Should we tweet this? Generative response modeling for predicting reception of public health messaging on Twitter

Abraham Sanders  
Rensselaer Polytechnic Institute  
Troy, New York, USA  
sandea5@rpi.edu

John S. Erickson  
Rensselaer Polytechnic Institute  
Troy, New York, USA  
erickj4@rpi.edu

Debmani Ray-Majumder  
Rensselaer Polytechnic Institute  
Troy, New York, USA  
raymad@rpi.edu

Kristin P. Bennett  
Rensselaer Polytechnic Institute  
Troy, New York, USA  
bennek@rpi.edu

ABSTRACT

The way people respond to messaging from public health organizations on social media can provide insight into public perceptions on critical health issues, especially during a global crisis such as COVID-19. It could be valuable for high-impact organizations such as the US Centers for Disease Control and Prevention (CDC) or the World Health Organization (WHO) to understand how these perceptions impact reception of messaging on health policy recommendations. We collect two datasets of public health messages and their responses from Twitter relating to COVID-19 and Vaccines, and introduce a predictive method which can be used to explore the potential reception of such messages. Specifically, we harness a generative model (GPT-2) to directly predict plausible future responses and demonstrate how it can be used to optimize expected reception of important health guidance. Finally, we introduce a novel evaluation scheme with extensive statistical testing which allows us to conclude that our models capture the semantics and sentiment found in actual public health responses.

CCS CONCEPTS

- Computing methodologies → Discourse, dialogue and pragmatics;  
- Information systems → Sentiment analysis.

KEYWORDS

tweet response generation, sentiment analysis, public health, COVID-19, vaccines

1 INTRODUCTION

During the COVID-19 pandemic, Twitter and other social media messaging by public health organizations played a significant role in their strategies to enact proposed mitigations to potential risks with varying effectiveness [23]. As such, recent works have focused on topical, semantic, and sentiment analysis of COVID-19 and vaccine related Twitter discourse, many leveraging natural language processing (NLP) technologies. For example, Sanders et al. [20] clustered tweets relating to mask-wearing in the early days of the COVID-19 pandemic to discover prevalent themes, perceptions, and sentiments. Cotfas et al. [7] applied machine learning for vaccine stance detection using tweets collected in the month following the announcement of a COVID-19 vaccine. Our study follows similar motivation - to investigate the way the general population reacts to messaging from major public health agencies (e.g., US CDC, European CDC, and WHO) on a variety of topics including COVID-19 and vaccines. Unlike previous work in this area, we investigate the feasibility and utility of using state-of-the-art text generation models to directly simulate typical response distributions to novel public health messages on Twitter. These simulations, combined with sentiment analysis, can be used to help public health organizations understand the specific opinions and concerns of their audience in order to develop more effective health messaging strategies.

In this study, we collect two datasets of public health tweets: (1) COVID-19 related public health messages from March 1st, 2020 to September 30th, 2020, and (2) vaccine-related public health messages from October 1st, 2021 to January 31st, 2022. These datasets include the original messages and samples of their responses, both in the form of direct replies and quote-tweets (retweets with comments). Using each dataset, we fine-tune a GPT-2 [16] language model to predict responses to the public health tweets and evaluate its effectiveness in terms of semantic and sentiment similarity with known responses. To evaluate the models, we establish "ground-truth" baselines for semantics and sentiment on each dataset by comparing two distinct samples of known responses to each message. We also establish "random-chance" baselines by likewise comparing each sample of known responses with a sample of random responses for any message in each dataset. We then use our models to generate responses to each test message compare them with the known response samples. Through rigorous statistical testing we find that our models are able to generate responses consistent with known samples in terms of semantics and sentiment. Thus, insights on perceptions toward particular public health issues can be gained from analyzing the generated response distributions. We envision our methods being able to aid public health decision makers and social media content managers proactively model how the public will react to future messages, increasing the likelihood that their tweets are well received and have the intended impact.

The remainder of this paper is organized as follows: (1) we present two datasets of Twitter public health messages and their responses, one related to COVID-19 and one related to Vaccines; (2) we fine-tune GPT-2 to generate responses on each of these datasets, and construct upper (ground-truth) and lower (random-chance) bound baselines against which to evaluate it; (3) we visually demonstrate the capabilities of our models using test set examples and
walk through our envisioned public health use case; (4) we perform extensive statistical testing to compare our models against the baselines, finding that GPT-2 can effectively capture semantics and sentiment in typical response distributions to messages in our test sets; and (5) we conclude with a discussion of limitations and future directions of our work including a review of related works and sentiment in typical response distributions to messages in our public health accounts.

2 DATA COLLECTION

As in [20], we used the Twitter streaming API to collect a random sample of tweets during the collection periods for each respective dataset (COVID-19 & Vaccine public health messages). We collected these datasets by filtering the streaming API using COVID-19 and Vaccine related keywords, respectively. Since we aim to study the response distributions to public health tweets, we focus only on those tweets which have responses either in quote-tweet or direct reply format. Collection of these tweets and their responses was done via Tweepy, a python library for accessing the Twitter API, and they were stored in Elasticsearch for efficient search and retrieval. Each dataset then filtered by screen name to include only tweets from public health organizations and their responses. The organizations selected and their respective accounts are shown in Table 1.

2.1 COVID-19 Public Health Messaging

Our dataset of COVID-19 related public health messages and their responses contains 8,475 original messages authored by these accounts and 70,331 responses to these messages. The original messages were authored between March 1st, 2020 and September 30th, 2020. The majority of the collected tweets originate from the WHO account, followed by CDCgov, as is the case in the COVID-19 dataset (see Figure 1). This dataset was collected by the same procedure outlined for the COVID-19 dataset in Section 2.1, with the only difference being the filter keyphrases. Here, all filter keyphrases were vaccine related, selected by doing a term-frequency analysis on a random sample of approximately 10,000 tweets collected using the keyphrase “vaccine” (see our code release for complete listing). Using these keyphrases we collected 52,282,174 original tweets before filtering for responses to the public health accounts.

3 EXPERIMENTAL SETUP

As discussed in Section 1, we train GPT-2 on the task of tweet response generation. This task is notably different from other text generation tasks in that it suffers from an extreme form of the one-to-many problem seen in dialogue response generation, where an utterance can have many equally valid responses [8, 10, 24]. Specifically, each public health message in our datasets has multiple responses, and we train GPT-2 to model the distribution of typical responses for each message. This means that the same message from the same author is repeated many times in the training set, each instance with a different target response. Once trained in this manner, temperature sampling can be used to generate a range of likely responses to an input author and message.

As previously mentioned, we evaluate this method by comparing model-generated responses to known responses. Specifically, given a known sample of responses to a particular message and author, we need to determine how well a model-generated sample of responses captures the semantics (e.g., meaning, topic, intent) and the sentiment polarity (e.g., positive, negative, neutral) of the known responses. This is akin to measuring retrieval recall - how well the model-generated response distribution “covers” that of the ground-truth. To measure sentiment we use a publicly available RoBERTa model.

(2) These tweets were filtered to keep only those that were in response to (either via quote or direct reply) a message from one of the public health accounts in Table 1; (3) As the streaming API returned quoted tweets but not (direct) replied-to tweets, these were separately requested using the Twitter Status API.

2.2 Vaccine Public Health Messaging

Our dataset of Vaccine related public health messages and their responses contains 3,060 original messages authored by the accounts in Table 1 and 61,099 responses to these messages. The original messages were authored between October 1st, 2021 and January 31st, 2022. The majority of the collected tweets originate from the WHO account, followed by CDCgov, as is the case in the COVID-19 dataset (see Figure 1). This dataset was collected by the same procedure outlined for the COVID-19 dataset in Section 2.1, with the only difference being the filter keyphrases. Here, all filter keyphrases were vaccine related, selected by doing a term-frequency analysis on a random sample of approximately 10,000 tweets collected using the keyphrase “vaccine” (see our code release for complete listing). Using these keyphrases we collected 52,282,174 original tweets before filtering for responses to the public health accounts.

| Public health account screen names | Other Public Health Accounts |
|------------------------------------|-----------------------------|
| European CDC Accounts 4            | WHO Injection/Safety        |
| ESCAIDE ETD PPH ECDC VPD ECDC HIV/AIDS ECDC Outbreaks | | |
We then fine-tune the 762 million parameter GPT-2 model\(^8\) on the response generation task. Each training example consists of a public health message, the author account’s screen name, and one response, delimited by three special tokens we add to the model’s vocabulary: (1) `<message>` to indicate the following text is the public health message; (2) `<author>` to indicate the following text is the screen name of the message author; and (3) `<response>` to indicate the following text is a response to the message. At inference time, we can randomly select two distinct samples for the ground-truth baseline.

After fine-tuning, 10% of the training set is held out as a validation set. Fine-tuning is then done with the AdamW optimizer\(^\[15\]\) with an initial learning rate of \(10^{-5}\) for a maximum of 15 epochs. Validation and checkpointing is done 4 times per epoch, and training is terminated early if three epochs elapse with no improvement in validation loss. Once training completes, the checkpoint corresponding to the lowest validation perplexity is selected as the final model.

3.3 Evaluation & Baselines

Finally, we use the test set of each dataset to establish the ground-truth and random-chance baselines which function as expected upper and lower bounds, respectively, for semantic and sentiment similarity measurements. For each message in the test set, we sample: (1) 60 known responses, and (2) 30 responses to random messages in the dataset. The 60 known responses are split into two distinct “ground-truth” sets - a Primary set and a Reference set used for establishing a baseline. Thus, for each test message we compare the 30 primary ground-truth responses with:

1. The 30 reference responses (ground-truth baseline).
2. The 30 model-generated responses (model evaluation).

### Table 2: Training & Test set statistics

|              | # of messages | # of responses | Avg. # resp. per msg. |
|--------------|---------------|----------------|-----------------------|
| **COVID-19 Public Health dataset** | | | |
| Train | 7,860 | 39,020 | 5 ± 8* |
| Test | 155 | 9,300 | 60** |
| **Vaccine Public Health dataset** | | | |
| Train | 27,900 | 27,878 | 10 ± 16* |
| Test | 149 | 8,400 | 60** |

\(*\): Mean ± standard deviation, both rounded to the nearest integer.

\(**\): Exactly 60 responses were sampled for each message in the test set.

### Table 3: GPT-2 training & inference examples

**Training Examples:**

- `<message>`: Is your child worried about #COVID19? Learn the facts so you can answer your children’s questions. Make sure to explain the simple things they can do, like washing their hands often. Learn more here: `<author>`.#CDCgov/<response>`: Great resource for parents and teachers.

- `<message>`: How children can #WearAMask properly. Read more about children & masks in relation to #COVID19 `<author>`.#WHO/`<response>`: How long back into school until kids are trading their avengers masks like pogs? `<author>`:

- `<message>`: How children can #WearAMask properly. Read more about children & masks in relation to #COVID19 `<author>`: #WHO/`<response>`: Children do not need to wear masks! Get last! `<author>`:

**Inference Example (not actual model output):**

Prompt: `<message>`: How will people respond to THIS? `<author>`: #CDCdirector `<response>`: Like this! `<author>`:

Output sample 1: And that! `<author>`:

Output sample 2: And that! `<author>`:

Output sample N: DefinitelyThis. `<author>`:

**Table 4: Model training statistics**

| Dataset            | # of train ex. | # of val ex. | # of epochs | Final val PPL. |
|--------------------|----------------|--------------|-------------|----------------|
| COVID-19           | 35,118         | 3,902        | 2           | 3.36           |
| Vaccines           | 25,090         | 2,788        | 2           | 2.82           |

\(\text{ex} = \text{examples}; \text{val} = \text{validation}; \text{PPL} = \text{perplexity}\)
Figure 2: An illustration of our model evaluation scheme for a single test message. For each test message, N=30 generated responses are compared with the primary ground-truth sample of N=30 known responses. The sentence embeddings and sentiment scores for each response are then used to perform statistical testing for semantic and sentiment similarity. The ground-truth and random-chance baselines are computed the same way, replacing the generated responses with the reference ground-truth and random response samples respectively.

(3) the 30 random responses (random-chance baseline).

Figure 2 illustrates this evaluation scheme. As shown, we use several statistical tests to measure semantic and sentiment similarity for the baselines and for model evaluation. Details of these tests and their results are provided in Section 5.

4 QUALITATIVE RESULTS

4.1 Model Output Examples

In Figure 3 we show primary ground-truth and model-generated responses for two messages from each test set (COVID-19 & Vaccines). For each message, we show the top five ground-truth responses ranked in descending order of mean cosine similarity (defined in Section 5) with the model-generated responses, and likewise we show the top five model-generated responses ranked in descending order of mean cosine similarity with the ground-truth responses. This filtering and ordering is done for the sake of brevity as it is not practical to fit all 60 × 4 responses in this document. We observe that the generated responses capture many of the same opinions and concerns present in the known responses. We summarize some of the key similarities evident in the examples:

The first example shows a test message from the COVID-19 dataset where CDCDirector recommends that schools can re-open safely. The known and generated responses both exhibit themes of mistrust toward the CDC (shown in red), allegations of bowing to the Trump administration (shown in orange), implication of shame and disgrace toward the CDC (shown in purple), concern for the well-being of school children (shown in brown), and references to loss of former respect (shown in blue). The second example shows a test message from the COVID-19 dataset where WHO calls for unity in the face of the pandemic. The known and generated responses both exhibit themes of mistrust toward the WHO (shown in red) and allegations of conspiracy with China (shown in blue). The third example shows a test message from the Vaccines dataset where CDCgov urges pregnant people and those planning to get pregnant to get vaccinated against COVID-19. The known and generated responses both exhibit themes of concern for the effects on unborn children (shown in red), concern for the vaccines getting FDA approval (shown in brown), and feelings of encouragement toward the recommendation (shown in blue). The fourth example shows a test message from Vaccines where CDCDirector discusses updating pediatric vaccine recommendations to include children 5-11 years old. The known and generated responses both exhibit feelings of gratefulness and acknowledgement of the importance of pediatric vaccination (shown in red).

4.2 Envisioned Use Case Walk-through

We anticipate that public health organizations may find value in being able to “preview” public reception to any potential message on health policy or recommendations. As such, we envision our method being incorporated as a social media insights tool that can help avoid negative receptions where possible with the aim to improve adherence to health guidance.

To demonstrate this use case, we invent a tweet encouraging vaccination against COVID-19 and we suppose it is being considered by the ECDC_EU content manager for future public release. We first use the generator model trained on the Vaccines dataset to predict a set of 30 responses. We then modify the tone of the message to reduce directness and implication of personal responsibility and use it to generate a new set of 30 responses. We use the RoBERTa sentiment classifier to score each response in both sets and compute the mean and standard deviation over the scores.
Message (CDCDirector; COVID-19 dataset):
Together, we can open schools safely. This fall, school may look different in some communities b/c of COVID19 as face masks, distancing, & hand hygiene meet bookbags & homework.
New CDC tools will help us make decisions for the upcoming school year.

Primary ground-truth responses:
@CDCDirector @CDCgov Shame on you and the @CDC. Trump canceled the @GOP convention, but it’s just fine sending our kids back to school. The curve has NOT been flattened and we are trending in the wrong direction, but send kids back to school? What has happened to the once great CDC? Shameful.
@CDCDirector @CDCgov Sad to see the once esteemed CDC become the pariah of an administration that doesn’t respect or understand data. Luckily, the American public does seem to understand death tolls.
@CDCDirector @CDCgov CDC is misleading America parents and putting children in great harm with its new guidelines for schools during this pandemic. There’s not enough data to support younger children’s are less likely to contract and spread this virus. YOU ARE PLAYING GOD.
@CDCDirector @CDCgov We the people are wondering if the @edc is taking bribes to cook the books...
@CDCDirector @CDCgov But the Republican Convention is not safe? The CDC is supposed to protect the American people and not do one person’s bidding. Remember your Hippocratic oath do no harm.

Generated responses:
@CDCDirector @CDCgov The CDC has lost its credibility. You should be removed immediately. You are not leading. You are a Trump puppet. Resign.
@CDCDirector @CDCgov How do you expect schools to be safe when you keep changing the testing recommendations? You are not a doctor, you are a politician. You have
@CDCDirector @CDCgov You should be removed from office immediately. You are a disgrace to the once respected CDC.
@CDCDirector @CDCgov Why are you putting children at risk? You say it will be safe, but you also say it will be deadly for children to be in schools. The CDC needs to be dismantled and put in place someone who can be trusted. You’re a disgrace.
@CDCDirector @CDCgov Why is your department still pushing for more testing? You are a disgrace to the once respected CDC by not standing up for science and the American people. You are complicit in thousands of deaths! Stand up to Trump and do your job!

Message (WHO; COVID-19 dataset):
"COVID19 has taken so much from us, but it has also given us a unique opportunity: -To put aside our differences -To break down barriers -To see and seek the best in each other -To lift our voices for #HealthForAll" @DrTedros during the One World, #TogetherAtHome

Primary ground-truth responses:
@WHO @DrTedros @COVID19 has taken so much from us, but it has also given us a unique opportunity -To reveal the corruption and incompetence of @WHO -To have a glimpse to the dictatorship of the CCP
@WHO @DrTedros Why not let China fund your escapade? You can’t have solidarity with people or an organization you don’t trust. #ChinaLiedPeopleDied @WHO Lied People Died
@DrTedrosWHO CCPVirus #XJlying #XiWeimethehefa
@WHO @DrTedros Keep the meaningless platitudes coming, I’m sure that’s the key to stopping this pandemic you opened the door to through your pandering to China. Bang up job.
@WHO @DrTedros Let’s be inclusive and include ! Amazing that the country EXCLUDED from the @WHO handed the Wuhan Virus the best.
@WHO @DrTedros To shill for China

Generated responses:
@WHO @DrTedros How dare you chime in on a global pandemic with the help of China in the middle of a global economic crisis. The Chinese Communist Party should be held fully accountable for this global tragedy.
@WHO @DrTedros #ChinaLiedPeopleDied China should be brought up on charges for this global pandemic. We need justice.
@WHO @DrTedros WHO & CHINA LIES & PEOPLE DIED! INVESTIGATE.
@WHO @DrTedros The Chinese are not your friend. You should be kicked out of every country of the world. You are the cause of this global crisis.
@WHO @DrTedros You must resign! You are a disgrace and a danger to the world! The whole world is suffering from your failures! China is suffering from you!

Message (CDCgov; Vaccines dataset):
URGENT. If you are pregnant, recently pregnant, or plan to get pregnant, get vaccinated against COVID19 as soon as possible. Getting vaccinated can protect you from severe illness from COVID-19, and a healthy parent is important for a healthy baby. More:

Primary ground-truth responses:
@CDCgov My pregnant wife, will not be getting vaccinated. Not enough studies on how it will affect our soon to be son!
@CDCgov Why are pregnant women and people who have already had covid vaccinated against covid sick and dying?

Generated responses:
@CDCgov How dare you chime in on a global pandemic with the help of China in the middle of a global economic crisis. The Chinese Communist Party should be held fully accountable for this global tragedy.
@CDCgov #ChinaLiedPeopleDied China should be brought up on charges for this global pandemic. We need justice.
@WHO @DrTedros WHO & CHINA LIES & PEOPLE DIED! INVESTIGATE.
@WHO @DrTedros The Chinese are not your friend. You should be kicked out of every country of the world. You are the cause of this global crisis.
@WHO @DrTedros You must resign! You are a disgrace and a danger to the world! The whole world is suffering from your failures! China is suffering from you!

Message (CDCgov; Vaccines dataset):
Together, we can open schools safely. This fall, school may look different in some communities b/c of COVID19 as face masks, distancing, & hand hygiene meet bookbags & homework.
New CDC tools will help us make decisions for the upcoming school year.

Primary ground-truth responses:
@CDCDirector @CDCgov Do not vaccinate children with this cv vaccine. They have natural immunity and the vac doesn’t work. HCG and Ivermetin does. Leave the Children alone you monsters. Remember your Hippocratic oath do no harm.
@CDCDirector @CDCgov CDC is misleading America parents and putting children in great harm with its new guidelines for schools during this pandemic. There’s not enough data to support younger children’s are less likely to contract and spread this virus. YOU ARE PLAYING GOD.
@CDCDirector @CDCgov We the people are wondering if the @edc is taking bribes to cook the books...
@CDCDirector @CDCgov But the Republican Convention is not safe? The CDC is supposed to protect the American people and not do one person’s bidding. Remember your Hippocratic oath do no harm.

Generated responses:
@CDCDirector @CDCgov The CDC has lost its credibility. You should be removed immediately. You are not leading. You are a Trump puppet. Resign.
@CDCDirector @CDCgov How do you expect schools to be safe when you keep changing the testing recommendations? You are not a doctor, you are a politician. You have
@CDCDirector @CDCgov You should be removed from office immediately. You are a disgrace to the once respected CDC.
@CDCDirector @CDCgov Why are you putting children at risk? You say it will be safe, but you also say it will be deadly for children to be in schools. The CDC needs to be dismantled and put in place someone who can be trusted. You’re a disgrace.
@CDCDirector @CDCgov Why is your department still pushing for more testing? You are a disgrace to the once respected CDC by not standing up for science and the American people. You are complicit in thousands of deaths! Stand up to Trump and do your job!

Figure 3: Example generated responses to test set messages on both datasets. Elements of each response that share similar themes are highlighted with the same color (e.g., red text in the first example indicates mistrust and red text in the last example indicates gratefulness). We observe that the generated responses capture many of the same opinions and concerns as the primary (known) responses.
We now describe in detail our statistical testing, the purpose of which is to confirm that our models capture the true semantic and sentiment distributions of known responses as we expect.

5.1 Semantic Similarity

For each test message, we aim to establish if the model generates responses that capture the semantics (e.g., meanings, topics, intents) present in the known responses. To do so, we compute the max pairwise cosine similarity between the sentence embedding of each known primary ground-truth response and those of the reference, generated, and random responses. This yields three sets of 30 max cosine similarity values for each test message - one for the ground-truth baseline, one for the model evaluation, and one for the random-chance baseline. We choose max instead of mean cosine similarity so that primary ground-truth responses will be considered “covered” by the model if at least one similar response shows up in the generated sample [10]. We then perform three statistical tests on each set to compare the model with the baselines: (1) a two-tailed paired t-test to compare the average achieved difference.

5 QUANTITATIVE RESULTS

We now describe in detail our statistical testing, the purpose of which is to confirm that our models capture the true semantic and sentiment distributions of known responses as we expect.

5.1 Semantic Similarity

For each test message, we aim to establish if the model generates responses that capture the semantics (e.g., meanings, topics, intents) present in the known responses. To do so, we compute the max pairwise cosine similarity between the sentence embedding of each known primary ground-truth response and those of the reference, generated, and random responses. This yields three sets of 30 max cosine similarity values for each test message - one for the ground-truth baseline, one for the model evaluation, and one for the random-chance baseline. We choose max instead of mean cosine similarity so that primary ground-truth responses will be considered “covered” by the model if at least one similar response shows up in the generated sample [10]. We then perform three statistical tests on each set to compare the model with the baselines: (1) a two-tailed paired t-test to compare the average achieved difference.

5.1.1 AUC-REC

We introduce the AUC-REC approach for assessing semantic similarity of the primary, reference, generated, and random response sets. Regression Error Characteristic (REC) curves generalize the principles behind the Receiver Operator Characteristic (ROC) curves to regression models [4]. The ROC curve is typically used to present the quality of a binary classification model by comparing its true-positive rate (along the y-axis) to its false-positive rate (along the x-axis). The area under the resulting curve (AUC-ROC) is a metric that summarizes the extent to which the classifier can correctly identify positive examples without mistaking negative examples as positive. The REC curve applies a similar premise to regression models: for each of an increasing series of error tolerances (along the x-axis) it shows the “accuracy” of the model within that tolerance (along the y-axis). Specifically, the accuracy is the percentage of examples for which the continuous target value can be predicted within the given error tolerance. The
area over the resulting curve approximates the total expected error of the model, and thus the area under the curve can be used to approximate model quality in the same manner as ROC curves. We use the REC curves to directly compare the ground-truth baseline (Primary vs. Reference), the model evaluation (Primary vs. Model), and the random-chance baseline (Primary vs. Random) using min cosine distance as the error metric. We construct each REC curve as follows: (1) we concatenate the sets of 30 max cosine similarity scores for each of M test messages, yielding a single list of cosine similarities for all M × 30 primary ground-truth responses (e.g., for the COVID-19 dataset, this yields 155 × 30 = 4,650 max cosine similarities); (2) we normalize the resulting list so that the highest score is 1; and (3) we subtract all values in the list from 1 to convert them to cosine distances. All three resulting lists (one for the model evaluation and two for the baselines) are then used to construct the REC curves and AUC values as described in [4]. Figure 5 shows the curves with corresponding AUC measurements for the model and baselines on both datasets. In Table 5 we report the REC-AUC scores (e.g., on the full COVID-19 test set we have Model % Difference = 71.7% for AUC-REC and 70.8% for the paired t-tests). We observe an absolute difference of less than 1% between the Model % Differences obtained for the paired t-tests and those obtained for the AUC-REC scores (e.g., on the full COVID-19 test set we have Model % Difference = 71.7% for AUC-REC and 70.8% for the paired t-tests, and on the full Vaccines test set we have Model % Difference = 66.7% for AUC-REC and 67.6% for the paired t-tests). This provides confirmation for the conclusions drawn from the AUC-REC results; that is, that the model can meaningfully capture and reproduce response semantics for test messages and authors.

Table 5: Semantic Similarity: AUC-REC Results

| Comparison             | ALL   | WHO   | CDCgov | CDCDirector |
|-----------------------|-------|-------|--------|-------------|
| Primary vs. Reference | 0.571 | 0.565 | 0.558  | 0.610       |
| Primary vs. Model     | 0.539 | 0.517 | 0.544  | 0.595       |
| Primary vs. Random    | 0.458 | 0.442 | 0.466  | 0.500       |
| Model % Difference*   | 71.7% | 61.0% | 84.8%  | 86.4%       |

Vaccines dataset

| Comparison             | ALL   | WHO   | CDCgov | CDCDirector |
|-----------------------|-------|-------|--------|-------------|
| Primary vs. Reference | 0.616 | 0.653 | 0.599  | 0.626       |
| Primary vs. Model     | 0.592 | 0.620 | 0.576  | 0.609       |
| Primary vs. Random    | 0.544 | 0.546 | 0.538  | 0.559       |
| Model % Difference*   | 66.7% | 69.2% | 62.3%  | 74.6%       |

* Model % Difference: 100 * (Model - Random) / (Reference - Random)

5.1.2 Paired t-Tests. We follow up the REC-AUC analysis with confirmatory two-tailed paired t-tests to directly compare the differences in average max cosine similarity between the model evaluation and the baselines. We again concatenate the sets of 30 max cosine similarity scores for each of M test messages, yet this time we do not normalize them or convert them to cosine distance. This yields three lists of M × 30 max cosine similarities (one for the model evaluation and two for the baselines), and we run two t-tests: (1) comparing difference in mean between the lists for both baselines, and (2) comparing the difference in mean between the model evaluation list and the random-chance baseline list. Each test is run with the null hypothesis that there is no difference between the means of the lists, giving a p-value at the 5% significance level for any observed differences.

In Table 6 we report the results of these tests for both datasets. We again report results for each full test set (ALL) and breakdowns for each twitter account with at least 20 messages in the test sets (WHO, CDCgov, CDCDirector, REC plots for these individual accounts are provided in Appendix A.

To compare model performance across datasets and test accounts, we compute the Model % Difference, which is the position of the model evaluation AUC relative to the distance between the upper and lower bounds established by the two baselines (e.g., 100% indicates model equals reference, and 0% indicates model equals random). Note that for both datasets and for each account, the min cosine distance AUC for the model evaluation is much closer to that of the ground-truth baseline than to that of the random-chance baseline (e.g., Model % Difference = 71.7% for COVID-19 and 66.7% for Vaccines). This indicates that the model is able to capture and reproduce the true semantics of typical responses to messages and authors in our test sets. In the COVID-19 dataset, the model had an easier time reproducing the semantic content of responses to the CDCgov and CDCDirector accounts compared to the WHO and account (e.g., Model % Difference = 86.4% for CDCDirector, 84.8% for CDCgov, and only 61.0% for WHO). However in the Vaccines dataset, the model had the easiest time with CDCDirector, followed by WHO and then CDCgov (e.g., Model % Difference = 74.6% for CDCDirector, 69.2% for WHO, and only 62.3% for CDCgov).

5.1.3 Correlation. Finally, we perform a correlation study between the max cosine similarity scores of the ground-truth baseline (Primary vs. Reference) and those of the random-chance baseline (Primary vs. Random). The purpose of this study is to identify the base level of semantic relatedness that any pair of random responses (to any message) has in each dataset, and investigate the degree to
which this increases for pairs of responses to the same messages. This captures the difficulty inherent in learning to predict semantics conditional on individual messages and authors. For example, imagine a degenerate dataset in which all responses are the same regardless of the message; in such a scenario there would not be much for the model to learn, and we would see a perfect linear correlation between the two baselines.

We use the same concatenated lists of $M \times 30$ max cosine similarities used in the t-tests, this time only using the ones for the ground-truth and random-chance baselines. For each dataset, we compute the Pearson’s correlation coefficient $r$ between these two lists. As seen in Figure 6, we observe that COVID-19 has more semantically diverse responses with correlation $r = 0.58$ p-value $< 2.2 \times 10^{-16}$ between the ground-truth and random-chance baselines, while Vaccines is much less so with $r = 0.71$ p-value $< 2.2 \times 10^{-16}$ between baselines. This indicates that Vaccines presents an "easier" problem for the model with respect to learning semantic distributions. This explains why model evaluation metrics are better for Vaccines (e.g., lower validation perplexity, higher AUC) than for the COVID-19 dataset, yet we see higher Model % Differences for COVID-19.

In Table 6 we report the percentage of test tweets for which there is no significant difference in the sentiment distribution between the compared sets. In Table 7 we report the percentage of test tweets for which there is no significant difference in sentiment distribution on the basis of failures to reject the null hypothesis in pair wise Chi-Square tests, for Primary vs. Reference, Primary vs. Model and Primary vs Random comparison sets, for both the COVID19 and Vaccine datasets. Analyzing the percentage values in each column for the organizations, the Primary vs. Reference comparison set provides the greatest match in sentiments distributions (72.9% for COVID-19 and 63.6% for Vaccines, for ALL organizations) followed by Primary vs. Model (55.5% for COVID-19 and 52.8% for Vaccines, for ALL organizations). The Model has a greater match of sentiments than the Primary vs. Random test (43.9% for COVID-19 and 43.6% for Vaccines, for ALL organizations).

5.2 Sentiment Similarity

Having established that the model effectively generates semantically similar responses to messages from the different accounts, we now analyze the sentiments that are reflected by the modeled responses and compare them against the sentiments reflected in Primary, Reference and Random responses. We assess if the sentiments expressed by the Model and the Primary, Reference and Random populations are distributed similarly.

As discussed in Section 3, we use RoBERTa to assign sentiment scores for each response. We bin the score $(s)$ of each primary, reference, generated, and random response into three classes: (1) Negative, where $-1 \leq s < -0.25$, (2) Neutral, where $-0.25 \leq s \leq 0.25$, and (3) Positive, where $0.25 < s \leq 1$. We then perform three Chi-square tests for each test message to compare the class distribution of its primary ground-truth responses and those of its reference, generated, and random responses. The Chi-squared statistic represents the difference that exists between observed and expected counts, if there is no relationship in the population. The null hypothesis of each test assumes there is no difference in class distribution, and the p-value gives the probability that any observed differences are due to chance. This yields three p-values for each message - one for the ground-truth baseline, one for the model evaluation, and one for the random-chance baseline. The percentage of messages where we fail to reject the null hypothesis with a significance level of 5% is counted for the model and the baselines. These percentages reflect the proportion of messages for which there is no significant difference in the sentiment distribution between the compared sets.

Table 6: Diff. in Mean for max cosine Sim. (paired t-test)

| Dataset          | ALL  | WHO  | CDCgov | CDCDirector |
|------------------|------|------|--------|-------------|
| GT vs. Random Baselines | +0.113 | +0.124 | +0.093  | +0.110       |
| ME vs. Random Baseline    | +0.080 | +0.076 | +0.078  | +0.095       |
| Model % Difference*      | 70.8% | 61.3% | 83.9%   | 86.4%        |

* Model % Difference: 100 * (ME vs. Random Baseline) / (GT vs. Random Baselines)

Table 7: Sentiment Similarity: Results of pair wise Chi-Square tests

| Dataset          | ALL  | WHO  | CDCgov | CDCDirector |
|------------------|------|------|--------|-------------|
| Primary vs. Reference | 72.9% | 77.1% | 70.4%   | 65.4%       |
| Primary vs. Model   | 55.5% | 49.4% | 65.9%   | 57.7%       |
| Primary vs. Random   | 43.9% | 41.0% | 65.9%   | 19.2%       |

5.2 Sentiment Similarity

Having established that the model effectively generates semantically similar responses to messages from the different accounts, we now analyze the sentiments that are reflected by the modeled responses and compare them against the sentiments reflected in Primary, Reference and Random responses. We assess if the sentiments expressed by the Model and the Primary, Reference and Random populations are distributed similarly.

As discussed in Section 3, we use RoBERTa to assign sentiment scores for each response. We bin the score $(s)$ of each primary, reference, generated, and random response into three classes: (1) Negative, where $-1 \leq s < -0.25$, (2) Neutral, where $-0.25 \leq s \leq 0.25$, and (3) Positive, where $0.25 < s \leq 1$. We then perform three Chi-square tests for each test message to compare the class distribution of its primary ground-truth responses and those of its reference, generated, and random responses. The Chi-squared statistic represents the difference that exists between observed and expected counts, if there is no relationship in the population. The null hypothesis of each test assumes there is no difference in class distribution, and the p-value gives the probability that any observed differences are due to chance. This yields three p-values for each message - one for the ground-truth baseline, one for the model evaluation, and one for the random-chance baseline. The percentage of messages where we fail to reject the null hypothesis with a significance level of 5% is counted for the model and the baselines. These percentages reflect the proportion of messages for which there is no significant difference in the sentiment distribution between the compared sets.

In Table 7 we report the percentage of test tweets for which there is no significant difference in sentiment distribution on the basis of failures to reject the null hypothesis in pair wise Chi-Square tests, for Primary vs. Reference, Primary vs. Model and Primary vs Random comparison sets, for both the COVID19 and Vaccine datasets. Analyzing the percentage values in each column for the organizations, the Primary vs. Reference comparison set provides the greatest match in sentiments distributions (72.9% for COVID-19 and 63.6% for Vaccines, for ALL organizations) followed by Primary vs. Model (55.5% for COVID-19 and 52.8% for Vaccines, for ALL organizations). The Model has a greater match of sentiments than the Primary vs. Random test (43.9% for COVID-19 and 43.6% for Vaccines, for ALL organizations).

Thus, the sentiment analysis results on the model-generated responses reflect that the model mostly captures the sentiment distributions of the known ground-truth responses. Only in one instance, Vaccine data set for WHO, the model generated responses
yield a worse percentage than Random when compared against the Primary sentiment distribution.

To further investigate how close the sentiment values from the Model, Primary and Random responses are, we looked at the density distribution of the raw sentiment values from RoBERTa for ALL organizations. Figure 7 represents the density distribution of the sentiment scores provided for the Primary, Model (generated) and Random responses for ALL tweets for each data set.

The density distribution of sentiments from the Primary, Model and Random responses reflect highest density peaks for negative sentiments (peaking close to sentiment value of -1.0). To understand if this is due to the relative differences of public message reception from different organizations, we investigate the density distribution obtained from the sentiments from Primary ground truth messages and responses for each public health organization in Figure 8.

We note that there seem to be more negative sentiments in the ground truth responses for CDCgov and CDCDirector accounts, when compared with that for the WHO. It is important to note that our models are text (response) generators and not directly trained to predict sentiment class likelihood. Also, since the models are not trained separately for each organization, the relative differences in response sentiments between WHO and other organizations may contribute to the diminished performance we observe capturing the true sentiment distribution in responses to WHO messages (as reflected in results from Vaccine data in Table 7).

6 RELATED WORK

We review relevant works which introduce methods for generating social media text (e.g., tweets), or which use social media text as a basis for learning to generate conversational responses. DialoGPT [25] is a GPT-2-based dialogue response generator trained on 147 million "conversations" constructed as distinct paths through comment chains on Reddit. PLATO [1], PLATO-2 [2], and BlenderBot [17] are recent open-domain neural conversational models that also use social media responses in pre-training (PLATO uses both Reddit and Twitter, the others use Reddit only). Cetinkaya et al. [6] propose a neural encoder-decoder-reranker framework for building chatbots that can participate in Twitter conversations by learning to take a side on a controversial issue (e.g., gun control). Tran & Popowich and Roy et al. both explore methods for generating tweets to notify the public about traffic incidents [19, 21]. Lofi & Krestel propose a method to use open government data to generate tweets that inform the public on ongoing political processes [14]. Finally, in perhaps the most related work to ours, Garvey et al. [9] propose a system designed to aid social media content managers design tweets that will be received well and become popular. Their system includes semantic and sentiment analysis components capable of estimating a tweet’s target engagement, which is used in turn with a custom probabilistic generative model to synthesize tweets. Although we share the same motivations and envisioned use cases, what differentiates our work is that Garvey et al. use generative modeling to help a user craft a proposed message and assign it an estimated engagement score, while our method generates responses
to a proposed message. This provides users with a view of what people might actually say if the message goes public, offering crucial insights into the specific concerns that lead to a message being received well (or not). We believe that our methods complement Garvey et al. well - specifically, an organization which adopts both tools might craft promising candidate tweets via Garvey et al. and then preview their reception with our models.

7 CONCLUSION
We conclude with a summary of our contributions and a discussion of limitations, future directions, and ethical considerations.

7.1 Contributions
Our main contributions are as follows: (1) we collected two datasets of public health messages and their responses on Twitter, one in context of COVID-19 and one in context of Vaccines; (2) we trained two GPT-2 text generators - one for each dataset - both capable of capturing and reproducing the semantic and sentiment distributions in known responses to public health messages; (3) we demonstrate our envisioned use case in which a public health organization uses our models to optimize expected reception for important health guidance; and (4) we introduce a novel evaluation scheme with extensive statistical testing to confirm that our models capture semantics and sentiment as we qualitatively observe.

7.2 Limitations and Future Directions
Here we note several key limitations of our study and discuss ways in which future work may address them. Specifically, we discuss the issues of: (1) factuality of generated responses; (2) quality of semantic and sentiment similarity measurement; (3) opportunities for further evaluation; and (4) generalization beyond this study.

7.2.1 Generated Response Factuality. Language models such as GPT-2 are prone to generate factually inaccurate output, often times "hallucinating" details (e.g., names, places, quantities, etc.) in the absence of external knowledge [12]. For example, many of the generated responses in Figures 3 and 4 tag users and/or display hashtags that do not make sense considering the response text. Additionally, our response generator models are prone to temporal drift unless continually re-trained on up-to-date samples from Twitter. For example, our COVID-19 dataset was collected during the spring and summer of 2020 (the early months of the pandemic) and thus would not generate accurate responses to tweets concerning late-pandemic issues such as vaccine boosters, relaxed mask recommendations, and return-to-office policies.

A potential remedy for language model hallucination and temporal drift is to take advantage of recent generative models capable of integrated information retrieval from knowledge bases (e.g., RAG [11]). Retrieval-augmented response generation would allow response predictions to incorporate rapidly evolving information (e.g., breaking news updates) without needing constant re-training, and could increase the general correctness of generated responses with respect to common world knowledge. Additionally, maintaining an up-to-date knowledge base of current events requires less computational resources than continually training language models.

7.2.2 Semantic & Sentiment Measurement Quality. In our study we use off-the-shelf pre-trained models for computing sentence embeddings and sentiment scores. Specifically, the MiniLM sentence embedding model was pre-trained on over one billion sentence pairs from 32 distinct datasets. These include Wikipedia, various Q&A collections, comments from social media forums such as Reddit, Quora, Stack Exchange, and Yahoo Answers, and many others. The RoBERTa sentiment classifier was fine-tuned on the dataset used in the TweetEval sentiment analysis task, which is the SemEval 2017 Twitter sentiment analysis dataset [18] with over 50,000 labeled English tweets. Since these models have already been exposed to large-scale corpora and perform well on their respective benchmarks, we deemed them sufficient for measuring semantic similarity and sentiment on our datasets. However, this comes with the limitation that these models have not seen specialized terminology (e.g., COVID-19) or entities (e.g., names, places) of significance beyond the topical and temporal scope of their training sets.

To mitigate this, it is possible to do additional fine-tuning of the sentence embedding and sentiment models on the collected tweets to ensure the most robust semantic and sentiment comparisons. We encourage any future work employing our methods to explore this avenue. We note that sentence pairs for further training of the embedding model can be mined from raw tweet collections automatically (e.g., select positive example pairs from the same reply thread and negative example pairs at random), but labeling new tweets for sentiment polarity requires manual effort.

7.2.3 Opportunities for Further Evaluation. In this work we evaluate one type of model for response generation (GPT-2). We recognize that response generation is a well studied area, specifically in conversational contexts (e.g., see Section 6), and thus there is opportunity to compare different response generation models on this task. For example, more recent, larger-scale generative models (e.g., GPT-3 [5]) are likely to produce higher quality responses at the cost of increased compute for training and evaluation. However, we note that new language models are constantly being developed and improved, and our proposed methodology supports the replacement of GPT-2 with any current or future text generation model without changing the nature of the task, evaluation, or its use-cases.

We also note the need for end-user validation by our target audience (e.g., public health social media managers). This could be in the form of a trial where users use our models, perhaps in combination with the tools of Garvey et al., to respond to their messages and then compare the system’s predictions with actual responses on Twitter. Such a study may yield valuable information regarding the effectiveness of our methods as day-to-day tools and produce directions for future improvement.

7.2.4 Generalization Beyond This Study. Our statistical evaluations demonstrate the effectiveness of generative response modeling in reproducing the sentiment and semantics of public health responses on Twitter. However, as briefly noted in Section 4.2, there is ample opportunity to generalize our method to other settings. Some possibilities for future work include: (1) allowing the response generator to be conditioned on attributes of the responder (e.g., geographical region, age, etc.) to provide insights into how targeted populations might react to a message; (2) training expanded models on additional author types beyond public health organizations (e.g.,
We recognize the potential dangers presented by the use of language models such as GPT-2 to emulate unfiltered public discourse as we do in this study. The examples in Figure 3 make evident the degree to which such models can be prompted to emit vitriol in this setting, and there is obvious direction for misuse. We take this opportunity to reiterate that our intended use case is to allow social media representatives for impactful organizations to gain accurate perspectives on the way their messages may be received by the public, which requires preserving the real semantics and sentiment of social media discourse regardless of its toxicity. We do not support or condone the use of our methods, models, or data for any purpose that may directly or indirectly cause harm to others.

ACKNOWLEDGMENTS

This study was supported by the Rensselaer Institute for Data Exploration and Applications (IDEA), the Rensselaer Data INCITE Lab, and a grant from the United Health Foundation. Additionally, we thank Brandyn Sigouin, Thomas Shweh, and Haotian Zhang for their participation in the exploratory phase of this project via the Data INCITE Lab.

REFERENCES

[1] S. Bao, H. He, F. Wang, H. Wu, and H. Wang. 2020. PLATO: Pre-trained Dialogue Generation Model with Discrete Latent Variable. In Proc of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, 85–96. https://doi.org/10.18653/v1/2020.acl-main.9

[2] S. Bao, H. He, F. Wang, H. Wu, W. Wang, Z. Guo, Z. Liu, and X. Xu. 2021. PLATO-2: Towards Building an Open-Domain Chatbot via Curriculum Learning. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, Online, 2513–2525. https://doi.org/10.18653/v1/2021.findings-acl-222

[3] F. Barbieri, J. Camacho-Collados, L. Espinosa Anke, and L. Neves. 2020. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. In Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, Online, 1644–1650. https://doi.org/10.18653/v1/2020-findings-emp1-148

[4] J. Bi and K. P. Bennett. 2003. Regression error characteristic curves. In Proceedings of the 20th Int Conf on machine learning (ICML-03). 43–50.

[5] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. 2020. Language models are few-shot learners. Advances in neural information processing systems 33 (2020), 1877–1901.

[6] Y. M. Cetinkaya, I. H. Tchorou, and H. Davuduc. 2020. Developing a Twitter bot that can join a discussion using state-of-the-art architectures. Social network analysis and mining 10 (2020), 1–21.

[7] L-A Cotfas, C. Delcea, I. Roxin, C. Ioanăș, D. S Gherai, and F. Tajariol. 2021. The Longest Month: Analyzing COVID-19 Vaccination Opinions Dynamics From Tweets in the Month Following the First Vaccine Announcement. IEEE Access 9 (2021), 33203–33223. https://doi.org/10.1109/ACCESS.2021.3059821

[8] M. Galley, C. Brockett, A. Sordoni, Y. Ji, M. Auli, C. Quirk, M. Mitchell, J. Gao, and B. Dolan. 2015. deltaBLEU: A Discriminative Metric for Generation Tasks with Intrinsically Diverse Targets. In Proc of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th Int’l Joint Conf on Natural Language Processing (Volume 2: Short Papers). Association for Computational Linguistics, Beijing, China, 445–450. https://doi.org/10.3115/v1/P15-2073

[9] M. D. Garvey, J. Samuel, and A. Pehlau. 2021. Would you please like my tweet?! An artificially intelligent, generative probabilistic, and ecometric based system design for popularity-driven tweet content generation. Decision Support Systems 144 (2021), 113497. https://doi.org/10.1016/j.dss.2021.113497

[10] P. Gupta, S. Mehta, T. Zhao, A. Pavel, M. Eksenas, and J. Bigham. 2019. Investigating Evaluation of Open-Domain Dialogue Systems With Human Generated Multiple References. In Proc of the 20th Annual SIGdial Meeting on Discourse and Dialogue. Association for Computational Linguistics, Stockholm, Sweden, 379–391. https://doi.org/10.18653/v1/W19-5944

[11] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W-Y. Yih, T. Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems 33 (2020), 9459–9474.

[12] T. Liu, Y. Zhang, C. Brockett, Y. Mao, Z. Sui, W. Chen, and B. Dolan. 2021. A token-level reference-free hallucination detection benchmark for free-form text generation. arXiv preprint arXiv:2104.08704 (2021).

[13] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. CoRR abs/1907.11692 (2019). https://arxiv.org/abs/1907.11692

[14] C. LoB and R. Kroeste. 2012. Participating on Twitter: Feelings and Engagement. In Proceedings of the 16th Conference of the European Chapter of the Assoc for Computational Linguistics: Main Volume. Association for Computational Linguistics, Online, 300–325. https://doi.org/10.3115/v1/eacl-2014.24

[15] S. Rosenthal, N. Farra, and N. Pavlovic. 2017. TweetEval-2017 Task 4: Sentiment Analysis in Twitter. In Proc of the 3rd Int’l Workshop on Semantic Evaluation (SemEval-2017). Association for Computational Linguistics, Vancouver, Canada, 502–518. https://doi.org/10.18653/v1/S17-2088

[16] B Roy, P. Dutta, and S. Choudhury. 2022. Probabilistic Traffic Tweet Generation Model. In Advances in Distributed Computing and Machine Learning; J. P. Sahoo, A. K. Tripathy, M. Mohaanta, K-C. Li, and A. K. Nayak (Eds.). Springer, Singapore, 250–260.

[17] B. R. Sanders, R. C. White, L. S. Severson, R. Ma, R. McQueen, H. C. A. Paula, Y. Zhang, J. S. Erickson, and K. P. Bennett. 2021. Unmasking the conversation on masks: Natural language processing for topical sentiment analysis of COVID-19 Twitter discourse. In AMIA Annual Symposium Proc, Vol. 2021. American Medical Informatics Association, 555.

[18] K Tran and F. Popowich. 2016. Automatic Twitter Generation From Traffic Incident Data. In Proc of the 2nd Int’l Workshop on Natural Language Generation and the Semantic Web (WebNLG 2016). Association for Computational Linguistics, Edinburgh, Scotland, 59–66. https://aclanthology.org/W16-3512

[19] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou. 2020. MinuLM: Deep Self Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. arXiv:2002.10957 [cs.CL]

[20] Y. Wang, H. Hao, and L. Sundahl Platt. 2021. Examining risk and crisis communications of government agencies and stakeholders during early-stages of COVID-19 on Twitter. Computers in Human Behavior 114 (2021), 106568. https://doi.org/10.1016/j.chb.2020.106568

[21] B. Wei, S. Lu, L. Mou, H. Zhou, P. Pospurt, G. Li, and Z. J. Min. 2019. Why Do Neural Dialog Systems Generate Short and Meaningless Replies? A Comparison Between Chinese and English. In IJCAI-19 (2019). https://doi.org/10.24963/ijcai.2019/8628

[22] Y. Zhang, J. Sun, M. Galley, Y-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan. 2020. DISTILGPT : Large-Scale Generative Pre-training for Conversational Response Generation. In Proc of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Association for Computational Linguistics, Online, 270–278. https://doi.org/10.18653/v1/2020.acl-demos.30
A  ADDITIONAL REC PLOTS

The following shows the semantic similarity REC curves for each individual public health organization account with at least 20 messages in the test set of both datasets. Figure 9 shows REC curves for the COVID-19 dataset and Figure 10 for the Vaccines dataset. The REC curves for each full test set (left-most plot in each figure) are provided here again to facilitate comparison.

Figure 9: REC curves using the min cosine distance error metric on the test set of the COVID-19 dataset.

Figure 10: REC curves using the min cosine distance error metric on the test set of the Vaccines dataset.