular account metadata & ratios (no tweets/images)	followers, following, statuses, entropy, verified/protected etc.	**Weak** (BotArtist predictions)
B	**Fox8-23** (2023, 2,280 users)	Tweets + some profile info (gold labels)	tweet embeddings (mean/attention pooled per user), simple metadata	**Gold** (human-verified)
C	**BotSim-24** (synthetic, 2024)	Simulated interactions & behaviors	inter-tweet times, diurnal patterns, reply/retweet ratios, cascade	**Gold** (by design)
D	**Dracewicz & Sepczuk** (synthetic , 2024)	Profile screenshots	image embeddings of avatar/banner/UI	**Gold** (by design)

Pipeline Overview

1. Data Preparation – feature normalization, embeddings, outlier detection.
2. Model Training per Modality: **Model Training per Modality**
   - **Metadata:** models trained on profile and account-level features.

   - **Text:** models leveraging tweet content and linguistic representations.

   - **Behavior:** models analyzing temporal, interaction, and coordination patterns.

   - **Visual:** models using profile and image-based representations.
3. Prediction Aggregation – collect each model's output P(bot | modality) across datasets.
4. Meta-Ensemble / Majority Voting
   - Level 1: majority vote of base models
   - Level 2: stacked meta-learner (e.g. Logistic Regression / LightGBM)
5. (Optional) human-labeled holdout validation (400 accounts) for calibration.
6. Evaluation and Explainability – ROC/PR-AUC, F1, Brier score etc.
7. Reproducibility – version-controlled notebooks, fixed seeds, configuration files.

Expected Results and Evaluation Metrics

Baselines and KPIs

Model	Baseline Score (AUC / F1)	Target Score (AUC / F1)
Metadata Model	0.90 / 0.88	0.92 / 0.90
Text Model	0.93 / 0.90	0.95 / 0.92
Behavior Model	0.88 / 0.86	0.91 / 0.89
Visual Model	0.85 / 0.80	0.88 / 0.83
Meta-Ensemble	—	0.96 / 0.94

Domain-Specific Lecture

Course: S_5690 – Critical Thinking and Social Media (WU Wien)

Status: Completed – Summer Semester 2025

his course explored misinformation, persuasion, and algorithmic influence—providing the ethical and social context for analyzing GenAI-driven automation and synthetic influence networks. As part of the course, the student conducted an empirical study on the human ability to distinguish between real and AI-generated images in a social-media news context, highlighting the challenges of visual misinformation and reinforcing the motivation for multimodal detection methods in this project.

References

- OpenAI. (2024). Influence and cyber operations: An update. OpenAI Threat Intelligence Report. https://cdn.openai.com/threat-intelligence-reports/influence-and-cyber-operations-an-update_October-2024.pdf
- Meta Platforms. (2024). Quarterly adversarial threat report – Q1 2024. Meta Platforms. https://md.teyit.org/file/meta-threat-report.pdf
- Centre for Emerging Technology & Security (CETaS). (2024). AI-enabled influence operations: Safeguarding future elections. The Alan Turing Institute. https://cetas.turing.ac.uk/publications/ai-enabled-influence-operations-safeguarding-future-elections
- Choi, J., Fink, C., & Carley, K. M. (2024). Analyzing Russia's propaganda tactics on Twitter using mixed-methods network analysis and NLP. EPJ Data Science, 13(47). https://epjdatascience.springeropen.com/articles/ 024-00479-w