Tech Review: Detection of Texts Generated by Language Models

Kurt Tuohy
CS 410, Fall 2021
University of Illinois at Urbana-Champaign
ktuohy@illinois.edu

This paper compares several studies which aim to detect text generated by modern large language models such as GPT-2. The paper focuses on classification of such deepfake text on Twitter, and compares it with a reader-aid model used to detect deepfakes of more traditional text. First, however, the paper discusses the need to detect deepfake text at all.

Outside of Twitter's own efforts to identify and ban malicious bot accounts, services like Bot Sentinel identify the likelihood that any given account is a bot. In addition, guidelines exist to help Twitter users judge whether a given account is a bot, such as these guidelines published by Norton.

The guidelines, and much automated detection effort, focus on behavior of suspicious Twitter accounts as much as on the content of their tweets. For example, accounts that tweet the same comments as one another may be bots operating in concert.

Transformer language models such as BERT and GPT-2 potentially offer bots new ability to mimic human accounts. For example, bots based on transformer models could vary the content of their tweets more and post more natural-sounding language, making the task of detection harder (Harrag, Debbah, Darwish, & Abdelali, 2020). Any such bot accounts that slipped through detection measures would probably be harder for humans to detect as well. In a study performed by the authors of GLTR (Giant Language model Test Room), test subjects only exhibited 54% accuracy in distinguishing transformer-generated paragraphs from human-written paragraphs (Gehrmann, Strobelt, & Rush, 2019).

Aside from the human consequences, one reason to detect auto-generated text is that such text may be unintentionally used to generate new language models. BERT and GPT are trained on enormous, unlabeled text corpora, and that can include the output of older transformer language models. If this happens, the new models may train to mimic older versions of themselves as much as to mimic humans.

Researchers have developed classifiers to flag tweets and other text as bots or human, but GLTR takes a different approach. Given a block of text, GLTR highlights words and outputs simple charts to help the reader judge whether the text is deepfake.

GLTR highlights words based on their conditional probability in the context of the piece of writing. GLTR also judges the entropy of the word predictions. According to the authors, human writing features a larger fraction of low-probability words and involves fewer low-entropy predictions, and the highlighting is designed to make that apparent to readers (Gehrmann, Strobelt, & Rush, 2019).

The same subjects who judged text as machine-generated with 54% accuracy improved to 72% when using GLTR. Unfortunately, the authors knew in advance the model used to auto-generate their text samples, and they only hypothesize that there would be a benefit when the generating model is unknown.

In addition, they used well-formed text such as New York Times cooking articles in their samples. Tweets are a very different form of text. As Richard Nordquist points out, a tweet may contain no valid words in its native language at all (Nordquist, 2020). The authors of GLTR indicate that language models attempting to defeat GLTR by sampling more-low probability words would produce less coherent text, and humans could more easily identify it as machine-authored. However real tweets sometimes exclusively contain low-probability words, so this failsafe may not apply.

Another factor to consider is that, like any guidelines-based approach, GLTR depends on the motivation of the user to determine whether a text is deepfake. Users may be more motivated to check text that has neutral emotional content, but a short, punchy tweet that provokes agreement or outrage might elicit a less measured response.

Deepfake classifiers have the potential to identify deepfake tweets with nonstandard language and to circumvent the issue of reader motivation. This paper considers two such classification experiments: one on (presumably) English-language tweets and one on Arabic tweets.

The study of Arabic-language tweets was motivated in part because social media is the primary news source for a large part of the Arabic-speaking population (Harrag, Debbah, Darwish, & Abdelali, 2020). To collect their data, the researchers harvested recent tweets from accounts which had been manually labeled as human in a 2015 study (Alerekhi & Elsayed, 2015). The older study was conducted before transformer-based language models existed, which made labeling easier. The original Twitter accounts were chosen at random.

The deepfake researchers generated their own set of tweet-like text using a white-box GPT-2 model, seeded with the human generated tweets. They then used two types of models for detection: one set of RNN-based models and the AraBERT language model.

Both types of detection showed striking success, with AraBERT outperforming the neural networks to reach above 98% accuracy, precision, recall, and F1 score.

In contrast, the English-language study aimed to produce a dataset of actual deepfake tweets, rather than generating text themselves (Fagni, Falchi, Gambini, Martella, & Tesconi, 2021). Their Tweepfake dataset is available on Kaggle.

The researchers identified the bots from the human Twitter accounts that they mimicked. The human accounts were chosen because their content referred to auto-generated text technologies in some way.

In many cases the researchers were able to identify the models which generated the fake tweets. Some model types were unknown, and all models were black-box. Among the known models were RNNs and GPT-2 models. Some of the latter were fine-tuned on the human Twitter profiles they mimicked.

The Tweepfake study employed multiple classifier types to detect the deepfakes, from simple classifiers like SVM to large, fine-tuned language models. Most of the latter were based on BERT, as was the AraBERT model in the Arabic-language study.

The fine-tuned models met with success in classifying many of the deepfake sources, with the striking exception of the deepfakes known to be generated by GPT-2. For those, the BERT-based XLNet model reached only 78% accuracy, with other models performing less well. This contrasts with the success that the Arabic-language researchers met with in using BERT to classify their GPT-2-source deepfakes.

Both studies, though, were highly successful in classifying RNN-based tweets.

What could account for the difference on GPT-2-source tweets? There are many differences between the two studies, including:

- Language of tweets: Arabic vs. English.
- Knowledge domain of tweets. The Arabic-language Twitter accounts were selected randomly, but the English-language accounts included content on machine-generated text – and the deepfake tweets mimicked those specific accounts.
- White-box vs. black-box deepfake sources. The Arabic-language researchers generated their own tweets, whereas the English-language study harvested theirs from Twitter.
- Fine-tuning of deepfake models. Some of the English-language deepfake generators are known to have been fine-tuned on their target human Twitter accounts, but the Arabic deepfake generator was not fine-tuned.

A more detailed comparison of the two classification studies is in Table 1 below.

Unfortunately, their original tweets are not available for viewing. The individual subjects and styles are unknown. It would be potentially fruitful to try to characterize some of the tweets which were easy to classify and those that were hard. Did English-language GPT-2 more faithfully mimic its human targets? Did nonstandard Twitter-style language play a role?

Both studies utilized older models such as BERT to detect deepfakes generated by the newer GPT-2. It may also be fruitful to test how accurately GPT-2 can detect its own deepfakes.

Neither study utilized behavioral features of Twitter bots, such as response speed and posting frequency. It's possible that the English-language classifiers would improve if bots using sophisticated deepfake techniques resorted to standard botlike behavior.

It's difficult to compare the classification-study results to those of GLTR. The GLTR study involved traditional text rather than tweets, although it was used to detect deepfakes generated by GPT-2.

It's worth noting that GLTR plays an entirely different role than the classifiers. Any text that GLTR displays has already reached the user's eyes, making GLTR one of the last firewalls between a deepfake and the reader's brain. The classifiers, on the other hand, could help to flag deepfakes before they ever reach a reader's Twitter timeline.

Before this could work, though, the researchers would need to test their methods on one another's datasets to judge their efficacy and to adapt or expand their approaches to accommodate the new problem domains. In the meantime, large language models may be making it easier for humans to be fooled into retweeting machine-generated text, thereby giving it a human stamp of approval.

		Tweepfake	Arabic-Language Deepfake Tweets			
Tweet Language		English (presumed)	Arabic			
Tweet Counts	*Human-generated*	12,786	4,196			
	Model-generated	12,786	3,512			
Tweet Sources	*Human tweets*	Identified human accounts which mentioned auto-generated text-related technologies at some point	Collected from Twitter accounts which were labeled as human in earlier study.			
	Model-generated tweets	Identified bot accounts that mimicked the collected user accounts	Generated by researchers			
Generating Model Characteristics	*Models used to generate text*	GPT-2, RNN, LSTM, Markov chains, others	GPT2-Small-Arabic			
	Were generating models fine-tuned?	Some models fine-tuned on human Twitter profiles	Not fine-tuned			
	Were generating models seeded with initial text?	Unknown	Seeded with the human tweets			
	Black-box vs. white-box generating models	Black-box	White-box			
Data Preprocessing	*Hashtags*	Removed hashtags	Split hashtags			
	URLs	Replaced with "__url__"	Removed URLs			
	Username mentions	Replaced with "__user mention__"	Replaced with "USER"			
	Emoticons	Kept as separated tokens	Unknown			
Classifier Model Types	*Simple classifiers*	Logistic regression, random forest, SVM	-			
	Deep neural networks	CNN and GRU with character encoding	4 RNN models: LSTM, GRU, bidirectional LSTM, bidirectional GRU			
	Large language models, non-fine-tuned	BERT with simple classifiers: SVM, etc.	-			
	Large language models, fine-tuned	XLNet, BERT, DistilBERT, RoBERTa	BERT (AraBERT)			
Performance Detecting GPT-2 Generated Text		Accuracy	Accuracy	Precision	Recall	F1
	Simple classifiers	65.0-76.0%	-	-	-	-
	Deep neural networks	68.0-82.0%	94.7-96.3%	91.2-96.9%	95.2-98.2%	94.6-96.0%
	Large language models with classifier, not fine-tuned	63.0-68.0%	-	-	-	-
	Large language models, fine-tuned	73.0-78.0%	98.7%	98.9%	98.5%	98.7%

Table 1: Comparison of Arabic-language and English-language deepfake tweet classification experiments

References

Alerekhi, H., & Elsayed, T. (2015). Detecting Automatically-Generated Arabic Tweets. *Asia Information Retrieval Symposium (AIRS)*, 123-124.

Fagni, T., Falchi, F., Gambini, M., Martella, A., & Tesconi, M. (2021). TweepFake: About detecting deepfake tweets. *PLoS One*.

Gehrmann, S., Strobelt, H., & Rush, A. M. (2019). GLTR: Statistical Detection and Visualization of Generated Text. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 111-116.

Harrag, F., Debbah, M., Darwish, K., & Abdelali, A. (2020). BERT Transformer model for Detecting Arabic GPT2 Auto-Generated Tweets. *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, 207-214.

Nordquist, R. (2020, 2 12). *What Is a Tweet? How Twitter Is Changing Our Language*. Retrieved from ThoughtCo.: https://www.thoughtco.com/tweet-definition-1692478