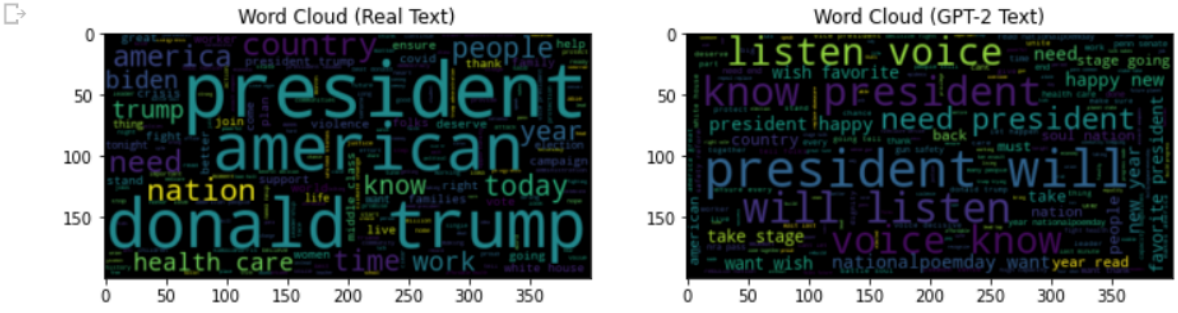
Augmented Chabots:
Personalizing Text With Natural Language Generation

Aashish Nair, M.S.
The George Washington University

Abstract

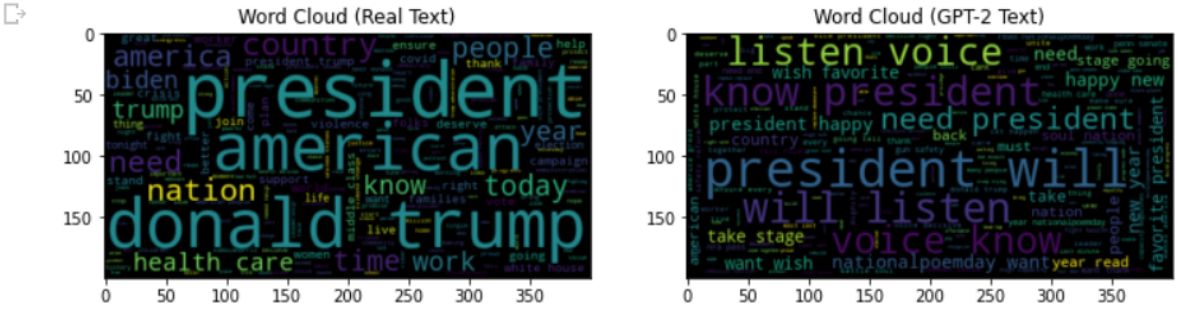
The advent of chabots, AI avatars, voice assistants, and other query based systems have improved operations in many industries. These products help save businesses significant amounts of time, money, and labor that can be allocated towards other areas. However, this technology has hit a plateau, with bots only being able to perform a limited range of functions. This project aims to explore the prospect of enhancing the performance of a bot by training it to generate personalized text as opposed to pre-written responses. Such a feat will require the use of natural language generation. In the study, 2 recurrent neural networks and a GPT-2 model were trained with tweets from 30 celebrity twitter accounts and were evaluated based on how their generated text compared to the original text in terms of writing structure and emotional content. The thorough data analysis and data modeling concluded with the GPT-2 model being deemed as the optimal model. The GPT-2 model's generated text met the requirements for sentence length in 16 out of the 30 twitter accounts. Its sentiment scores also met the requirements in 15 out of the 30 twitter accounts. Finally, its subjectivity scores met the requirements in 10 out of the 30 twitter accounts.

Keywords: natural language generation, text generation, recurrent neural network, GPT-2



Introduction

Question-and-answer systems have emerged as a highly demanded tool in many businesses and organizations. By using computer programs and artificial intelligence to supplant human workers with bots, companies can engage with their consumers and clients at a much larger scale. Automating this part of the business model allows for consumers to submit their queries and receive a response with much greater ease and in much less time. Chabots, for instance, have soared in popularity to the point where "nearly 40% of internet users worldwide prefer interacting with chatbots than virtual agents" (Insider, 2021). It also spares companies the burden of delegating manpower to cater to much of the people's queries. Thus, bots have become ubiquitous in many fields, such as banking and finance. However, despite the numerous benefits that come with this technology, bots do not replicate the experience that one receives from human interaction. Tech giants such as Apple, Google, and Amazon acknowledge that consumers value and expect "a consumerized, personalized experience" from their services, which creates a need to customize bots for each individual (Sudhakar, 2021). After all, no service or product should impede a company's ability to provide satisfying customer experience. Since reverting back to using manpower to perform menial tasks is infeasible in this technological era, businesses can only aim to enhance their chatbots and avatars to a level where their performances better replicate human interaction. Making a transition to a creative chatbot that

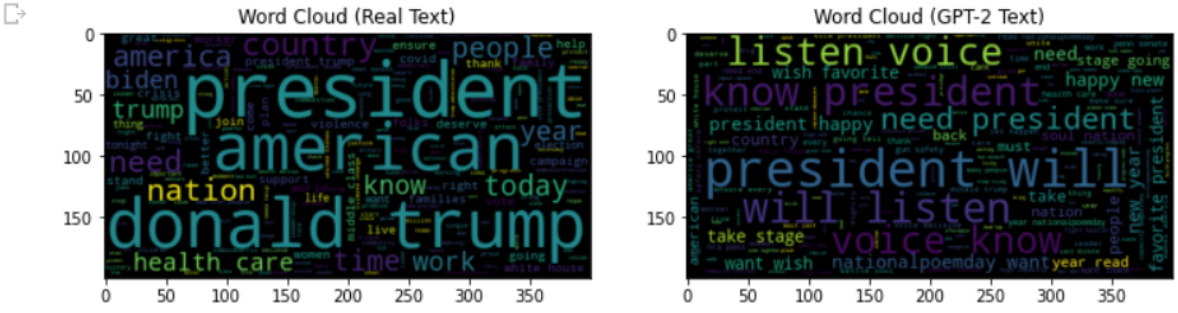


does not rely on pre-written responses brings about a plethora of challenges. A crucial step that must be taken to develop chatbots is to personalize them.

This project explores the idea of personalizing bots through the use of natural language generation. Instead of relying on a bot that contains a limited set of text responses, having a bot that can project a "personality" can grant users a more satisfying experience. This begs the question: Can a bot, trained with a specific body of text, generate its own text with a writing structure and personality that accurately resembles the original text? The project aims to answer this by taking text from famous celebrities and generating text that is similar to the content and structure of the corresponding celebrities' texts. Attempting to build a creative chatbot that can write its own words and sentences will give more insight into how realistic the prospect of a personalized chatbot is and what obstacles must be overcome to realize this technology.

Literature Review

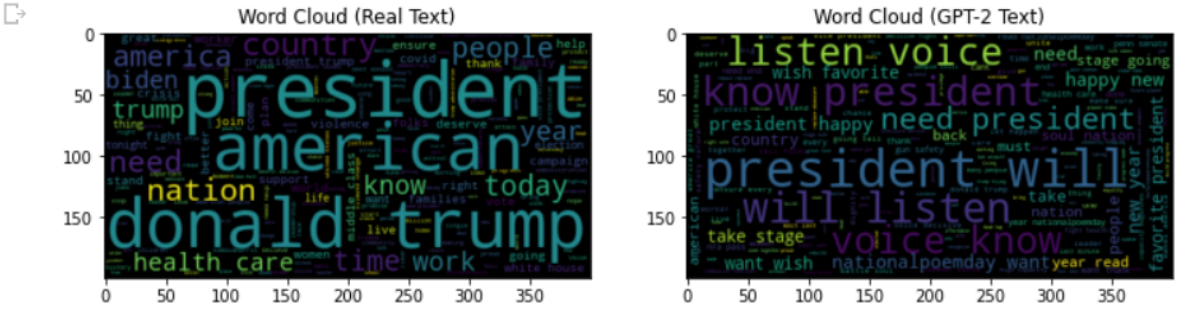
Text generation entails the use of Natural Language Generation (NLG) algorithms that generate text after being trained with a text corpus. NLG deviates from traditional NLP in that NLP focuses on collecting data from given text, whereas NLG creates new data from given text (SkimAI, 2020). Analysts utilizing NLG stress the importance of training their models to generate text that can pass off as literature created by a human. This requires emphasis on the factors such as subject topics, sentence patterns, word choice, and syntax. Text generation can render traditional "one size fits all" bots obsolete. A successful NLG model can customize its responses depending on the user and the context of the situation. Automated responses that take into account these complexities will improve the customer experience, thereby increasing customer loyalty. Chatbots and other question-and-answer systems can be personalized if they are trained with NLG algorithms that effectively generate text that consider context and adhere to the required format of the text.



There are several established approaches towards developing models that can generate text. Recurrent neural networks (RNN) are often relied upon in NLG. The long short term memory (LSTM) model, in particular, excels at predicting future sequences by retaining information on previous inputs (Li, 2018). When trained with a large enough text corpus, it will be able to generate data with creativity while retaining structure. Another architecture that has been tested is the generative adversarial network (GAN). A GAN is composed of a generator and a discriminator. A generator creates the artificial text, while the discriminator distinguishes between the real and the generated text. The constant cycle of text generation and identification improves the performances of the generator and discrimator alike (Lu, 2018). By the time the GAN is finished training, it should be able to deliver state-of-the-art results.

Popularized pre-trained models suited for NLG were also examined in addition to the aforementioned LSTM model and GAN. One such model is OpenAI's GPT-2 model, a text generation model trained with 1.5 billion parameters. The model uses an extensive range of vocabulary and has a strong grasp of context (Singh, 2019). It is capable of creating large bodies of text after being trained with a short text sample. The GPT-2 model has many parameters when run on Python. The primary parameters to consider are the "temperature" and the "top_k" parameters. The "temperature" parameter affects the randomness of the word selection, with a larger value yielding more random completions. The "top_k" parameter considers the number of tokens considered before selecting the next word. GPT-2 models have the potential to create systems like AI writing assistants and dialogue agents (OpenAI, 2019). The model is fitting for this project, which aims to generate words and sentences based on a specific body of text.

Text generation deviates from other AI challenges (e.g. classification, regression) since it incorporates elements of freedom and creativity. Models that perform tasks like classification can be assessed on the differences between their predictions and the actual outcomes. However, text generation algorithms do not aim to replicate text verbatim. It seeks to generate different text while maintaining the style of writing and content in the original text. As a result, text generation

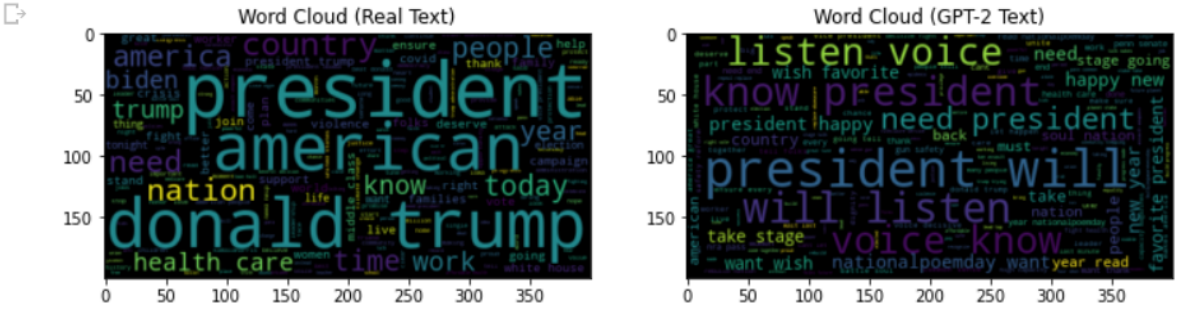


models are difficult to evaluate. Data scientists have utilized various methods for examining text generated by their models. Celikyilmaz suggests using a combination of machine-learned evaluation metrics and human judgement (2020). Machine-learned metrics such as sentence semantic similarity and factual correctness are quantifiable and will help evaluate a NLG algorithm's performance. Human centric evaluation is also necessary in such a study. Since humans will ultimately be reading the generated text, they will also be needed to adjudicate the quality of the generated text. The study by Celikyilmaz asks multiple people to critique samples of generated text based on fluency, coherence, and correctness (2020). The subjects are asked to rate each criteria from 1 to 5 in surveys. This approach allows researchers to quantify variables that are otherwise immeasurable.

Data

The data used for the project comes in the form of tweets from popular celebrity Twitter accounts. Twitter was chosen as the primary source of text data as it offers a database of texts from different personalities that can not be found in other mediums. Common sources of text such as books and articles often tend to be neutral and would not suit a study of personalized text generation. Celebrities that write their tweets are often unrestrained and speak as they wish (e.g. slang, swear words, improper grammar) as a much more suitable source of data. One caveat of Twitter is that the users are forced to write in a certain format that does not completely align with traditional writing or speech.

The tweets used for the project come from 30 people from different backgrounds. The twitter accounts in this study belong to Barack Obama, Joe Biden, Elizabeth Warren, Pope Francis, LeBron James, John Cena, Kevin Durant, Ronda Rousey, Anthony Joshua, Kevin Hart, Emma Watson, Neil Patrick Harris, Harry Styles, Dwayne Johnson, Ellen Degeneres, Jimmy Fallon, Oprah Winfrey, Conan O'Brien, Gordon Ramsay, Daniel Tosh, Jeff Weiner, Bill Gates, Elon Musk, Kylie Jenner, Tim Cook, Lady Gaga, Wiz Khalifa, Louis Tomlinson, Alicia Keys,



and Mariah Carey. These individuals are famous identities that belong to one of 6 categories: leader, athlete, entertainer, TV personality, entrepreneur, and artist.

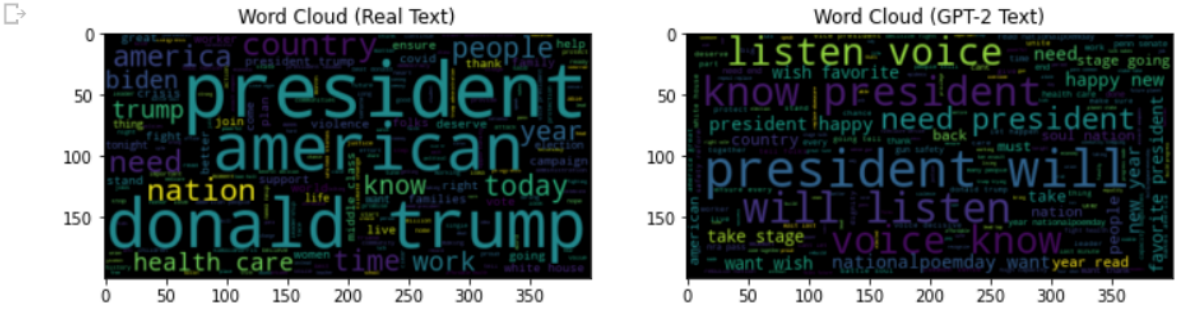
The project utilizes a large pool of subjects in order to identify potential trends or patterns in the text generated for each celebrity. For example, is there a correlation between the occupation of a celebrity and the type of tweets (e.g. sentence length, word choice) they write? Are there large disparities in the level of performance of the text generated by the models? If so, what could explain such disparities.

Most of the celebrities' tweets were procured through a tweet scraping software called Seobot. The only exceptions to this were the tweets from Donald Trump and Joe Biden, which were obtained from Kaggle.com. Over 50,000 tweets were collected for the project. For each tweet, the data contains 6 features: the contents of the tweet, the Twitter account's user handle, the username, the date of posting, the number of likes, and the number of retweets.

Research Methodology

The bulk of the project was executed using Python 3.6. The code was written with Google Colab and with Jupyter Notebook (.ipynb format). The preprocessing and the exploratory data analysis portion of the project mainly utilized the Numpy, Pandas, NTTK, and Spacy packages. The data modeling was conducted with the Keras Framework. The study also includes transfer learning, where a miniaturized version of the GPT-2 model (124 million parameters) was trained to generate celebrity tweets. The results of the EPA and text generation were visualized using a combination of packages in Python (e.g. seaborn, matplotlib) and Tableau.

Details regarding the code can be found on the github repository: https://github.com/anair123/DATS-6501---Capstone-Project---Aashish-Nair. The tableau workbook that visualizes the results can be seen here: https://public.tableau.com/profile/aashish2358#!/vizhome/CapstoneProjectDashboard_16179011454220/Dashboard1. Finally, access to the website that covers the project can be viewed with the following link: https://anair123.github.io/Augmented-Chatbots/.

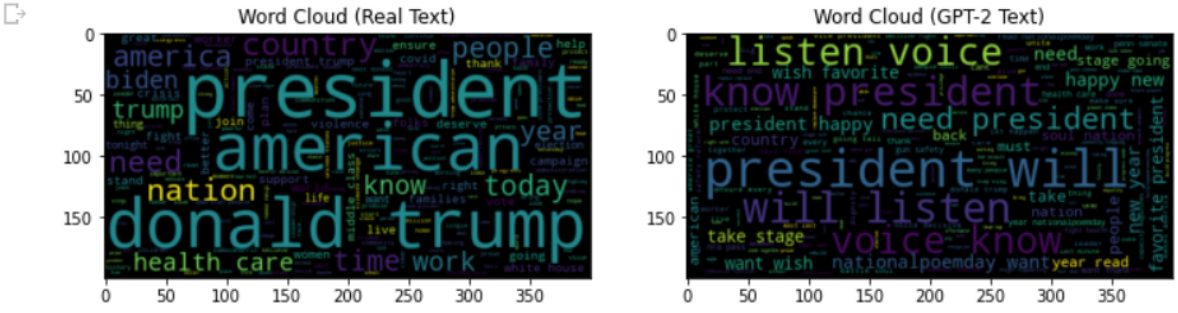


Data Analysis

The data is first subject to a thorough data preprocessing procedure. Tweets that were too short (i.e. 3 words or less) or tweets that were classified as retweets were omitted. Within the remaining tweets, all links, emojis, and characters that were not alphabets, numbers, or punctuation were removed. By the end of the data preprocessing, the dataset had 36,343 rows and 6 columns.

In search of the best model for generating text based on tweets, 3 models were utilized. The first 2 were LSTM models built from scratch. One LSTM model is a baseline model with minimal complexities, consisting of only 1 hidden layer. The other LSTM model was subject to more experimentation. Different combinations of hidden layers, neurons, and activation functions were tested in search of the best performing model. The final modified LSTM model had 3 hidden layers with 256, 128, and 64 neurons, respectively. Each hidden layer used the 'relu' activation function and was followed by a dropout layer.

The project also relied on transfer learning with the inclusion of OpenAI's GPT-2 model. Since the original GPT-2 model is massive (1.5 billion parameters), training it with each celebrity's data would require an exorbitant amount of time and computation. Thus, a miniaturized version of the GPT-2 model (124 million parameters) was used instead. The text generation with the GPT-2 model was tested with different "temperature" and "top_k" values. The generated text was compared to the original tweets with various approaches. The first method entails evaluating the generated text with quantitative metrics such as sentence length and variation since different people write texts at different lengths. A confidence interval of the average sentence length (alpha=0.99) will be calculated for the real text and the generated text for each celebrity. If the confidence intervals overlap, it will signal that the text generation model successfully replicated the celebrity's text in terms of sentence length. Furthermore, the emotional content of the original and the generated text are also compared through their



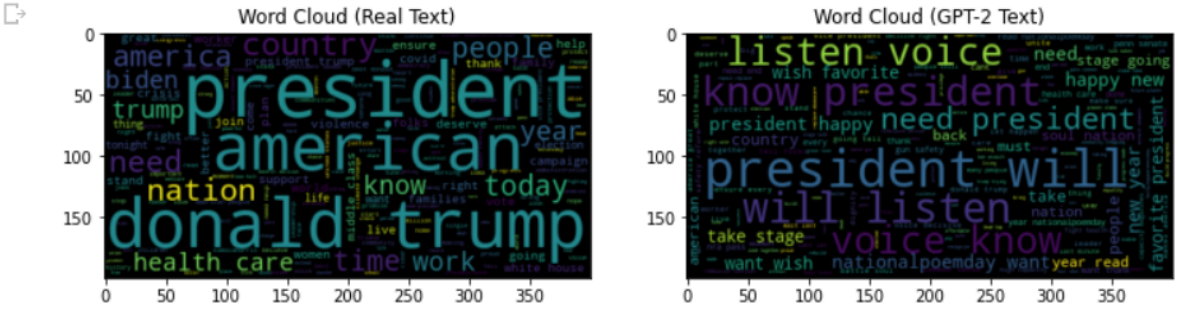
sentiment and subjectivity scores, which are obtained with the TextBlob package in Python. The generated text of a celebrity's tweets will be deemed to have similar emotional content if its sentiment score and subjectivity score differs from that of the original by less than 0.1.

In addition to quantifiable metrics, the generated text is gauged with human judgement. In other words, if a person were to read the text created by the model, would they be able to distinguish between the original and generated texts? Such an analysis will require scrutiny of the writers' grammar and diction. Word clouds are built from the original and generated text to serve as visual aid during the assessment.

Key Findings

During the exploratory data analysis, there was no clear correlation between the occupation of the celebrity and the makeup of the text. Celebrities' texts did not share similar sentences in terms of the mean and standard deviation of the sentence lengths compared to texts of celebrities in the same industry. For example, figures like Joe Biden, Elizabeth Warren, and Donald Trump, who are all prominent figures in US politics, have sentences of contrasting lengths and variation. This suggests that occupation or other external variables do not need attention when attempting to personalize text.

In the data modeling phase, the text generated from the LSTM models built from scratch yielded underwhelming results and, as a result, highlights the magnitude of the challenge that comes with text generation. For all personalities, many sentences were incoherent and sometimes even incomprehensible. A typical string of generated text consists of multiple typos and grammatical errors. Even if the text is decipherable, it can not qualify as material that can be used in any type of AI software. Some models generated text that were completely indecipherable, generating a seemingly random sequence of characters. Common phrases such as "thank you" and "I love you" were generated accurately for the twitter accounts that frequently used those phrases, but other less popular phrases were not reproduced by the model adequately.



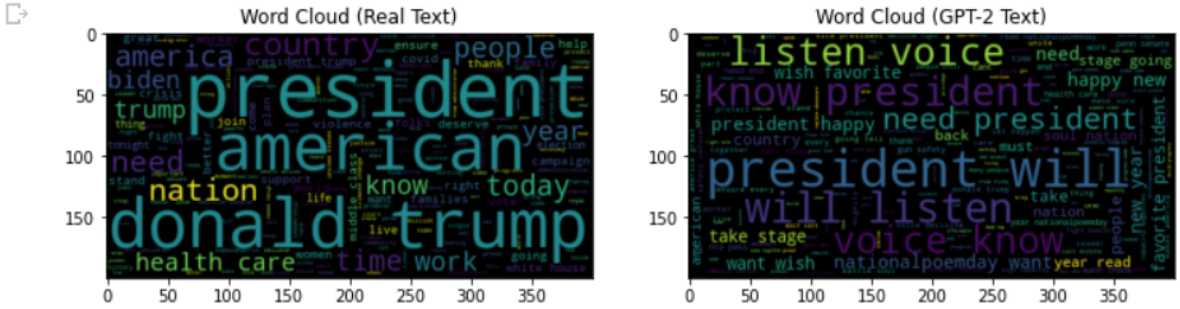
Due to the numerous typos in the generated text from the LSTM models, the models fail from the perspective of a human. Thus, calculation of the performance metrics such as the sentiment score and the use of visualization tools such as word clouds are also infeasible.

The GPT-2 model, on the other hand, had a more promising outcome. It generated text that, on surface level, appeared to resemble the text written by the celebrities. From a human standpoint, the sentences were coherent, the word choices of both texts were comparable, and the emotional content from both texts were similar. On occasion, it did generate absurd sentences. For instance, in a few cases, words were repeated multiple times in a sentence.

The GPT-2 model's generated text delivered mixed results when evaluated with quantifiable metrics. Firstly, the confidence intervals of the sentence lengths of the original and generated text overlapped for 16 out of the 30 personalities, meaning that only half of the celebrity's tweets were properly generated in terms of sentence length. Next, out of the 30 generated texts for the personalities, 15 yielded sentiment scores that were similar to the original (i.e. had a difference of less than 0.1). Unfortunately, only 10 out of 30 had similar subjectivity scores compared to the original (i.e. having a difference of less than 0.1), suggesting that the GPT-2 model does not properly capture the subjectivity displayed in the celebrities' text.

Finally, although different Twitter accounts yielded different results based on the evaluation metrics, they performed similarly in terms of diction and word choice. The word clouds for most of the original text corpuses are dominated by 2-3 words in large font, with a range of other words taking up a much smaller space in the diagram. However, for the generated text, the words that dominate the word clouds rarely match those of the original text. The word clouds illustrate that although the sentences generated by the GPT-2 model appear to be similar to the original text, they do not perfectly encapsulate the personalities' language and vocabulary.

The text generation of US president Joe Biden can serve as an example to illustrate the performance level of the GPT-2 model. From a human centric perspective, the text generated by the model resembles that of the original twitter text. Sentences generated by the GPT-2 model



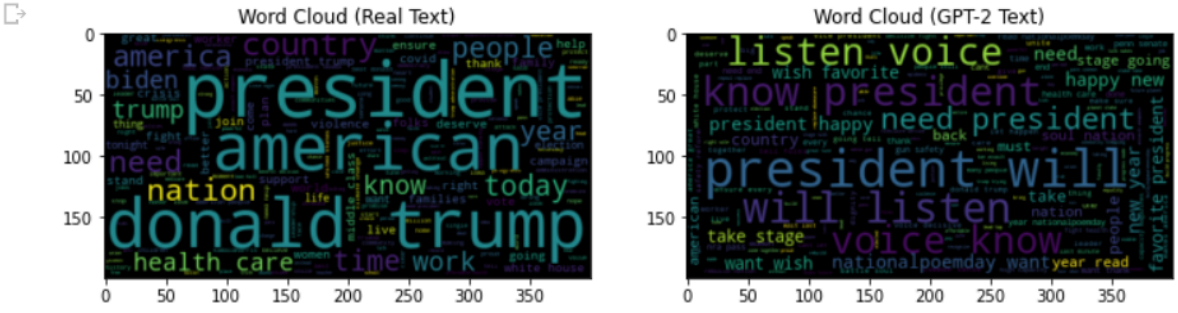
with Joe Biden's data include: "we need a president who will unite us and lead us through this moment", "as i take the stage, i'm going to tell folks that there is no place for racism in america", and "every american deserves to have the opportunity to vote". Although these sentences seem convincing and could even pass off as the original, the evaluation metrics tell a different story. Firstly, the confidence intervals of the average sentence length of the original and the generated text overlap, meaning that the model effectively replicates Biden's text in terms of sentence length. However, the sentiment and subjectivity scores of the real and generated texts deviate by more than the established 0.1 threshold. Finally, as can be seen from the word clouds shown below, the GPT-2 model does not accurately capture the word choice of Joe Biden.

Figure 1: The word clouds of the real and generated text of Joe Biden

Overall, despite the generated text of Joe Biden being convincing on eye level, the evaluation metrics suggest that it is not at a level fitting for real world application. The evaluation of the text generated with other celebrities' tweets lead to the same conclusion. Evidently, capturing the personality of an individual through their text will require more in depth research and experimentation.

Recommendations

The study proves that realizing a personalized chatbot will be a daunting feat. The simple task of generating text with no syntax, spelling, or grammatical errors will pose a challenge. Making the generated text resemble the style of the original will introduce additional complexities that will in turn give rise to more obstacles. Ultimately, there is a massive disparity between developing a bot that returns pre-written responses and a bot that returns creative responses with context that aligns with the personality they are trying to portray. Such an

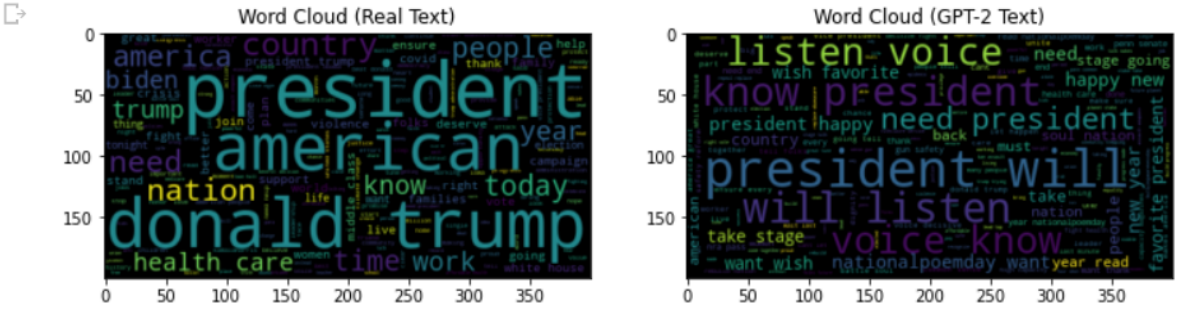


endeavor will require two crucial elements: a massive amount of training data and a deep and thorough experimentation procedure.

The recurrent neural network models and the GPT-2 model in this study were trained with text data consisting of only around 1000-7000 tweets per celebrity. While a data sample of this size will enable the models to generate sentences with some cohesion, it will not train the model to perfectly replicate the writing style and patterns exhibited in the original text. Factors like word choice and sentence length can not be properly accounted for by the generated text without sufficient data. Increasing the number of tweets used to train the model should yield better results.

In addition to a larger training dataset, the research on personalized bots can also be aided by more thorough experimental trials. Given the timeframe of the project as well as the time and computational demands that come with training these models, only a limited number of architectures and combinations of parameters could be tested. Only a few parameters (e.g. number of layers) and hyperparameters (e.g. activation function) were examined when optimizing the model. Thus, even though the LSTM models in the study failed to show promise, it is unreasonable to conclude that such an architecture is unsuitable for NLG. Furthermore, the GPT-2 model, despite its success with generating sentences resembling the original text, also was not perfectly optimized, with a few parameters (i.e. temperature, top_k) being tested. For future reference, it would be beneficial to experiment with the GPT-2 model and other pretrained models to a greater extent to see if the issues it faced during this project could be resolved. Other architectures not used in this project, such as GAN, should also be explored.

Finally, it is important to implement a more robust evaluation of the generated text. As human judgement is essential for gauging the sentences made by the bot, a large number of people should examine and adjudicate the generated text for a more reliable analysis. Furthermore, the peoples' evaluation of the text should be recorded in surveys, in which key elements of the text like grammar and vocabulary are given a numerical rating. Future analysts



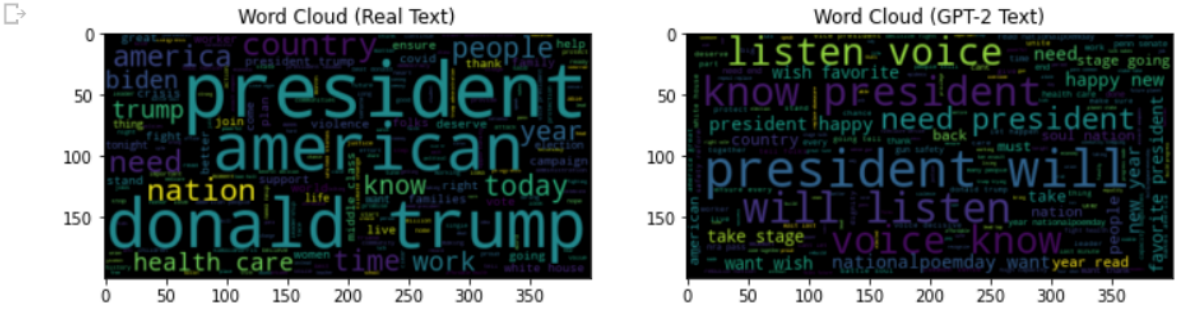
building text generation models should also entertain the idea of subjecting the people to a Turing test. For this test, people who are shown texts would have to decide if the text originated from a celebrity's tweet or if it was generated by a model. Being unable to distinguish between real and fake text would be a sign of a successful text generation model.

Conclusion

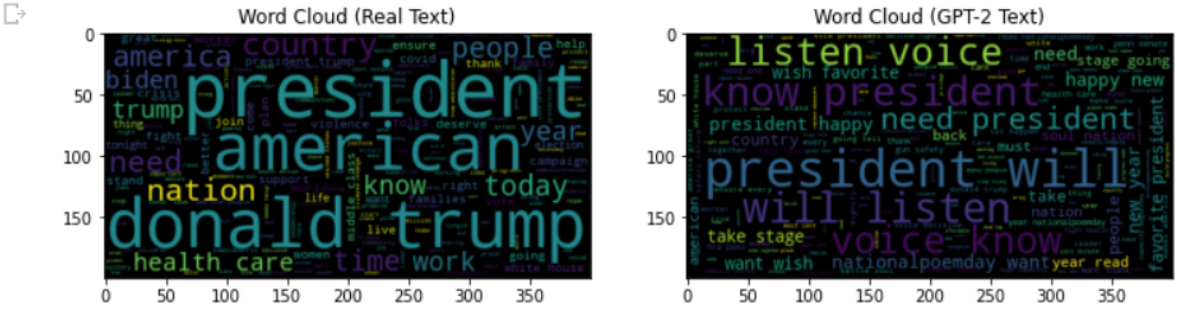
To conclude, the GPT-2 model shows most promise with regards to generating personalized text. The model created sentences that, on surface level, matched the text that it was trained with in terms of fluency and cohesion. However, in terms of quantifiable metrics, it returned mixed results. The sentence length of the generated text matched that of the original for only 16 out of the 30 celebrities. Furthermore, the sentiment and subjectivity scores of the generated text came close to that of the original for only 15 and 10 celebrities, respectively. With such results, it is difficult to currently build and execute a personalized chatbot. For future reference, this experiment should be rerun at a much larger scale. This entails collecting an exorbitant amount of text data, training text generation models with different architectures and parameters, and evaluating them with the help of a third party. Constructing a model with the ability to personalize text will expand the realm of possibilities of businesses that seek to incorporate AI into their business models to streamline their processes. Accomplishing this feat can signify a big milestone in the current rise of artificial intelligence.

Biography

Aashish Nair is a graduate student pursuing a master's degree in data science at George Washington University. His interests lie in machine learning and natural language processing. He



fervently monitors every new milestone or breakthrough made in the artificial industry. Outside of his academic and career related pursuits, he spends his free time watching sports and cooking.



References

Celikyilmaz, A. (2020, June 26). Evaluation of Text Generation: A Survey. arXiv. https://arxiv.org/pdf/2006.14799.pdf

Intelligence, I. (2021, February 8). *Chatbot market in 2021: Stats, trends, and companies in the growing AI chatbot industry*. Business Insider. https://www.businessinsider.com/chatbot-market-stats-trends.

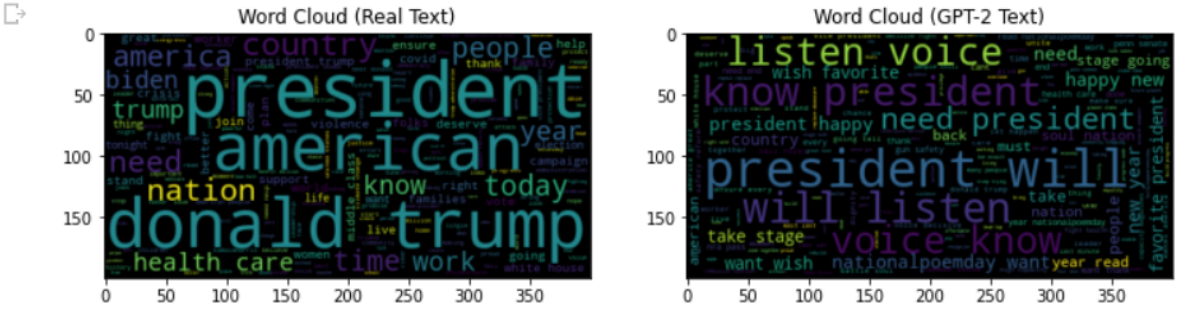
Li, Y. (2018, March 22). A Generative Model for category text generation. Elsevier. https://faculty.ist.psu.edu/szw494/publications/CS_GAN.pdf

Natural Language Generation and Its Business Applications. Skim AI. (2020, December 31). https://skimai.com/natural-language-generation-business-applications/.

Radford, A. (2020, September 3). *Better Language Models and Their Implications*. OpenAI. https://openai.com/blog/better-language-models/.

Sudhakar, M. (2021, January 19). *Council Post: Chatbots: The Great Evolution To Conversational AI*. Forbes. https://www.forbes.com/sites/forbestechcouncil/2021/01/20/chatbots-the-great-evolution-to-conversational-ai/?sh=57cc54464d0e.

Singh, S. (2020, December 23). *OpenAI's GPT-2: Building GPT-2 AI Text Generator in Python*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2019/07/openai-gpt2-text-generator-python/



Tam, S. (2020, May 7). *Tweet generation with Neural Networks: LSTM and GPT-2*. Medium. https://towardsdatascience.com/tweet-generation-with-neural-networks-lstm-and-gpt-2-e163bfd3fbd8.

Woolf, M. (2020, January 16). *How to Build a Twitter Text-Generating AI Bot With GPT-2*. Max Woolf's Blog. https://minimaxir.com/2020/01/twitter-gpt2-bot/.