Identifying Bot Tweets: Political and Nonpolitical Contexts

William Johnson and Andrew Hatfield

COLGATE UNIVERSITY

Introduction

As society is exposed to more social media and bot accounts become prevalent, the task of identifying and removing bots from these platforms becomes more significant. As these bots improve, **are LMs able to categorize tweets as human or bot-made when trained on these types of tweets?**

Do bot detectors trained on more specific contexts perform better on a bot classification task than more generally trained models? More specifically, this experiment explores **whether models trained on just political or nonpolitical bot tweets and human tweets perform better on those categories than a model trained on both types.**

b'Ah these Barca players have made me angry why they didnt do some of the preseason'

b'A bunch of my buddies coast to coast blowin' me up about seeing' the RITZ promo's in the grocery stores...who woulda thunk!? #spon'

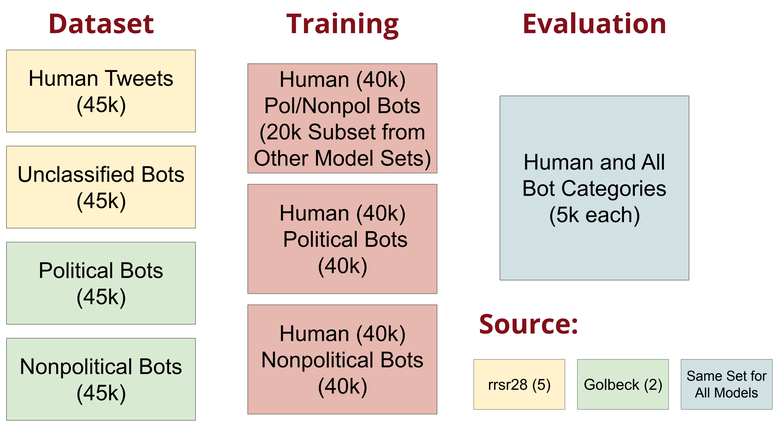
We predicted that across all model types, training on a **narrower scope of data** would **improve** classification **accuracy**. If this prediction were accurate and generalized to other contexts (besides political/nonpolitical), future bot detectors could use this concept to improve performance and protect users (as long they could differentiate tweet categories).

Experimental Setup

The experiment sourced data from Golbeck et al. and a Twitter Bot Detection Github Repository (2,5). It was conducted using both a causal (**distilgpt2**) and masked (**distilbert**) model within NLPScholar's text classification mode to assess extrapolation of results across model types (3, 6, 7). A **7:1** training to validation ratio was used to optimize weights and hyperparameters during training, and accuracy on each category is used as the primary evaluation criteria.

Retweets by bots, which overwrite text written by humans, were **removed** so the models could more directly learn textual patterns and styles of human and bot tweets.

To evaluate user-specific bias, the experiment was run once where tweets from specific users were randomized across training/evaluation data, and again where training/evaluation data was nonrandomized, containing all the tweets from the same users.



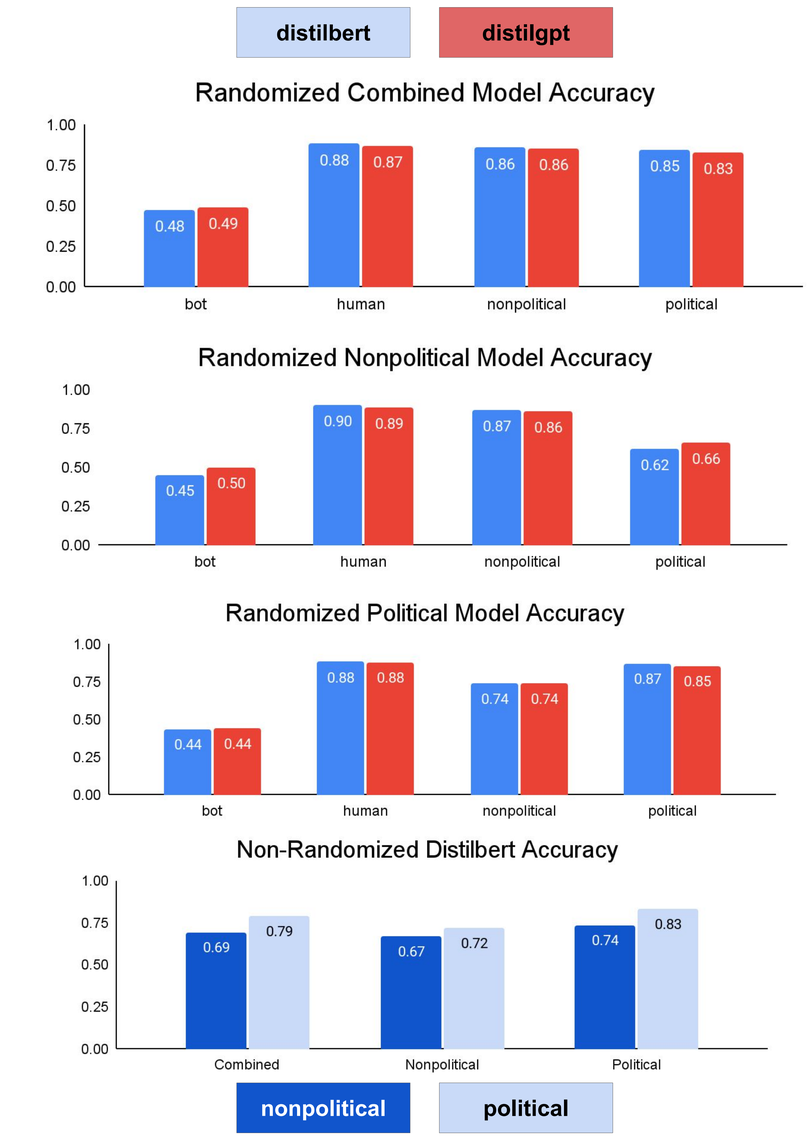
Background

Bot detection is a complex task with multiple unique approaches. In addition to tweets themselves, Feng et al. analyzed user metadata (**# of followers, following, # of posts, time between posts, etc.)** to find similarities between accounts used in training to classify them (1).

Other multimodal approaches build on this idea by observing the **types, durations, and number of tweets** made by users with the idea that stronger repetitive patterns may indicate bot behavior. Ilias et al. represented account metrics graphically and used comparisons between accounts to indicate bot likelihood (4).

A more **holistic approach** to bot detection should be used as bots become harder to detect. Combined with Ilias et al.'s classification of types of tweets made by different accounts, if narrower training results in better performance, such models could be **combined** with user metadata to produce a **new generation** of bot detectors.

Results



Summary + Conclusion

Overall, the models were highly successful at identifying bots, with many achieving **accuracies over 80%**. However, randomized models performed only marginally better than the combined model on the classes they were trained on. Additionally, all models performed poorly on bots from a different dataset, suggesting poor generalization.

Models trained on tweets from the same users (nonrandomized) performed worse and did not perform better on the specific category they were trained on. In other words, models are far **better at identifying the same users** than identifying any user as a bot. All nonrandomized models identified political bots slightly more successfully than nonpolitical, showing that **political bots are more distinctive** than nonpolitical to LM's.

The results **did not confirm our hypothesis** that LM's trained on a specific class of bots will perform better on that class. Overall, the data showed only insignificant improvements on specific classes, and these improvements were erased completely when the models had not seen those specific users. While models were highly successful at identifying bots from the same sample and tweets from the same user, **poor generalization across samples and users suggests limited applicability** against bots which are constantly adapting.

This experiment was limited by the availability of data from platforms like X, which have **costly API's**. As a result, human and nonpolitical/political bot tweets came from different datasets, which may have structural differences in sampling. Furthermore, **human tweets classified** as political and nonpolitical were also **not available** for this experiment.

Future Directions

While we tested for generalization across causal and masked models, future experiments could **repeat** this experiment on **larger** and **newer models** like BERT or GPT-5.

Bots are an issue on other platforms like Tiktok and Instagram. Future experiments could test whether **bot behavior** is **consistent** and similar **across platforms**.

Generalization across datasets was **weak** in this experiment, likely due to differences in **types** of users observed or **time horizon** of tweets. Companies like **X** have **broader access** to user metadata and tweet databases (due to user privacy). Using their resources, they can gather a more **comprehensive** set of tweets that could **amplify improvements** made by narrowing the scope of bot detection by tweet category. Utilizing the **organization** of tweets into **categories** by Ilias et al, X could assess **abstraction** of results across tweet categories and time horizons.

References

1. Feng, S., Wan, H., Wang, N., Tan, Z., Luo, M., & Tsvetkov, Y. (2024). What does the bot say? Opportunities and risks of large language models in social media bot detection. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Long Papers) (pp. 3580–3601). Association for Computational Linguistics.

2. Golbeck, J., et al. (2021). Tweets and Social Network Data for Twitter Bot Analysis. University of Maryland.

3. Grusha Prasad and Forrest Davis. 2024. Training an NLP scholar at a small liberal arts college: A backwards designed course proposal. In Proceedings of the Sixth Workshop on Teaching NLP, pages 105–118, Bangkok, Thailand. Association for Computational Linguistics.

4. Ilias, L., Kazelidis, I. M., & Askounis, D. (2024). Multimodal detection of bots on X (Twitter) using transformers. In arXiv preprint (arXiv:2308.14484). arXiv.

5. rrsr28. (2025). Twitter-Bot-Detection [Computer software]. GitHub. https://github.com/rrsr28/Twitter-Bot-Detection/blob/main/Twitter_Bot_Detection.ipynb

6. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter [Conference paper]. arXiv. https://arxiv.org/abs/1910.01108

7. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilGPT-2 [Large language model]. Hugging Face. https://huggingface.co/distilgpt2