Analyzing COVID-19 Tweets with Transformer-based Language Models

Philip Feldman[1, 2] **Sim Tiwari** [2] **Charissa S. L. Cheah** [2] **James R. Foulds** [2] **Shimei Pan** [2]
[1] ASRC Federal, Beltsville, Maryland, USA
[2] University of Maryland Baltimore County, Baltimore, Maryland USA
[1]philip.feldman@asrcfederal.com, [2]{simt1, ccheah, jfoulds, shimei}@umbc.edu

Abstract

This paper describes a method for using Transformer-based Language Models (TLMs) to understand public opinion from social media posts. In this approach, we train a set of GPT models on several COVID-19 tweet corpora that reflect populations of users with distinctive views. We then use prompt-based queries to probe these models to reveal insights into the biases and opinions of the users. We demonstrate how this approach can be used to produce results which resemble polling the public on diverse social, political and public health issues. The results on the COVID-19 tweet data show that transformer language models are promising tools that can help us understand public opinions on social media at scale.

Introduction

Large-scale research based on feedback from humans is difficult, and often relies on labor-intensive mechanisms such as polling, where statistically representative populations will be surveyed using phone interviews, web surveys, and mixed-mode techniques. Often, for longitudinal studies, participants in a survey may need to be recontacted to update responses as a result of changing environments and events (Fowler Jr 2013).

As social media has become ubiquitous, many attempts have been made to determine public opinion by mining large amount of social media data, often spanning years, which is available from online providers such as Twitter, Facebook and Reddit, e.g. (Colleoni, Rozza, and Arvidsson 2014; Sloan et al. 2015). Though social data can be mined in a variety of ways, answers to specific questions frequently can not be obtained without expensive manual coding.

This may be ready to change with the emergence of large transformer-based language models (TLMs) like the GPT series (Radford et al. 2018) and BERT (Devlin et al. 2018). These models are trained on massive text datasets such as BookCorpus, WebText and Wikipedia. They implement a transformer-based deep neural network architecture which uses attention to allow the model to selectively focus on the segments of the input text that are most useful in predicting target word tokens. A pre-trained GPT model can be used for generating texts as a function of the model and a sequence of word tokens provided by users. We call the sequence of words provided by users a "prompt" or a "probe", which is specifically designed to set up the theme/context for GPT to generate sentences. Since the model is not trained using any hand-crafted language rules, it effectively learns to generate natural language by observing a large amount of text data. In doing so, it captures semantic, syntactic, discourse and even pragmatic regularities in language. GPT models were shown to generate text outputs often indistinguishable from that of humans (Floridi and Chiriatti 2020).

As such, these models contain tremendous amounts of information that can be used to answer questions about the content and knowledge encoded in the training text. Unfortunately, the knowledge captured in these TLMs is latent (e.g., millions of neural network model parameters), can be difficult to interpret. In this study, by using carefully constructed prompts/probes, we "poll" the model to gain access to the latent knowledge contained in the model in ways that are analogous to accessing the knowledge in a population by surveying a random sample of individuals.

We can illustrate this process against a known ground truth (the spatial relationships between countries) by polling the GPT-3[1] repeatedly with the prompt "A short list of countries that are nearest to ___, separated by commas:". The prompt is initially seeded with a value such as "United States". Recursive responses create a graph, and a force-directed layout approximately reconstructs the original country relationships[2] (Figure 1).

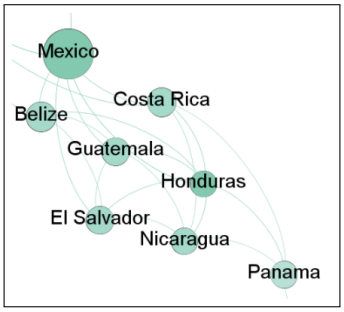

Figure 1: Central America reconstruction.

Copyright © 2021, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1]GPT-3 accessed using the OpenAI beta API
[2]Full map at tinyurl.com/gptworldmap

Using this approach, we can create "maps" that approximate the local spatial relationships of $\approx$ 75% of current countries. The presence or absence of a country is correlated with population. This appears to be a bias in the model.

In this paper, we employ a similar approach that combines a repeated-prompt based query method with data analysis and visualization to analyze the beliefs, biases and opinions of Twitter users in the context of the COVID-19 pandemic. We describe our method for GPT model training and fine tuning, determining probes, and analyzing and visualizing the output from the model. Lastly, we discuss its limitations and implications for this approach to computational sociology.

Related Work

Since the introduction of the transformer model in 2017, TLMs have become a field of study in themselves. The transformer uses self attention, where the model computes its own representation of its input and output (Vaswani et al. 2017). So far, significant research has been in increasing the performance of these models, particularly as these systems scale into the billions of parameters, e.g. (Radford et al. 2019). Among them, BERT (Devlin et al. 2018) and GPT (Radford et al. 2018) are two of the most well known TLMs used widely in boosting the performance of diverse NLP applications.

Understanding how and what kind of knowledge is stored in all those parameters is becoming a sub-field in the study of TLMs. Among them, (Petroni et al. 2019) used probes that present a query to the mode as a cloze statement, where the model fills in a blank (e.g. "Twinkle twinkle _____ star"). Research is also being done on the creation of effective prompts. Published results show that mining-based and paraphrasing approaches can increase effectiveness in masked BERT prompts over manually created prompts (Jiang et al. 2020). In another study using GPT models fine-tuned on descriptions of chess games, it was shown that models trained on a corpora of approximately 23,000 chess games accurately replicated human gameplay patterns (Feldman 2020).

Using TLMs to evaluate social data is still nascent. A study by (Palakodety, KhudaBukhsh, and Carbonell 2020) used BERT fine tuned on YouTube comments to gain insight into community perception of the 2019 Indian election.

Lastly, we cannot ignore the potential weaponization of TLMs. OpenAI has shown that the GPT-3 can be "primed" using "few-shot learning" (Brown et al. 2020). (McGuffie and Newhouse 2020) primed the GPT-3 using mass-shooter manifestos with chilling results.

Dataset

In this study, we investigated the feasibility of employing GPT to analyze COVID-19 related tweets. Using the Twitter Search API, we collected tweets from the USA if they included at least one of the identified keywords/hashtags. As of this writing, the list of case-insensitive keywords/hashtags include "coronavirus", "covid19", "sars-cov-2", "pandemic", "chinavirus", "social distancing", "mask", "ventilator", "shelter in place" etc. Terms are reevaluated and updated monthly.

So far, we have retrieved a total of 18,703,707 tweets from Nov. 2019 to the time of this writing. For this research, we constructed three datasets to train three separate GPT models based on three COVID-19 keywords: *covid, sars-cov-2*, and *chinavirus*. The most common term by Twitter users to refer to the disease, was *covid*, and was thought to represent the perspectives of the general public; *sars-cov-2* was chosen for its greater use in science-related contexts; while *chinavirus* was chosen for its arguably racist connotations.

Table 1 shows the size of each dataset. As one would expect from its more common usage, we collected approximately 350 times more tweets associated with the *covid* than the other two tags. This would have some implications in the behavior of the trained models, which we will describe in the fine-tuning section. Please note the three datasets may be overlapping (e.g., a tweet tagged with both *chinavirus* and *covid*). There are also many pandemic tweets not in any of the datasets (e.g., those tagged with *social distance*.)

Table 1: Tweet counts in each dataset

chinavirus	covid	sars-cov-2
14,950	7,015,582	25,768

We show three tweets from each dataset. These tweets were posted at the beginning of April 2020, when the pandemic's first wave was starting to happen in the USA.

chinavirus dataset: *Just to remind everyone why social healthcare is terrible...the UK has a 7% mortality rate of #ChinaVirus #WuhanVirus #coronavirus while America is currently hovering around 1.7% of reported cases*

covid dataset: *This can't be happening. These are our first responders! If they aren't protected our entire country is at risk. These doctors and nurses are on the front lines risking their life's. With a RAIN PANCHO?! #coronavirus #COVID #COVID2019 https://t.co/KWPVGuIgPW*

sars-cov-2 dataset: *Currently, incubation period of coronavirus, SARS-CoV-2, is considered to be 14 days. Health authorities across countries are adopting a 14-day quarantine period, based @WHO guidelines.*

These samples suggest qualitatively differing perspectives: the *chinavirus* appears reactionary and nationalistic; the *covid* tweet is more emotional; while the *sars-cov-2* tweet is detailed and explicit.

Methods

In this section, we describe how we customize a pre-trained GPT model with COVID-19 tweets and how we design prompts to setup the context and probe the models to reveal the answers to our questions.

Model Training and Fine Tuning

For this research, we employ the pre-trained GPT-2 model with 12-layers, 768-hidden, 12-heads, 117M parameters hosted by Huggingface.com[3]. We then fine tune the pre-trained GPT-2 model using the three COVID-19 tweet datasets described above to produce three GPT models, one for each dataset. We did not use the latest GPT-3 model as it does not allow us to retrain it with new data.

To demonstrate the behavior of these models, we use the beginning of the same tweets that are shown in the previous section as the prompts to generated three synthetic tweets. In each generated tweet, the prompt is shown in the brackets, and the generated content is shown in *italics*:

> **Chinavirus model:** [Just to remind everyone why social healthcare] *doesn't work. Stop being afraid. @realDonaldTrump @FLOTUS https://t.co/lFh9R1KdU*
>
> **covid model:** [This can't be happening. These are our first responders!] *Please keep them in your thoughts and prayers! #NursesCOVID19 https://t.co/K8xnCfGfjC*
>
> **Sars-cov-2 model:** [Currently, incubation period of coronavirus, SARS-CoV-2, is considered to be 14 days.] *The duration of incubation is 12-14 days and the virus can persist for up to 18 hours (depending on temperature and humidity).*

We can see that each model has captured not only the format of tweets (e.g., using mentions such as @realDonaldTrump, @FLOTUS or short URLs such as https://t.co/lFh9R1KdU) but also subjective and semantic aspects of the original training data. As we can see, the models are not *reconstructing* the tweets. Rather, they are stochastically generating new text based on the prompts and latent knowledge in the model. This property allows us to obtain responses to our queries which may reflect the beliefs typical of the set of twitter users in the training data.

Prompt design

Prompts can be regarded as a form of conditioning or training, as discussed by (Brown et al. 2020). The length and content of the probe can be structured so that a desired response is more likely.

Finding probes that resulted in useful responses was an interactive process, where we would develop ideas for probes and try them out on the models in small batches. For example, the probe "Dr. Fauci is" allows the model to produce adjectives ("likable") adverbs ("very"), determiners ("a"), and verbs ("running") as the next token. Changing the probe to "Dr. Fauci is a" constrains the next word to more likely to be an adjective or a noun. If we use the next nouns or adjectives after the prompts as the responses to our inquiries, the probe "Dr. Fauci is a" may produce more direct answers.

Example output from the *covid* model is shown in Table 2.

For these relatively small models, we found that shorter probes, typically 3 - 7 words would produce useful results.

[3]huggingface.co/gpt2

Table 2: Similar probes and different GPT outputs. We bold face the words if the first nouns are extracted as the answers

Dr. Fauci is	Dr. Fauci is a
out of the **spotlight**	**hero** in the war against COVID
at it again	dangerous **man**.
100% correct	medical **genius**
on top of **everything**	**liar**. It was never about COVID19

By varying the prompts, we could tune the results to explore the latent knowledge captured by GPT.

Finally, we designed a set of prompts to probe the GPT model to reveal the Twitter public's opinion/attitude toward various COVID-19 pandemic-related topics such as: whom to blame for the pandemic? How do people feel during the pandemic? Is there any systematic bias towards certain demographic groups? For example, the prompt "During the pandemic, [xxx] have been feeling" would be filled in with the terms [Americans, Asians, Black folks] to compare the feelings of different people during the pandemic.

Response/Answer Extraction

There are multiple ways to extract answers to different queries from the GPT model. First, directly from the model. Transformer language models such as the GPT generate the next token in a sequence based on a probabilistic distribution computed by the model. We can directly use the output probability of a word given a prompt as the probability related to an answer. For example, given the prompt "Dr. Fauci is a", the model can directly output the probability of the word "scientist" or "liar" appearing as the next word. Second, extracting answers based on output sample analysis. For each prompt, we can generate a large number of representative tweets using the model. We can then compute statistics based on the generated samples. In this preliminary study, we adopted the second approach. We will explore the first approach in future work.

Specifically, for each prompt, each model generated 1,000 sample responses. Each response was stored in a database, along with parts-of-speech tags and a sentiment label (either positive or negative) automatically assigned by a sentiment analyzer (Akbik et al. 2019). Statistics were computed from the samples, focusing on the relationships between probes and models. Initial statistics were gathered on the first words directly after each prompt in each response as they were generated directly from the given probe, and are most likely to vary with and impacted by the probe. Since next word analysis is likely to produce high probable functional words that do not carry specific meaning such as "the" and "has", parts-of-speech tags were used to extract the first noun ("virus") or noun-noun combination ("chinese virus"), or adjective ("anxious"). Lastly, we computed the percentage of positive tweets in the 1000 samples generated per prompt per model.

One might ask why should we analyze synthetic tweets generated by GPT instead of the real tweets directly? For some insight, we can look at Table 3 which shows the num-

ber of times each probe appears exactly in the 18 million real tweets we retrieved:

Table 3: Probe frequency in 18 million COVID tweets

Probe	Count
Donald Trump is a	1,423
Dr. Fauci is a	427
The pandemic was caused by	6
For the pandemic, we blame	0

Though there are enough results to perform statistical analysis related to Donald Trump and Dr. Fauci, the data is not sufficient to support an analysis with statistical significance for the COVID-19 causes and blames. Because transformer language models such as the GPT create tweets based on statistically valid patterns (Feldman 2020), each synthetic tweet may encode information derived from a large number of real tweets that are syntactically and semantically related. Further, in principle, the models can generate unlimited datasets to facilitate robust data analysis.

Preliminary Results

In this section, we present results that explore biases in the tweet corpora that each model was trained on.

Polling the General Public on Twitter

The *covid* model was trained on the most data and more representative of mainstream Twitter users' opinion/attitude. As such, it was more able to provide more granular responses to our probes. Our first set of results will focus on the behavior of the *covid* model.

PROBE: "The pandemic was caused by" Based on the normalized count (i.e. percentage) of the first nouns to appear after this probe in the 1000 generated samples, we list the top ranked nouns in Table 4.

Table 4: "The pandemic was caused by" top responses

Cause	Percentage
Coronavirus	31%
Virus	11%
China	6%
Lab	2%
Accident	2%

The vast majority of the responses (31%) attribute the coronavirus explicitly, while an additional 11% referred to viruses in general. Typical tweets generated by the model after the prompt "The pandemic was caused by" include *a novel* ***coronavirus*** *known as SARS-CoV-2* and *a* ***virus*** *known as SARS-CoV-2*. Variations of these statements make up over half of the generated results ranging from fatalistic, *a* ***virus*** *that could not have been contained* to conspiratorial, *a novel* ***coronavirus*** *that originated in a lab in Wuhan*.

The next values appear to be more focused on human causes. For example, "caused by" ***China****, which had unleashed COVID19 on the world and is responsible*. Further down this list align with conspiracy theories, where "the pandemic was caused by" *a* ***lab*** *accident. You can bet there were dozens of other deaths*, and "caused by" *a 'genetic* ***accident****' of the genes of aborted babies*.

PROBES: Blame for the pandemic To see if the models would distinguish between *cause* and *blame*, we tried the probe "For the pandemic, we blame". The most common response was to blame President Trump (Table 5). Responses

Table 5: "For the pandemic, we blame" top responses

Cause	Percentage
Trump	35%
China	13.5%
COVID-19	5.7%
Media	5.3%
Government	2.6%

such as this one are common: "For the pandemic, we blame" ***Trump*** *for the catastrophic response*. Tweets that blamed the media often blamed the government as well: "we blame" *the* ***Media*** *and* ***Government*** *for not telling the truth*.

PROBES: How distinct groups are feeling To extract the public opinions about ethnic groups, we ran three probes, each of which began with "During the pandemic,", and finished with: 1) "Asians have been feeling" 2) "Black folk have been feeling" 3) "Americans have been feeling."

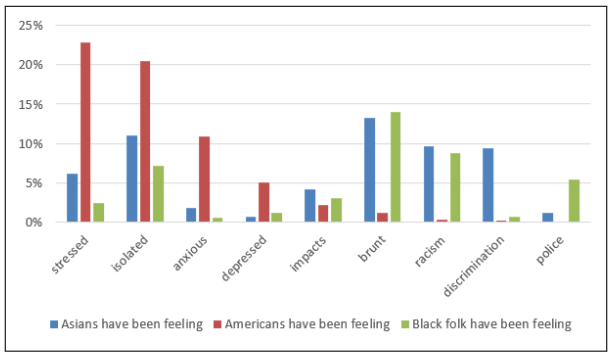

Figure 2: Asian/Black/American feeling, top responses. X-axis is sorted by the normalized frequency (percentage) of each word appearing in the tweets generated by the *covid* model with *Americans have been feeling* as the prompt.

The top results are shown in Figures 2. Common among the Asian and Black groups is the term "brunt". In the Asian results, tweets like "During the pandemic, Asian Americans have been feeling *the* ***brunt*** *of discrimination and harassment.*" are common. Alternatively, Black results emphasize "feeling *the* ***brunt*** *of racism, discrimination, and police brutality*". All groups have substantial numbers of responses that refer to isolation, with output like "have been feeling

isolated**, lonely & disconnected.*" However, the dominant term in the American responses is "stressed"[4]. For example, "Americans have been feeling *a lot more **stressed** and anxious about a new normal*", and "anxious**, stressed, hopeless, and depressed*". These generated texts indicate that the model is presenting a more subjective feeling set of responses for Americans in general, while ethnic sub-groups are feeling the brunt of external forces.

Please note that the above results are about "how groups feel" based on the Twitter general public. To poll the feeling of each ethnic group directly, we would need to use a prompt like "I am an Asian. I have been feeling".

Polling Different Populations

In addition to the "covid" model, we created models trained on tweets containing the "chinavirus" and "sars-cov-2" tags. In this section, we compare the outputs across the three models, similar to polling to different sub-populations on Twitter. Each model generated 1,000 synthetic tweets for each probe, allowing for direct comparison between the generated data.

Donald Trump and Dr. Fauci One of the most polarizing prompts that we found was "Dr. Fauci is a"[5]. This created distinct sets of responses for each of the models, as seen in Figure 3. The chinavirus model produced tweets such as "Dr. Fauci is a ***liar** and a demagogue #ChinaVirus*." Sorting by term frequency based on the outputs of this model produces an opposed trend in the sars-cov-2 model. Linear regression on each model's term frequency clearly shows this interaction. The dominant terms produced by the sars-cov-2 model for this prompt are *professor, scientist*, and *physician*. The generated content uses a more informational style, such as "Dr. Fauci is a ***professor** and **physician**. He authored and co-authored several papers published on SARS-CoV-2*"

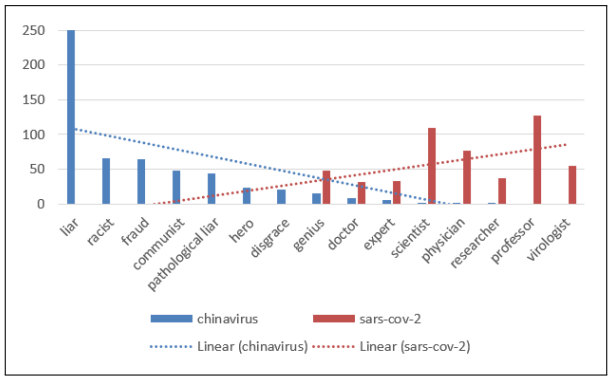

Figure 3: Fauci nouns extracted from different models, where the x-axis is sorted by their frequency computed from the Twitter samples synthesized by the "chinavirus" model.

To examine the differences in emotional tone that these models produced with the "Dr. Fauci is a" probe, We ran an additional probe, "Donald Trump is a", and compared the sentiment of the tweets synthesized from each probe across all three models, using an existing sentiment analyzer (Akbik et al. 2019). This is shown in Table 6:

Table 6: Trump / Fauci Positive Sentiment

Probe	chinavirus	covid	sars-cov-2
Dr. Fauci is a	13.3%	33.1%	53.6%
Donald Trump is a	44.4%	28.4%	27.1%

We see a similar pattern to that seen in Figure 3. In the response to the "Dr. Fauci is a" probe, the chinavirus model generates only 13% positive responses, while it generates approximately 45% positive text in response to "Donald Trump is a" (e.g. here is one such tweet produced by the model: "**Donald Trump is a** *great politician and a man of integrity*."). The covid model falls between the other two models, particularly with respect to the Dr. Fauci probe. It is not significantly different from the behavior of the sars-cov-2 model in response to the Donald Trump probe.

Polling Over Time

The proposed method can also be used to poll opinions at different times. We can have two different ways to poll the model over time. First, fine tune each model with time sensitive data (e.g., fine tune the model with pre-pandemic versus during and post pandemic data); Second, we may use time-sensitive prompts such as "in March, 2020, Americans have been feeling". In this study, we perform a coarse pre- and during-pandemic analysis. We therefore used the first method.

As shown in Figure 4, we summarize the public sentiment towards different demographic groups before and after the pandemic. The GPT2-large model on the left was trained on general web data before the pandemic while the other three models were fine-tuned with the twitter covid data we collected during the entire course of the pandemic to date. The probes used in the analysis include "[xxx] are like a" where [xxx] can be Americans, Hispanics, Asians, Blacks, Jews and Chinese. We use the GPT's responses to these metaphors to assess the public sentiment towards different demographic groups.

As shown in the chart, the sentiment in general is much more positive before (outputs from 2019 GPT2-large) than during the pandemic (outputs from all three pandemic-related models). This is true across all demographic groups we considered in the experiment. Moreover, before the pandemic, the sentiment towards "Americans" is the most positive while that towards "Blacks" is the most negative. During the pandemic, the sentiment towards Chinese has turned decisively negative. This is true across all three pandemic models. Inspecting tweets generated by different models towards the Chinese, the GPT2-large model generates tweets like "**We think Chinese are like** *a lot of other groups - very loyal to their own, have great energy*". However, during the pandemic, the chinavirus model generates tweets like "**We think Chinese are like** *a snake eating bats in a cauldron. #ChinaVirus*".

[4]for this analysis the counts for "stressed", "stress", and "strain" were combined

[5]Thanks to Dr. Peterson of Montgomery College for the suggestion

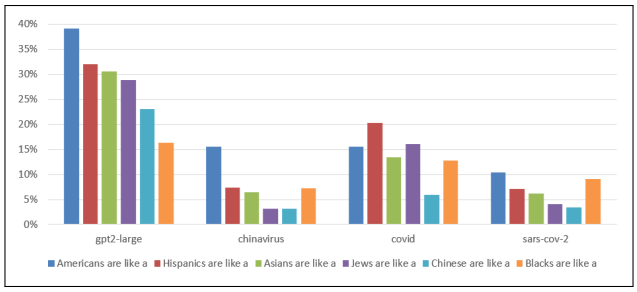

Figure 4: Public sentiment towards different demographic groups before and during the pandemic. The y-axis is the percentage of positive sentiment associated with the samples generated with each prompt by each model.

Discussion and Future Work

Polling transformer language models have provided us with a new lens to assess public attitude/opinions towards diverse social, political and public health issues. It is dynamic and can be used to answer diverse questions. It is computationally inexpensive and does not require any costly human annotated ground truth to train. It also can be used to support longitudinally studies via either prompt design (e.g., using a prompt like "in January 2020") or model tuning with time-appropriate data.

Polling transformer language models is very different from real polling. For example, results from GPT, particularly the small models like *chinavirus* and *sars-cov-2* are noisy. In particular, individual tweets stochastically generated by GPT may be incorrect. It is important that we rely on statistical patterns rather than individual tweets synthesized by these models to draw conclusions. In addition, prompt design is tricky. Small changes in prompts may result significant changes in results (e.g., "Dr. Fauci is a" verus "Dr. Fauci is"). Limitations of the TLMs themselves may also prevent them from providing accurate information. For example, although humans can link affordances (*I can walk inside my house*) and properties to recover information that is often left unsaid (*the house is larger than me*), TLMs struggle on such tasks (Forbes, Holtzman, and Choi 2019). TLMs are also vulnerable to *negated* and *misprimed* probes.

So far, we have only scratched the surface trying to probe and understand the latent knowledge captured in a transformer language model. To further this research, we plan to (a) develop a systematic approach for prompt design based on a deeper understanding of the relationships between prompts and responses as well as between prompts and context, (b) infer word/phrase-based probability directly based on the token probability generated by the GPT, (3) improve the NLP techniques used to extract answers from synthesized tweets. In this preliminary study, we employed very simple techniques such as extracting the first nouns or adjectives, or using existing sentiment tools. With more sophisticated syntactic analysis, we can extract more meaningful answers from model responses.

References

Akbik, A.; Bergmann, T.; Blythe, D.; Rasul, K.; Schweter, S.; and Vollgraf, R. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 54–59.

Brown, T. B.; et al. 2020. Language Models are Few-Shot Learners.

Colleoni, E.; Rozza, A.; and Arvidsson, A. 2014. Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *Journal of communication* 64(2): 317–332.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Feldman, P. 2020. Navigating Language Models with Synthetic Agents. *arXiv preprint arXiv:2008.04162* .

Floridi, L.; and Chiriatti, M. 2020. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines* 30(4): 681–694.

Forbes, M.; Holtzman, A.; and Choi, Y. 2019. Do Neural Language Representations Learn Physical Commonsense? *arXiv preprint arXiv:1908.02899* .

Fowler Jr, F. J. 2013. *Survey research methods*. Sage publications.

Jiang, Z.; Xu, F. F.; Araki, J.; and Neubig, G. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics* 8: 423–438.

McGuffie, K.; and Newhouse, A. 2020. The radicalization risks of GPT-3 and advanced neural language models. *arXiv preprint arXiv:2009.06807* .

Palakodety, S.; KhudaBukhsh, A. R.; and Carbonell, J. G. 2020. Mining insights from large-scale corpora using fine-tuned language models .

Petroni, F.; Rocktäschel, T.; Lewis, P.; Bakhtin, A.; Wu, Y.; Miller, A. H.; and Riedel, S. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066* .

Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. Improving language understanding by generative pre-training .

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1(8).

Sloan, L.; Morgan, J.; Burnap, P.; and Williams, M. 2015. Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data. *PloS one* 10(3): e0115545.

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

This paper is based upon work supported by the National Science Foundation under grant no. 2024124.