Trump or Computer Dump?

Akshay Trikha
Harvey Mudd College / Claremont, California
atrikha@hmc.edu

Abstract

Can we fine-tune the second generation of the Generative Pre-Trained Transformer (GPT-2) language model created by OpenAI to generate Trump tweets that sound authentic? Furthermore, can we gamify it and create an experiment with our classmates to see if they can tell the difference between the real Donald and his computerized comrade? Two versions of differing capacity of the GPT-2 language model were used to generate synthetic Trump tweets. The success of each model was evaluated using the scores from a survey evaluating people on how well they could discern between a model and ground truth tweets, as well as by analyzing the lexical similarity of each model's generated text and ground truth tweets.

1 Related Work

The task at hand has already been attempted in a fairly passable manner. One tried and tested method is using a Markov model of n-grams extracted from a training set of Donald Trump's tweets [Hráček, 2020]. This model captures the lexical structure of Trump's tweets somewhat well, but its overall sentences do not make enough sense semantically to trick people into believing that they are not computer generated. Examples of synthetic markov chain based tweets that speak to this are "Entrepreneurs: Everything starts with you, you need all the primary debates and you have no respect" and "The brand new and even less understanding of success-and very lazy!" - where it seems that the longer the tweet the less probability there is that it will semantically deceive a person [Hráček, 2020]. Furthermore, because a markov chain being used is based only on its training data, its vocabulary is also limited to whatever it was trained on - which were 1.163MB or roughly 9000 tweets.

GPT-2 is a neural network based language model that has already been trained on 40 billion tokens from the WebText dataset [Radford et al., 2019a]. This dataset was created using text scraped from the social media website Reddit. The model has four variants based on the number of parameters which range from 117 million to a staggering 1.5 billion parameters [Radford et al., 2019b]. Its performance has been demonstrably strong in not only in text generation tasks, but also in the similar problems of question answering, summarizing, and translation tasks [Radford et al., 2019b]. An established and well cited example of when it has been used successfully for question answering is in the context of task-oriented dialog, where authors measured performance based on whether or not the system was able to answer requested attributes [Budzianowski and Vulic, 2019]. Our investigation builds on ideas from this work to use the GPT-2 language model to generate even more realistic-sounding Trump tweets.

Though this paper will explore using GPT-2 more or less out of box, it is important to note that other researchers have already started looking at ways to improve it. Notably, OpenAI researchers have already gone one step further and used reinforcement learning techniques to further fine-tune their language model, but for text continuation tasks [Ziegler et al., 2019]. Another step that researchers at Uber AI have taken is to pair GPT-2 with a user specified bag of words to control the topic or sentiment of generated text [Dathathri et al., 2019].

2 Transformer Architecture

The transformer architecture for a neural network was first proposed by Google and University of Toronto researchers in late 2017. Its introduction was an important advancement for the natural

language processing community and people were amazed by reported performance particularly in machine translation [Vaswani et al., 2017].

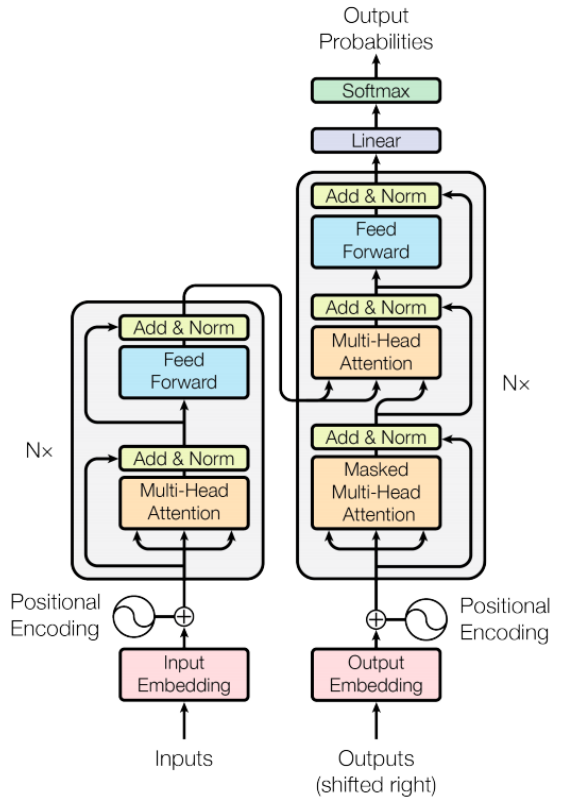

Figure 1: The transformer model architecture [Vaswani et al., 2017].

The transformer model's architecture is broadly split up into two encoder and decoder blocks. It allows input sequences to be passed in simultaneously and word embeddings also to be calculated simultaneously with one another. A unique part about this architecture is its incorporation of attention, which at a high level is defined as informing a model what part of an input sequence to focus on [Ajay Halthor, 2020]. This lets the model focus on a word in a given input sentence and calculate its relevance with respect to other words in the input.

GPT-2 goes a step further than usual by generating word embeddings as well as positional embeddings, meaning vectors capturing the context of a word within a sentence. This is particularly useful for the task at hand because it means that a GPT-2 model will take tweet structure into account when training, which is important for capturing the style of an author.

In any machine learning model, a parameter is a number that a model learns by constant updating during the training process. In a neural network, such as GPT-2, a parameter is the weight ascribed to the link between two neurons. It is then easy to marvel at how mind-boggling the size of the GPT-2 models are, and how computationally expensive they must have been to pre-train.

3 Methods

3.1 Data

Data was obtained from a GitHub repository historical records of Trump tweets dating from 2009 to 2018. This was due to time constraints and prioritizing GPT-2's training over time working with Twitter's API with its rate limits. Only tweets from 2015 till 2018 were considered as this would guarantee more political content, as well as the observation that Trump's characteristic style of communication blossomed in those years. A total of 17,797 ground truth tweets amounting to 2.409MB were used to train GPT-2. Retweets from the original dataset were discarded so as to focus on text generation from Donald Trump's perspective. Data was stripped of anything but the source text of the tweet, and all combined into a continuous text file separated by a newline delimiter.

3.2 GPT-2 Training

A wrapper for the GPT-2 library that OpenAI released on GitHub was used to access the pre-trained model as its functions were easier to use [Woolf, 2020]. It was then tweaked for our purposes by fine-tuning it with the Trump twitter source material. The 124 million parameter model (124M) and 774 million parameter model (724M) were used to explore the relationship in model capacity and quality of generated output. The 1.5 billion parameter was attempted to be fine-tuned but even while training on an NVIDIA Tesla T4 GPU with 27.4GB of RAM there wasn't enough computational power for a single iteration of training. Each model was trained on roughly 2000 iterations with 124M having a learning rate of 1E-4 and 724M having a learning rate of 5E-4 - which were determined empirically based on training time. After the models were trained, they were tasked to generate tweets. Interestingly, the models did not need to be limited on the number of words in generated - so as to adhere to Twitter's 280 character per tweet policy - as they learned what a reasonable tweet length was in their training. 9505 tweets or 1.115MB worth were generated from the smaller model, 124M. 9505 tweets or 1.115MB worth were

generated from the smaller model, 124M. 9362 tweets or 1.115MB worth were generated from the smaller model, 124M.

3.3 Human Verification

A series of artificially generated tweets mixed with ground ground tweets were displayed to human test participants to see how well each individual performed in discerning reality from computer generated text. GPT-2 generated tweets were selected by random using numpy.random.choice(). The test was split into two sections, one for each model, with each section containing 5 model generated tweets and 5 real tweets. This structure allowed for the study each model's performance independently of the other.

Each tweet was embedded into a real twitter tweet's HTML and its source code was modified to account for the generated or real tweets. Dates from Twitter were redacted from test participants because that contextual information would assist a human in determine the validity of a tweet. Timestamps, like count, retweet count, and quote counts from the original online tweet were left in place as they had less influence on a test participant's performance. Retweets had their name, twitter handle, and display picture further redacted so as for the focus of the tweet to be in its text. Example tweets that were displayed to participants are displayed in the Figure 2 below:

Figure 2: Sample tweets that were displayed to test participants. The first was generated by the 774 million parameter version of GPT-2 and the second was generated by its 124 million parameter version.

4 Results

4.1 Survey Results

Participants of the study found it to be challenging. The average score between 45 people was 10.76 out of 20 points with a median of 11 and a range of 5 to 18. A possible explanation for this surprisingly below-par performance is that Donald Trump's tweets are known to be erratic, so it is indeed quite difficult to discern if they were actually posted online. Participants were mostly in their early 20s and from the United States and Singapore - but pursuing college in the US - with a select few older adults from India and Singapore. The dist2ibution of scores is depicted in Figure 4 below.

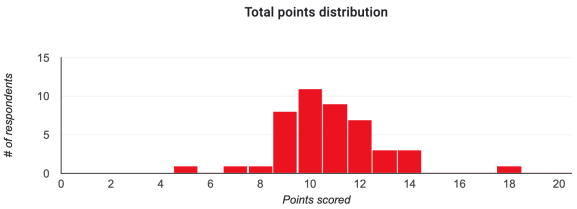

Figure 3: Score distribution of quiz participants.

The notable outlier who scored 18 out of 20 was the person who designed the test itself! Even he was fooled by it, admittedly embarrassingly. The other outlier with the poorest score was a 34 year old from the UK who identified as more left-leaning than right and reads the news every day. Without including myself, the average score was 10.59 out of 20 with a median of 11 and a range of 5 to 14.

The tweet that perplexed the most people is the first one shown Figure 2. It turns out there have been many similar tweets that have been tweeted in real life, with the most similar being "A TOTAL WITCH HUNT!!!" with 3 exclamation marks rather than 1. Given that the model was only fine-tuned with around 2.4MB worth of tweets it was reasonable for it to generate tweets that seemed like they were memorizing the data.

The tweet that had the best classification performance amongst participants was:

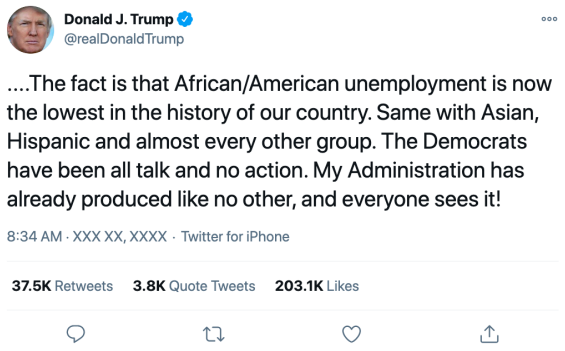

Figure 4: Tweet with best classification performance with 40 out of 45 participants correctly classifying it as real.

One potential explanation is that this tweet is long but still maintains sense throughout its sentences. Participants probably had intuition on how difficult it is for a computer to generate such a long and sensible tweet, and therefore tagged it as real.

Interestingly, participants had a more difficult time discerning 124M's tweets from reality. Their average score on tweets generated by 124M were 2.644 whereas the average score of participants on 774M's tweets was slightly lower at 2.133 - which makes sense because 774M was trained with around 650 million more parameters. Given that the study's sample size was only 44 and there were only 5 tweets each from GPT-2 model I think this gap was understated in this study. With more questions and a larger sample size it is my belief that the 774M would more clearly outperform the 124M one. However, something has to be said about the low overall scores overall. The classification task at hand is deemed a difficult one and a though this study was a good start a larger one with longer questions must be conducted for more solid conclusions.

4.2 Tweet Lexical Similarity

There are two categories of text similarity in natural language processing that people refer to, lexical and semantic. The former is concerned about the structure of language - sentence lengths, punctuation used - while the latter is concerned with the meaning between two given sentences. It was the aim of this project for the GPT-2 generated tweets to be novel, but still sound like they could have been authored by Donald Trump. As a result, it made more sense to analyze the structure of a model's tweets, to capture the style of the author, in comparison to ground truth tweets, instead of grouping a model's and ground truth tweets by their many topics and then asking how semantically similar they were.

While semantic similarity has many standard metrics like BLEU and ROUGE, which are widely used for machine translation, there isn't a standard metric for lexical similarity. As such, a handful of simple measurements were combined together to evaluate the similarity between 124M's tweets and ground truth tweets, and 774M's tweets and ground truth tweets.

One of the factors that was considered in the same vein as the Jaccard similarity coefficient was how many new words that a GPT-2 model incorporated in its tweets which were not presented in its fine-tuning corpus. For 124M, 4980 out of 17271 tokens, or 28.8%, weren't present in the ground truth corpus and the same for 774M were 7585 out of 17795 tokens or 42.6%. With this measure, we could say that 124M's generated text was more similar to the ground truth than 744M's was.

Another factor that was explored were how similar the most common unigrams, bigrams, trigrams, 4-grams, and 5-grams were. The table in the Appendix depicts these results from the 3 documents containing generated and ground truth tweets. Punctuation that is frequently used by Donald Trump were treated as separate words. There is a large overlap between the most frequent GPT-2 generated and real n-grams, but it is noticeable that the ground-truth ones read are more stable. This is meant in the sense that both GPT-2 models have tended to generate either long strings of a repetitive word or phrase or a long chain of separate tweets containing that phrase to the extent that no human would. Something that is potentially not a concern for this project is that it seems like both GPT-2 models might have been trained to the extent that they overfitted on the training data. Perhaps in the future it would be wise to train using less iterations.

The average word lengths over all tweets were quite similar. For 124M there were 4.971 characters per word while 774M had 4.577, and in reality it was 4.656 characters per word. The average tweet lengths had slightly more variance. For 124M there were 15.486 words per tweet while 774M had 17.757, and in reality it was 19.092 words per tweet. In both of these metrics, 774M outperformed 124M in in being closer to tweets from reality.

It also was quite surprising that a dispropor-

tionate number of tweets generated by both GPT-2 models were akin to Trump quoting a tweet that quoted another user - as opposed to directly retweeting them. 59.2% of ground truth tweets were like this, but 124M amplified this to 83.9% and so did 774M to a lesser extent of 72.3%.

A further interesting factor to consider that wasn't in this study due to time constraints is comparing each text source's entropy and relative entropies. This might lead to insights into just now novel GPT-2's text generation is. Finally, in the future it would be worthwhile to try and combine the aforementioned metrics together to obtain a final overall score for an easier comparison of the models.

5 Conclusion

The results from this study were even better than expected! Both models produced tweets that confused participants, with 774M being more successful in that regard. In terms of lexical similarity of the generated texts and ground truth corpus, 774M's output was again more structurally similar than 124M's. With GPT-3's release on the horizon, the world needs to be more prepared and educated for what's about to come. I envision a future of language model text generation paired with deep fake misuse as a serious threat to society that can only be countered again if enough diligent citizens understand the technologies creating them and always question the media they consume.

References

Filip Hráček. Automatic donald trump. *Website*, 2020.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019a.

A. Radford, Jeffrey Wu, R. Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019b.

Pawel Budzianowski and Ivan Vulic. Hello, it's GPT-2 - how can I help you? towards the use of pretrained language models for task-oriented dialogue systems. *CoRR*, abs/1907.05774, 2019. URL http://arxiv.org/abs/1907.05774.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *CoRR*, abs/1909.08593, 2019. URL http://arxiv.org/abs/1909.08593.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. *CoRR*, abs/1912.02164, 2019. URL http://arxiv.org/abs/1912.02164.

A Vaswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, AN Gomez, L Kaiser, and I Polosukhin. Attention is all you need. arxiv 2017. *arXiv preprint arXiv:1706.03762*, 2017.

CodeEmporium Ajay Halthor. Transformer neural networks - explained! (attention is all you need). *YouTube*, 2020.

Max Woolf. gpt-2-simple. *GitHub Repository*, 2020.

6 Appendix

6.1 Most Common N-grams by Text Source

Text Source	Unigrams	Bigrams	Trigrams	4-grams	5-grams
124M GPT-2	. the : ! ,	: @realDonaldTrump Thank you will be of the for the	I will be ! Thank you MAKE AMERICA GREAT AMERICA GREAT AGAIN GREAT AGAIN !	MAKE AMERICA GREAT AGAIN AMERICA GREAT AGAIN ! I will be interviewed will be interviewed on ! I will be	MAKE AMERICA GREAT AGAIN ! I will be interviewed on an ass! ! Of course ass! ! Of course he's ! Of course he's an
774M GPT-2	. the , ! :	: @realDonaldTrump Thank you will be of the for the	who who who ! Thank you I will be GREAT AGAIN ! MAKE AMERICA GREAT	who who who who MAKE AMERICA GREAT AGAIN AMERICA GREAT AGAIN ! "Obamacarefail" x 4 I will be on	who who who who who MAKE AMERICA GREAT AGAIN ! "Obamacarefail" x 5 will MAKE AMERICA GREAT AGAIN . MAKE AMERICA GREAT AGAIN
Ground Truth Tweets	. , the ! to	: @realDonaldTrump of the Thank you , and will be	! Thank you I will be . Thank you AMERICA GREAT AGAIN GREAT AGAIN !	MAKE AMERICA GREAT AGAIN AMERICA GREAT AGAIN ! Make America Great Again America Great Again ! ! I will be	MAKE AMERICA GREAT AGAIN ! Make America Great Again ! was my very great honor to I will be interviewed on . MAKE AMERICA GREAT AGAIN

Figure 5: The most common N-grams found for each model were similar to that of the fine-tuning corpus, with the exception of the 774M parameter model producing some unusual repetitive words and phrases.

6.2 GPT-2 Generated Tweets Shown to Test Particpants

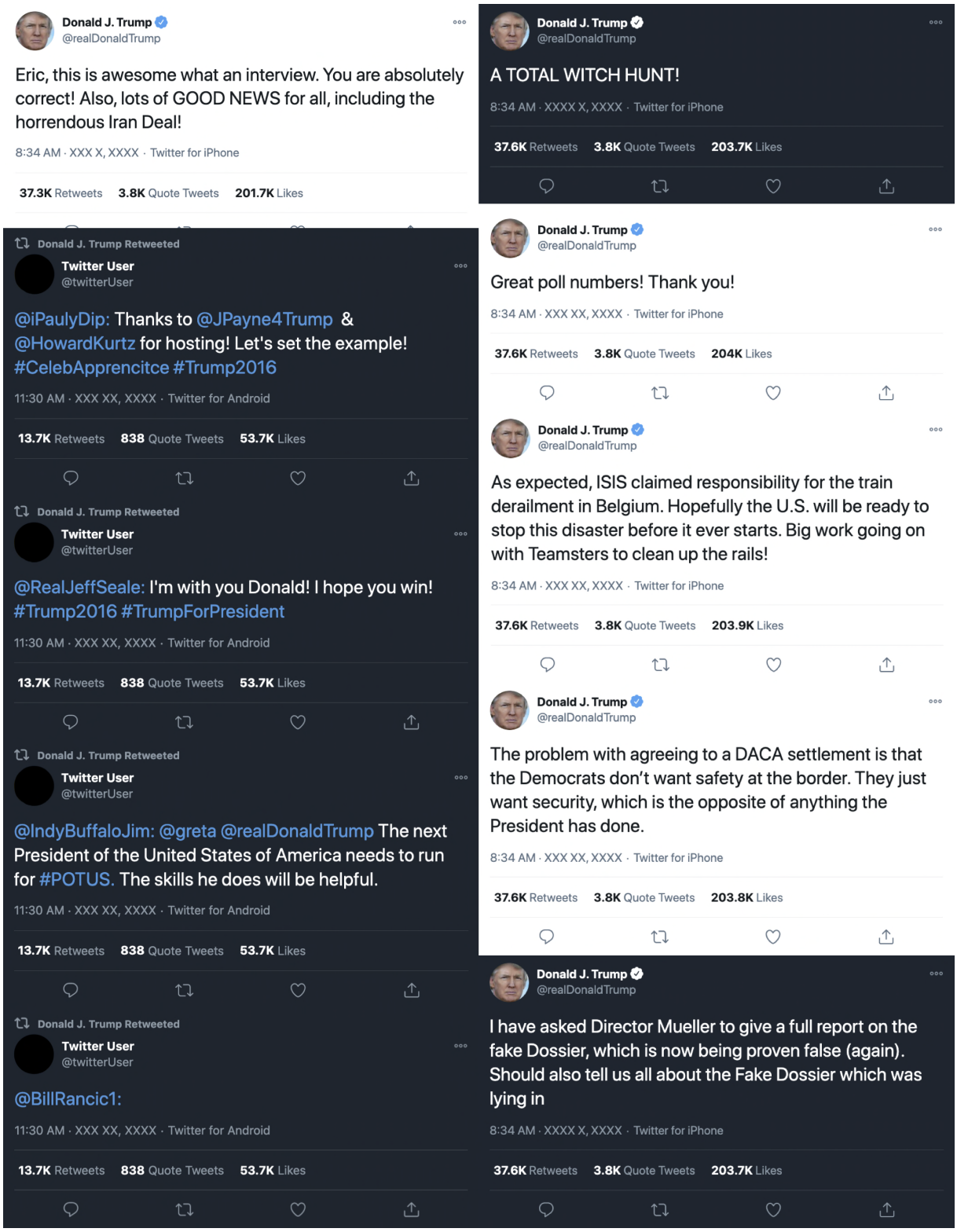

Figure 6: Left column tweets were generated by the 124M parameter model and the right column were from the 774M parameter model.