Generalized and Generative Bot Detection

Brodric Cormier
University of California, Berkeley

1. Abstract

We propose a novel training pipeline for bot detection using generative model manufactured training data inspired by the ethos of large language model research. We show the promise of the methodology in comparison to state of the art benchmarks in bot detection while using no metadata and only raw text as the features for classification input.

2. Introduction

With the increase in model generated text being passed off as human written texts in a variety of mediums, from academic papers[1] to twitter spam, it also becomes increasingly important to be able to detect the truth from the fakes. As a result there is a big research trend centered around social media, specifically Twitter[2] and Reddit[3]. While this is an interesting problem given the short form messages and abnormal syntax and semantics present in tweets there are key issues around training data collection, and many competitive models use other forms of metadata[4] for training features. The other issue that has come out of the Twitter research is overly specific language models[5] all the way down to COVID-Twitter-BERT[6]. To progress past these shortcomings we build on the research done on large language models with strong transfer learning abilities like GPT2[7], T5[8], BERT[9] and their subsequent improvements like RoBERTa[10]. Specifically in this work we generate a training set of "bot" text using a fine tuned GPT2 that is passed into a RoBERTA based deep neural classifier using only the raw text embeddings as input.

3. Background and Related Works

The researchers behind the TweepFake[11] paper set out with a similar objective that I did of detecting deep model generated fake text. In their pursuit they compiled a data set of

[1] Gpt Generative Pretrained Transformer, Almira Osmanovic Thunström, Steinn Steingrimsson. Can GPT-3 write an academic paper on itself, with minimal human input?. 2022. https://hal.archives-ouvertes.fr/hal-03701250v1

[2] Fagni T, Falchi F, Gambini M, Martella A, Tesconi M (2021) TweepFake: About detecting deepfake tweets. PLOS ONE 16(5): e0251415. https://doi.org/10.1371/journal.pone.0251415

[3] Saeed, Mohammad Hammas, et al. "TROLLMAGNIFIER: Detecting state-sponsored troll accounts on reddit." *arXiv preprint arXiv:2112.00443* (2021). https://arxiv.org/abs/2112.00443

[4] Bhatt, Paras and Anthony Rios. "Detecting Bot-Generated Text by Characterizing Linguistic Accommodation in Human-Bot Interactions." *FINDINGS* (2021). https://arxiv.org/pdf/2106.01170.pdf

[5] Nguyen, Dat Quoc et al. "BERTweet: A pre-trained language model for English Tweets." *ArXiv* abs/2005.10200 (2020): n. Pag. https://arxiv.org/abs/2005.10200

[6] Müller, Martin et al. "COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter." *ArXiv* abs/2005.07503 (2020): n. Pag. https://arxiv.org/abs/2005.07503

[7] Radford, Alec et al. "Language Models are Unsupervised Multitask Learners." (2019).https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[8] Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *J. Mach. Learn. Res.* 21.140 (2020): 1-67. https://arxiv.org/abs/1910.10683

[9] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018). https://arxiv.org/abs/1810.04805

[10] Liu, Yinhan, et al. "Roberta: A robustly optimized bert pretraining approach." *arXiv preprint arXiv:1907.11692* (2019). https://arxiv.org/abs/1907.11692

[11] Fagni T, Falchi F, Gambini M, Martella A, Tesconi M (2021) TweepFake: About detecting deepfake tweets. PLoS ONE 16(5): e0251415. https://doi.org/10.1371/journal.pone.0251415

manually compiled accounts generating deep fake tweets. To the best of our searching this is the closest dataset in character to the one we generate. They then trained a variety of classifiers over the data, with a RoBERTa based classifier achieving the best performance over a variety of metrics. This provides a useful comparison as an evaluation metric, but slightly different to this work because they are also training on their data while we are training on other data (the data generated by the trained GPT2 model).

Creci et al.[12] also compiled a fake tweet dataset. The Cresci set is much larger, but has a slightly different focus on more traditional spam bots and those used for marketing and advertising. It is often used for benchmarking methods for Twitter spam detection. They used crowdsourcing for labeling with less than perfect agreement between the labelers so we expect some items to be misclassified.

Yang et al.[13] is the state of the art in Twitter bot detection on multiple benchmarks and aims to build a lightweight classifier using a lower amount of features. This is a very good benchmark to compare against, but even with the lightweight random forest model they utilize 20 input features so the comparison against our model with only the raw text as input needs to keep that in consideration.

4. Methods

4.1 GPT2

We started by training a GPT2 model with the popular wiki-text dataset[14] that contains about 1.8 million examples of text from Wikipedia. This model can then be used to generate passages that look like Wikipedia texts. These texts allow us to generate an indiscriminate number of longer form text samples to act as “bot” generated text to train a classifier against. This idea is derived from the idea behind popular large language models that train on large diverse collections of text to provide performant embeddings in other areas. Due to compute and time constraints this set of experiments will not provide evidence around what the data and training size need to be, but will show some support for the methodology.

4.2 RoBERTa Baseline

Taking the GPT2 generated texts and pairing them with actual examples we can generate a training and validation set in equal parts or weighted more towards generated or “human” text. Given this dataset we trained a single layer hidden layer binary deep classifier with Adam[15] optimization on top of RoBERTa[16] to classify bot or human generated text. This simple, but powerful architecture is able to very successfully differentiate between bot and human on training and validation data from the created wiki dataset.

4.3 Baseline on Tweep Fake

One of the main goals of this project was to see if a similar idea of a large general data set actually generalizes well. To test this

[12] Cresci, Stefano, et al. "The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race." *Proceedings of the 26th international conference on world wide web companion*. 2017. https://arxiv.org/abs/1701.03017

[13] Yang, K.-C., Varol, O., Hui, P.-M., & Menczer, F. (2020). Scalable and Generalizable Social Bot Detection through Data Selection. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(01), 1096-1103. https://doi.org/10.1609/aaai.v34i01.5460

[14] Merity, Stephen, et al. "Pointer sentinel mixture models." *arXiv preprint arXiv:1609.07843* (2016). https://huggingface.co/datasets/wikitext

[15] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* (2014). https://arxiv.org/abs/1412.6980

[16] Liu, Yinhan, et al. "Roberta: A robustly optimized bert pretraining approach." *arXiv preprint arXiv:1907.11692* (2019). https://arxiv.org/abs/1907.11692

we pulled in the TweepFake[17] data set to use as a validation set. Immediate performance on the TweepFake validation set is not much better than random, which is initially a concerning result for the generalized form. But, after looking closer we can see that the model is thinking all of the twitter data is model generated. This makes a lot of sense because the twitter data is more erratic in ways that mirror some of the mistakes the generative model makes.

4.4 In-Domain Classifier

In an effort to keep the Twitter noise in check, but without overtraining on the TweepFake data we trained the same model architecture as the baseline on the spam and genuine data presented by Cresci et al.[18] This also achieves excellent results on training and validation data. When evaluating this model on the TweepFake data we see slightly better, but similar results as the wiki trained models. However the failure cases are in the opposite direction. The deep fake generated texts are more convincingly human to the model trained on more traditional types of spam tweets.

4.5 Combination Model

Because of the difference in behaviors of the wiki and in-domain models a model that can balance these aspects might have a good chance at successfully delineating deep fake from real tweets, as well as in other generalized situations. Thus, we trained the same model architecture on mixtures of the two data sets. We showed the model three combinations of datasets. One with twice as much wiki data as Cresci et al. tweet data, one with an even split, and one with twice as much tweet data as wiki. The model shown twice as much wiki data was the best performing, which shows some promise of the methodology towards generalized bot detection.

Results and Discussion

Model Data Ratios	**Tweep Evaluation Accuracy**	**Tweep Evaluation Precision**	**Tweep Evaluation Recall**
2 to 1 Wiki to Cresci	**0.6036**	**0.7040**	**0.3573**
1 to 1 Wiki to Cresci	0.5237	0.5885	0.1570
1 to 2 Wiki to Cresci	0.5377	0.6301	0.1822

Table 1: The evaluation metrics of the final combination models and their corresponding data splits. All models have the same architecture and hyperparameters.

In Table 1 we can see that the 2 to 1 wiki to Cresci model is the best performer among our experimental models, but does fall short of the in-domain, multi feature benchmark of Yang et al. The mix of wiki and Cresci data achieves better precision than the sole wiki data, but does lose out on recall, which was a strength (to a fault) of the exclusively wiki data trained model. This falls short of the 89% accuracy that the TweepFake authors were able to achieve with a specific training set. However, there is support that the methodology of using a pre-trained classifier on general text characteristics with additional in domain training data does help with bot detection performance. This is especially promising because the in-domain validation results of our initial cresci-17 trained data set is

[17] Fagni T, Falchi F, Gambini M, Martella A, Tesconi M (2021) TweepFake: About detecting deepfake tweets. PLoS ONE 16(5): e0251415. https://doi.org/10.1371/journal.pone.0251415

[18] Cresci, Stefano, et al. "The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race." *Proceedings of the 26th international conference on world wide web companion*. 2017. https://arxiv.org/abs/1701.03017

comparable to state of the art in Twitter bot detection by Yang et al. as shown by Feng et al.[19]

Model	Accuracy	F1
Our cresci-17 trained model	0.9822	0.9821
Yang et al. methodology	0.9847	0.9893

Table 2: Our cresci-17 trained model validation results compared with one of the states of the art in twitter bot detection by Yang et al. implemented and evaluated by Feng et al.

Admittedly this is not an apples to apples comparison because we do not know the exact makeup of the training and evaluation sets used by Feng et al. as compared to ours. Also Yang et al. use a 20 feature random forest model, while we use a deep neural classifier over just the raw text as a feature.

Future Works

While we made progress towards a general methodology for bot detection there is still much to be done and plenty more open questions that our experiments uncovered than closed. The obvious next steps are to dive deeper into the failures of our mixed training set models to get more performance in transfer learning tasks. This could include more experimentation with architecture and hyperparameters for the classification model, larger generated text set sizes for initial training, and different lengths and mixes of lengths for the generated passages. Experiments on the generated texts to optimize the downstream input for maximum generalizability would also likely provide huge value.

Conclusion

In this work we demonstrated the promise of a generalized bot detection model built on top of a text generation pipeline. While this approach in the early stages is not able to replace highly engineered models for bot detection it shows the promise of the methodology over even in-domain attempts at transfer tasks. It also drastically simplifies the modeling and approach towards bot classification by removing the need for extreme feature engineering. The engineering and development effort will also be simplified by abstracting the research of the effort of the classifier into two separate, independent, non-coordinated pipelines of the training data generation model and the fine tuning of the classification model.

References

Gpt Generative Pretrained Transformer, Almira Osmanovic Thunström, Steinn Steingrimsson. Can GPT-3 write an academic paper on itself, with minimal human input?. 2022. https://hal.archives-ouvertes.fr/hal-03701250v1

Fagni T, Falchi F, Gambini M, Martella A, Tesconi M (2021) TweepFake: About detecting deepfake tweets. PLOS ONE 16(5): e0251415. https://doi.org/10.1371/journal.pone.0251415

Saeed, Mohammad Hammas, et al. "TROLLMAGNIFIER: Detecting state-sponsored troll accounts on reddit." *arXiv preprint arXiv:2112.00443* (2021). https://arxiv.org/abs/2112.00443

Bhatt, Paras and Anthony Rios. “Detecting Bot-Generated Text by Characterizing Linguistic Accommodation in Human-Bot Interactions.” *FINDINGS* (2021). https://arxiv.org/pdf/2106.01170.pdf

Nguyen, Dat Quoc et al. “BERTweet: A pre-trained language model for English Tweets.”

[19] Feng, Shangbin, et al. "Twibot-20: A comprehensive twitter bot detection benchmark." *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 2021. https://arxiv.org/abs/2106.13088

ArXiv abs/2005.10200 (2020): n. Pag. https://arxiv.org/abs/2005.10200

Müller, Martin et al. “COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter.” *ArXiv* abs/2005.07503 (2020): n. Pag. https://arxiv.org/abs/2005.07503

Radford, Alec et al. “Language Models are Unsupervised Multitask Learners.” (2019).https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *J. Mach. Learn. Res.* 21.140 (2020): 1-67. https://arxiv.org/abs/1910.10683

Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018). https://arxiv.org/abs/1810.04805

Liu, Yinhan, et al. "Roberta: A robustly optimized bert pretraining approach." *arXiv preprint arXiv:1907.11692* (2019). https://arxiv.org/abs/1907.11692

Cresci, Stefano, et al. "The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race." *Proceedings of the 26th international conference on world wide web companion*. 2017. https://arxiv.org/abs/1701.03017

Yang, K.-C., Varol, O., Hui, P.-M., & Menczer, F. (2020). Scalable and Generalizable Social Bot Detection through Data Selection. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(01), 1096-1103. https://doi.org/10.1609/aaai.v34i01.5460

Merity, Stephen, et al. "Pointer sentinel mixture models." *arXiv preprint arXiv:1609.07843* (2016). https://huggingface.co/datasets/wikitext

Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* (2014). https://arxiv.org/abs/1412.6980

Liu, Yinhan, et al. "Roberta: A robustly optimized bert pretraining approach." *arXiv preprint arXiv:1907.11692* (2019). https://arxiv.org/abs/1907.11692

Feng, Shangbin, et al. "Twibot-20: A comprehensive twitter bot detection benchmark." *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 2021. https://arxiv.org/abs/2106.13088

Appendix

GPT2 Generated Texts	
. This has to do with two factors. First of all, there may not be enough fat soluble lipids in the blood to compensate for weight loss in athletes who have diabetes. Secondly there is not enough fat soluble lipids in the bloodstream to compensate for weight loss in athletes who have diabetes who do not have Type 2 diabetes. As a result, there may not even be enough saturated fatty acids in the bloodstream to compensate for weight loss in athletes who have Type 2 diabetes. There is some debate as to what is wrong with Type @-@ 2 diabetes. = = Epidemiology = = According to epidemiologic studies, there have been over 3 @,@ 000 documented cases of Type @-@ 2 diabetes in the U.S. According to the Centers for Disease Control and Prevention (CDC), there have been 2 @,@ 062 hospitalizations per 1 @,@ 000 people as of 2011 for Type @-@ 2 diabetics. There have also been 18 hospitalizations per 1 @,@ 000 people as of 2011 for Type @-@ 2 diabetics. According to the World Health Organization (World Health Organization) there have been 2 @,@ 516 traffic accidents in	, " and " The Man With The Golden Claws ". On February 2, 2012, Beyoncé performed " Don 't Stop Believin'" on The Ellen DeGeneres Show. = = Formats and track listings = = = = Charts and certifications = = Notes ^ a signifies a co @-@ producer ^ b signifies an additional co @-@ producer ^ c signifies an additional producer ^ d signifies an additional producer ^ e signifies an additional producer ^ f signifies an additional producer ^ g signifies an additional producer ^ h signifies an additional producer ^ i signifies an additional producer ^ j signifies an additional producer ^ k signifies an additional producer ^ l signifies an additional producer ^ m signifies an additional producer ^ o signifies an additional producer ^ p signifies an additional producer ^ q signifies an additional producer ^ r signifies an additional producer ^ s signifies an additional producer ^ t signifies an additional producer ^ u signifies an additional producer ^ v signifies an additional producer ^ w signifies an additional producer

Appendix 1: Example of good and bad generated text. Good examples are hard even for researchers to differentiate between real and generated text, while bad examples can produce strange long blocks of formatting and ancillary document structure unlikely to happen in a real page.

Link to project repository for reference: https://github.com/bcormier1/generalizedBotDetection

Link to TweepFake datasets download:
https://www.kaggle.com/datasets/mtesconi/twitter-deep-fake-text/versions/5?resource=download&select=validation.csv

Link to wiki-text Hugging Face dataset card: https://huggingface.co/datasets/wikitext

Link to Cresci-17 dataset download: https://botometer.osome.iu.edu/bot-repository/datasets.html

Wiki-text Examples	
On its day of release in Japan , Valkyria = = As Atlanta = = The brothers Asa and Nelson Tift received the contract to convert the blockade runner into an ironclad in early 1862 with the name of Atlanta , after the city in Georgia . This was largely financed by contributions from the women of Savannah . Fingal was cut down to her main deck and large wooden sponsons were built out from the sides of her hull to support her casemate . After the conversion , Atlanta was 204 feet (62 @.@ 2 m) long overall and had a beam of 41 feet (12	= = Track listings = = " There 's Got to Be a Way " (Original album version) – 4 : 52 " There 's Got to Be a Way " (7 " remix) " There 's Got to Be a Way " (12 " remix) " There 's Got to Be a Way " (Alternative Vocal Dub Mix) = = Charts = = = Nebraska Highway 88 = Nebraska Highway 88 (N @-@ 88) is a highway in northwestern Nebraska . It has a western terminus at Wyoming Highway 151 (WYO 151) at the Wyoming – Nebraska state line . The road travels eastward to N @-@ 71 ,

Appendix 2: Two training examples from the wiki-text dataset for comparison to the generated text.