Extracting Sentiment-Polarizing Tokens from COVID-19 Topic Tweets

Alice Chen ayjchen@mit.edu
Keenly Chuang kschuang@mit.edu
Emily Jiang emji@mit.edu
Filip Ryzner ryznerf@mit.edu

Abstract

Our research investigates the polarization of public sentiment surrounding the COVID-19 pandemic and related public health guidelines. In particular, we develop a method to isolate polarizing terms, or words that are important in influencing both the positive and negative sentiment of a text. We use the well-studied, lightweight transformer model DistilBERT to train a multiclass sentiment classifier on a dataset of pandemic-related tweets. We then design and implement a novel method of interpreting self-attention weights to quantify relative polarization between terms. For each word, we analyze the self-attention weights for each tweet it appears in, using KL-divergence to isolate the most informative self-attention head per layer. We compute the variance in self-attention scores across tweets to compare the polarity of different words. Overall, our self-attention interpretation method isolates many pandemic-related terms as most polarizing, with words relating to quarantine, vaccines, and historically controversial COVID-19 drugs showing significant contribution toward determining a tweet's sentiment in both positive and negative directions.

1 Introduction

As social media platforms assume an increasingly prevalent role in modern sociopolitical discourse, the largely unstructured data found in their text posts, comments, and reactions offer insight into how current events are perceived among different groups of users worldwide as well as how that perception evolves over time. In particular, throughout the ongoing COVID-19 pandemic, social media use has increased significantly as people shifted more of their social interactions online during periods of isolation; for instance, US users spent a daily average of 65 minutes on social media in 2020 compared to 54 minutes the previous year [5]. As such, we expect popular micro-blogging platform Twitter, with over 150 million daily users, to be a useful source of natural language data that reflects real-time sentiment of users about the ongoing pandemic.

In this paper, we examine how tweet language shapes specific public sentiment by exploring tweets posted during and about the COVID-19 pandemic. We are not as concerned with what kind of language engenders negative or positive sentiment, which is a highly-researched problem, but rather what terms or phrases are associated with particularly *polarizing* sentiment. Over the past few years, pandemic policies and public health guidelines in the US have become politicized along party lines, resulting in certain concepts being perceived highly differently across sociopolitical groups [18]. This division in opinion also caused divided perception of similar issues, thus we aim to use natural language processing models trained on recent Twitter data to search for the most polarizing words.

To this end, we develop a sentiment analysis model on the text of English-language tweets flagged as being associated with COVID-19. We analyze the model's parameters and weights in order to identify the most polarizing COVID-related words; that is, words that significantly influence sentiment in both very positive and very negative tweets, with the assumption that such words are controversial in some sense. We hypothesize that the most polarizing terms identified would be topics that different people have highly contrasting preconceptions about, like masks, vaccines, and social distancing or gathering guidelines. In general, the sentiment of pandemic-related vocabulary provides important insight into understanding individual perceptions of this globally momentous era. Additionally, the task of identifying extreme-

sentiment terms is significant because pertinent messages about public health are unlikely to reach a broad audience if they are delivered using unintentionally polarizing language.

2 Related Works

Following the groundbreaking Attention is All You Need paper [21], transformer-based models have become the standard approach to many NLP tasks, including sentiment analysis. For the related task of multiclass sentiment classification on coronavirus tweet data, [22] compare the performance of vector space models (Bag of Words and TF-IDF) with Support Vector Machines, word embeddings (Word2Vec [11] and GloVE [14]) with Long Short-Term Memory [6], and BERT [4]. They found that BERT achieves the best performance with a weighted F1 score of 0.85 for the 3-class classification problem. The research additionally found that BERT had an even clearer advantage than the other two model types when working with *five* sentiment classes. They also concluded that DistilBERT [16] performed similarly to BERT, despite its smaller size [22]. This motivates our decision to choose deep contextual language models like DistilBERT over traditional context-independent systems such as TF-IDF.

Transformers achieve state-of-the-art results in no small part due to the self-attention mechanism. Analogously to how humans use attention as a means of selective focus, both standard attention and self-attention have been used to interpret how models weigh particular tokens for determining their objectives. Intuitively, the higher the attention weight on a particular token, the more it contributes towards the output of that layer. In text classification tasks for Long Short-Term Memory (LSTM) and Recurrent Neural Networks (RNN) models, attention does not seem to play a significant role in determining the final output. In multiple studies, researchers were able to create adversarial weights or use uniform weights to achieve similar results to a trained model with attention, indicating that the embedding of the text is more significant for final results [7; 20]. In transformer models however, the self-attention mechanism does not replicate this behaviour. Studies involving BERT show that the self-attention weights are integral to the performance of the model, and performance suffers immensely when they are uniformly or randomly set [20]. Moreover, attention weights have been cited as being effective in determining the usefulness of particular components in a model, even in sentiment analysis [3; 17]. In particular, where attention weights were higher on words predicting both negative and positive sentiment. The interpretability of neural nets with attention mechanisms is an active area of research, and our project aims to contribute by providing a novel and empirically insightful approach.

Interpretation of the attention head weights generated by the transformer models for analysis of sentiment token importance has also been explored by previous research. Specifically, [12] use RoBERTa to perform sentiment analysis on text extracted from Russian-language news articles. The researchers investigated the interpretability of weights across self-attention heads, with the hypothesis that BERT-based models pay more attention to "sentiment" words than neutral ones in determining sentiment. They interpreted the self-attention weights by essentially computing the average attention vector across all the heads for each token, and through analysis of Kullback–Leibler divergence (KL divergence) from each attention head and the average attention vector, found that some (but not all) heads focus on sentiment words, especially negative words [13]. This paper motivates our analysis of attention weights by inspiring the use of computing an average of weights across attention heads for an particular token (e.g. the classifier [cls] token) as a reasonable representation of how much other tokens relate to that token overall.

Considering our research topic in the context of other COVID-19 sentiment analysis works, existing research in this space establishes the precedent of using Twitter as a data source, but has primarily focused on regions or research topics different from ours. For example, [2] compare the performance of several popular natural language processing models in training COVID-specific sentiment classifiers. The data used were tweets posted in India during lock-downs in 2020, where each tweet text was labelled as expressing one of fear, sadness, anger, or joy. The study found that the Bidirectional Encoder Representations from Transformers (BERT) model significantly outperformed logistic regression, SVM, and LSTM models, achieving an accuracy of 89% [2]; this matches the expectation of transformer-based models to be the gold standard for sentiment analysis tasks.

Most other papers aim to provide insight on current societal attitudes, rather than to optimize a classifier. Specifically, [1] evaluate public Twitter sentiment toward vaccination up to the global vaccination roll-out in 2021, using topic modeling and VADER sentiment analysis to predict sentiment on a continuous scale. The paper's literature review also highlights the findings of similar NLP studies on social media data that explore vaccine sentiments in various geographical scopes, compare rates of vaccine hesitancy between different major cities, and document increases in vaccine opposition over time, among other focuses [1]. Meanwhile, Jang et. al. have broader scope, identifying major pandemic-era public concerns in North America and analyzing the sentiment related to each one. The researchers had an additional goal of investigating anti-Asian racism in the context of the pandemic, and indeed found negative sentiments associated with the overall outbreak, misinformation, and Asians, while positive sentiments were associated with physical distancing [8].

Across the literature, while there is much focus on quantifying public sentiment towards pandemic-related issues, there is a lack of research on identifying *controversial* terms or topics. We hope that our research focus will expand upon these existing sentiment analysis models, with an emphasis on identifying polarizing topic phrases and their influence on sentiment trends. Recognizing sentiment as highly dependent on individual influences, we believe our approach adds nuance to the current state of pandemic-related NLP research.

3 Data

Our text data was sourced from the Coronavirus (COVID-19) Tweets Dataset [9], which is available on IEEE Dataport for use under the Creative Commons Attribution license. The dataset consists of English-language tweets that were selected from the global Twitter feed if they contained any of 90+ keywords or hashtags commonly associated with the pandemic. The dataset itself contains the tweet IDs and a corresponding sentiment score from -1 (most negative) to 1 (most positive).

We downloaded over 300,000 tweet IDs from this dataset across various periods during the pandemic. We obtained the tweet text and metadata corresponding to those IDs using the Hydrator desktop application [19]. However, the majority of the downloaded tweets were retweets, so we removed duplicate occurrences of the same tweet text. Following deduplication, we were left with around 85,000 tweets with sentiment scores.

3.1 Text Pre-Processing

Tweet text is often noisy, containing abbreviations, hashtags, and mentions of other Twitter users. To streamline our data before training, we applied standard tweet pre-processing techniques, including converting all text to lowercase, removing punctuation, user account mentions (words starting with @), and links. Finally, we performed word lemmatization, a common technique that reduces variations of words into their root form (e.g. *doesn't* would be lemmatized into *do not*). Table 1 lists a few examples of original tweet texts, and Table 2 shows the resulting text of the same tweets after applying our pre-processing steps.

Raw Tweet Text
Sorry to disappoint, but I've not made ...
@mayawiley Patches. She is with us ...
DocSend CEO peers into a post-pandemic ...

Table 1: Sample Raw Tweet Texts

Preprocessed Tweet Text
sorry to disappoint but I ve not make ...
patch she be with we ...
docsend ceo peer into a post pandemic ...

Table 2: Sample Pre-Processed Tweet Texts

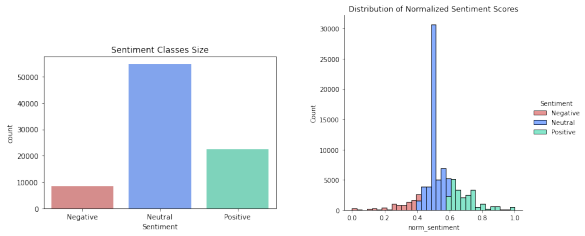
3.2 Sentiment Score Pre-Processing

As outlined above, the ground truth sentiment scores for individual tweets were initially in the interval of [-1, 1]; we normalized the sentiment values to fall within the range [0, 1]. Since training a regression model using continuous sentiment values generally requires much more data to achieve sensible results, we simplified the sentiment scores to three categorical values: Negative, Neutral, and Positive Sentiment. The category membership was determined based on the value of the normalized sentiment, where tweets with normalized sentiment in the range [0, 0.4] were labeled as having Negative Sentiment, tweets with normalized sentiment in the range [0.4, 0.6] were labeled as Neutral, and the remainder was labeled as having Positive Sentiment.

The introduction of the Neutral Sentiment category was motivated by our research goal. As we aim to determine the most polarizing words or words responsible for the tweet sentiment, we only want to base our result on tweets that can be clearly labeled as negative or positive.

3.3 Dataset Preparation

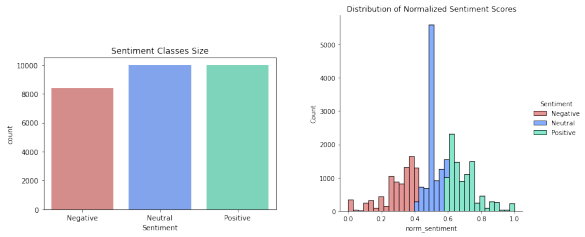
During the exploratory analysis of the dataset of 85,000 tweets, we found that the vast majority have Neutral Sentiment. In Figure 1a, which depicts the sentiment occurrences counts in the full dataset, we observe that both the Negative and Positive Sentiment tweets are significantly underrepresented. Figure 1b shows that the normalized sentiments resemble a normal distribution with extremely positive kurtosis driven by the large proportion of neutral-sentiment tweets.



(a) Sentiment Counts **(b)** Sentiment Distribution

Figure 1: Full Dataset Sentiment Distribution

Training a classifier on an unbalanced dataset is problematic as it is likely to result in a biased model. To alleviate this problem, we downsampled [15] the Neutral and Positive classes to 10,000 samples and kept the 8,401 Negative ones. We believe that given the potential diversity of tweets, it did not make sense to up-sample the under-represented classes as that would not add any diversity to the data. The results of the down-sampling procedure are depicted in Figure 2a, which displays the class balance and in Figure 2b, which depicts the new normalized sentiment distribution.



(a) Sentiment Counts **(b)** Sentiment Distribution

Figure 2: Downsampled Sentiment Distribution

4 Methodology

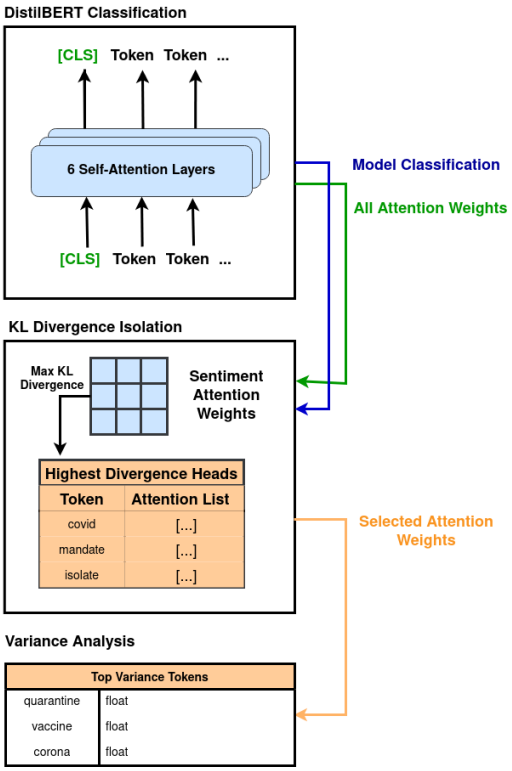


Figure 3: The data processing pipeline runs in three main stages. Firstly, the input sequences are fed through the DistilBERT transformer model to obtain sentiment classifications and attention head weights. Second, the weights and classification are modified and the highest KL-divergence head is selected. The dataset is populated with attention weights from that head, and the variance for attention heads per token is calculated.

4.1 Context-Independent Baseline Approach

Our baseline approach uses two common word representations: Bag-of-Words (BoW) and Term Frequency–Inverse Document Frequency (TF-IDF). For both representations, we train a binary logistic sentiment classifier using the `sklearn` Python package. After training the model, we extract the vector of weights w where each position represents a coefficient in the regression model corresponding to a specific word token. Our assumption is that these coefficients represent the sentiment of the word in the model, as these coefficients were obtained from a model that was specifically trained to determine the overall tweet sentiment.

However, this approach is limited by the context-independent nature of its word representations — by assigning each word only a single sentiment

coefficient, we by definition cannot fulfill our research goal of identifying *polarizing* terms across different tweet contexts. With that, we look more closely into context-dependent transformer-based approaches.

4.2 Transformer-Based Approach

We refine our model to use a transformer-based approach trained on 3-class sentiment of positive, negative, and neutral sentiment (in contrast to the binary positive/negative sentiment used for the baseline model). We chose multiclass over binary sentiment because our aim is to isolate words that contribute strongly to extreme sentiment (positive or negative) tweets, and we thus prefer to explicitly label neutral tweets during training rather than forcing a positive/negative label.

We use the pretrained DistilBERT transformer from the HuggingFace platform. Since we are performing a classification task, we add the classification token [CLS] as the first of every tokenized tweet; additionally, the model automatically adds the end-of-sentence [SEP] marker as the last token. We train the transformer using the default architecture of 6 layers and 12 self-attention heads.

4.2.1 Analyzing Attention Weights

The self-attention weights between tokens (specifically, between the classification [CLS] token and all other word tokens in the tweet) can be interpreted to isolate polarizing terms. Prior to extracting the polarizing terms, however, we first perform pre-processing on the raw attention weights.

1. **Merging split-token attention weights**. The HuggingFace tokenizer splits large words like "quarantine" into multiple tokens for training, such as "qu", "##aran", "##tine", which each have their own corresponding attention values. We perform a merging procedure to take the summation of these fragment attentions for each large word in a input sentence.

2. **Averaging attentions**. We average all self-attention weights across the 12 heads in each layer. This results in 6 attention average values for each of the 6 transformer layers.

3. **Computing Kullback–Leibler divergence**. For each layer, we computed the Kullback-Leibler (KL) divergence (using the `torchmetrics.functional` package) between each self-attention head and the layer's average self-attention vector calculated in the previous part. Again, we only consider the self-attention vectors on the [CLS] token since the classification token represents the tweet's sentiment. The KL-divergence metric measures the distance between two probability distributions (the attention vectors in our case). The head with the highest KL-divergence in each layer is predicted to be generally more important for sentiment [13], as it likely captures more unique relationships than other heads.

   For each non-stopword in the corpus, we store the self-attention score it is assigned by the highest-divergence head, for all of the tweets it appears in. In the list of attention scores for each word, we negate the values from tweets classified as negative.

4.3 Isolating Polarizing Terms

After performing the above preprocessing steps, we end up with a list of attention scores for each word, where attentions corresponding to negative tweets are negated, as aforementioned. We then identify a list of candidates for polarizing words by calculating the variance of the attention list for each word. Our assumption is that words with the highest variance in attentions contribute significantly to sentiment in both positive and negative tweets, and thus polarize sentiment.

To ensure the polarizing words have a sufficient balance of positive and negative sentiment, we filtered out words that have high variance but do not positive sentiment tweets or negative sentiment tweets. We conclude that these tokens may be strongly sentiment-driving but are not *polarizing* as we desire since they primarily influence sentiment in only one direction. We perform this filtering by setting the parameter `attn_threshold=0.001` and determining the validity of the current candidate word as follows:

$$\textbf{valid}(w_i) = \min(|\min(a_i)|, \max(a_i))) > \text{attn\_threshold}$$

where a_i is the self-attention vector for the ith word.

After filtering, we sort the candidates in order of decreasing variance to isolate the most polarizing terms per transformer layer.

5 Results and Discussion

5.1 Baseline Results

The sentiment analysis from our baseline bag-of-words and TF-IDF word representation models with logistic regression, trained with ***binary*** sentiment values (0/1) over 2000 iterations and with normalized final sentiment coefficients, are shown in Tables 3 and 4 below, respectively. To sanity check our results, we selectively examined the sentiment score of some key terms related to COVID-19.

Table 3: Bag-of-Words Sentiment Scores (Binary)

Word Token	Sentiment Score
covid	0.070
mask	0.006
quarantine	0.285
pandemic	0.040
virus	-0.057
vaccine	0.121
distancing	0.133
corona	-0.025

The training accuracy with the **bag-of-words** representations was **0.949**, and the testing accuracy was **0.879**.

Table 4: TF-IDF Sentiment Scores (Binary)

Word Token	Sentiment Score
covid	0.645
mask	0.154
quarantine	0.478
pandemic	0.284
virus	0.101
vaccine	0.322
distancing	0.086
corona	0.29

The training accuracy for **TF-IDF** transformed count matrices was **0.987**, and the testing accuracy was **0.882**.

Though our baseline results have some outliers, the majority of sentiment scores for our selected COVID-19 related words are relatively small compared to other words. For example, the highest magnitude of the top 5 most positive and most negative sentiment score for the BOW approach were above 0.69, whereas the highest magnitude score for our COVID-19 terms is "quarantine", with a score of 0.285. Likewise, for the TF-IDF approach, the highest magnitude for the top 5 most positive and negative sentiment words are above 0.56, and the highest magnitude for our COVID-19 terms is also "quarantine", with a score of 0.478. Interestingly, the scores for nearly all of the COVID-19 related terms are positive, with only "virus" and "corona" having negative sentiment for the BOW approach. Our baseline results suggest that most COVID-19 related terms are not positively or negatively charged as they produce nearly neutral sentiment scores, but our model fails to capture how these words are being used in both negative and positive sentiment tweets.

5.2 Transformer Model Results

During our analysis, we found that the model heavily weighted sentiment adjectives and adverbs, which led to the highest variance being in the last two layers. However, the model also seemed to attend to semantically similar sets of words in other layers. In particular, COVID-19 related terms appeared frequently among the highest variance terms in the first and second layers, while words relating to nationality and ethnicity were the highest in the subsequent two layers. The earlier layers had overall lower self-attention magnitudes on the [CLS] token once stop words were removed, and thus had lower variance in sentiment attention as well.

5.2.1 Polarizing Terms

Upon observing that the first and second transformer layers, in contrast to the other layers, appeared to attend most strongly to nouns and COVID-related words in general, we analyzed the self-attention variances for Layer 0. As Section 4.3 describes, we first filter out insufficiently polarizing words; we also only consider word tokens that occur in the corpus at least more often than average, or at least **11** times for our tweet corpus.

Under this criteria, we identified the top 50 words with highest attention variance. We observed that a significant proportion (72%) of these words were reasonably relevant to the pandemic, and we manually removed unrelated words, such as common terms like "tweet", "af", and "another", as well as unrelated abbreviations like "aur" and "gst," which may have been slang, typos, or noise.

Some particularly interesting COVID-19 words identified as most polarizing in tweets by the transformer model are displayed in Table 5 below in or-

der of decreasing variance. For similar words (e.g. "sanitize," "disinfectant," "sanitizer"), we only list the most polarizing variant in the table. We note that the magnitudes of the variance are not particularly meaningful; for the task of sentiment analysis, the transformer model tends to heavily weight high-sentiment adjectives and adverbs, which indeed have high self-attention weights in the last two layers. In contrast, while the COVID-19 related terms may be controversial, the overall sentiment of a tweet is determined more by sentiment terms. However, isolating earlier layers reveals that our model does group terms by semantic meaning.

The full list of the top 50 highest-variance words, with unrelated words filtered out, is reproduced in full in Table 7 of the appendix.

Table 5: Transformer Layer 0 Polarizing Terms With Above-Average Occurrence (Abridged)

Term	**Variance** $\times 10^5$
hydroxychloroquine	4.69
sanitize	2.59
remdesivir	2.34
quarantine	2.33
unvaccinated	2.22
astrazeneca	1.81
asymptomatic	1.59
curfew	1.32
fauci	1.29
covidiot	1.09
pfizer	1.02

Notably, our model identifies drugs and sanitation measures related to COVID-19 as the terms with highest variance. The highest variance term, "hydroxychloroquine," is not a drug recommended by medical professionals or the World Health Organization to treat symptoms of COVID-19, but the drug historically sparked debate online as an alternative treatment to vaccination [10]. Additionally, "remdesivir" is a anti-viral medication whose efficacy for COVID-19 treatment was debated. We also see a number of vaccine-related controversial words (e.g. "unvaccinated," "astrazeneca," "pfizer," as well as "hesitancy," "inject," and more from the full list), and words related to public policy (e.g. "curfew," "fauci") which is in line with our hypothesis in Section 1.

We also experimented with keeping the same attention threshold but only considering words occurring more often than one standard deviation above average, or at least **88** times in our tweet corpus.

Under these parameters, we observed that 54% of the top 50 highest-variance words were relevant to the pandemic; the full list with unrelated terms removed is reproduced in Table 8 of the appendix. We again isolate a selection of the most interesting polarizing terms in Table 6 below. In particular, "quarantine", "symptom", and "vaccination" were the highest variance terms, which were all heavily discussed and debated over Twitter during the pandemic. "Pfizer", "distancing", and "pandemic" also had high variance.

Table 6: Transformer Layer 0 Polarizing Terms With 1 Standard Deviation Above Average Occurrence (Abridged)

Term	**Variance** $\times 10^5$
quarantine	2.33
symptom	1.44
vaccination	1.04
pfizer	1.02
distancing	0.86
pandemic	0.82
covid19	0.65
modi	0.62
immunity	0.59
lockdown	0.38
ban	0.36
biden	0.35
wave	0.15

In addition to commonly controversial words like "quarantine" and "pfizer" that also appear using the less-aggressive min-occurrence threshold, we especially notice that the tokens "biden" and "modi" are identified as high variance here, which corroborates the controversy over both leaders' policies and approach to handling the COVID-19 pandemic crisis in their respective countries. Overall, using the higher min-occurrence threshold observes a lack of overly specific drug titles, pharmaceutical company names, and COVID-19-centric slang language compared to the more relaxed threshold. Instead, we find more general polarizing terms that characterize the sociopolitical tension of the pandemic as a whole among public Twitter platforms.

Comparing the two min-occurrence thresholds, we find that the lower min-occurrence threshold filtered for higher variance words, but the higher threshold filtered for more general COVID-19 related terms. In particular, the higher min-occurrence threshold selected for more general terms that were likely to be commonly used in tweets; the lower min-occurrence threshold allowed terms that came up less frequently to appear. In the extreme case, removing the min-occurrence criteria allowed terms with only two or three instances (often not relevant to the pandemic) throughout the dataset to appear as high variance. Overall, adjusting the min-occurrence threshold appears to change the balance between highly-contested, more niche topics and general controversial discussions.

6 Conclusion

By collecting the self-attention weights corresponding to each word for all the tweets it appears in, we effectively isolated polarizing (i.e. high variance in attention weights across tweets) COVID-19 related terms from Layer 0 in the transformer network. Particular words of interest included "hydroxychloroquine," "quarantine," and words related to vaccines. By adjusting the threshold for min-occurrence, we were also able to adjust the balance of the magnitude to which terms are polarizing as well as their frequency of use and genarality. Based on how our results were able to capture and isolate both COVID-19 related terminology as well as highly debated topics, we conclude that our novel method of analysis was effective in determining polarizing terms from the Twitter dataset.

Given more time and computational resources, a natural extension of this work would be to perform similar analysis on larger language models and datasets or models with more parameters or pre-training. Another potential direction for further research in this area would be to introduce a time-based parameter in order to investigate trends in controversial terms over the timeline of the pandemic. This is an especially pertinent topic for COVID-19 research, as different controversies surrounding major worldwide events and wide-reaching policies have arisen and passed since 2020. We previously discussed Chandrasekaran et. al.'s work on public sentiment toward vaccination over time [1]; we expect that similarly, a time-based analysis of polarizing tweet language will yield results of sociopolitical significance. Additionally, we can extend the interpretation of our model attentions by experimenting with other aggregation and divergence methods beyond the simple average of attention head weights per layer and KL-divergence in our current analysis.

References

[1] CHANDRASEKARAN, R., DESAI, R., SHAH, H., KUMAR, V., MOUSTAKAS, E., ET AL. Examining public sentiments and attitudes toward covid-19 vaccination: infoveillance study using twitter posts. *JMIR infodemiology 2*, 1 (2022), e33909.

[2] CHINTALAPUDI, N., BATTINENI, G., AND AMENTA, F. Sentimental analysis of covid-19 tweets using deep learning models. *Infectious Disease Reports 13*, 2 (2021), 329–339.

[3] CÓRDOVA SÁENZ, C. A., AND BECKER, K. Assessing the use of attention weights to interpret bert-based stance classification. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (2021), pp. 194–201.

[4] DEVLIN, J., CHANG, M.-W., LEE, K., AND TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[5] DIXON, S. Social media use during coronavirus (covid-19) worldwide, Oct 2022.

[6] HOCHREITER, S., AND SCHMIDHUBER, J. Long short-term memory. *Neural computation 9*, 8 (1997), 1735–1780.

[7] JAIN, S., AND WALLACE, B. C. Attention is not explanation. *arXiv preprint arXiv:1902.10186* (2019).

[8] JANG, H., REMPEL, E., ROTH, D., CARENINI, G., JANJUA, N. Z., ET AL. Tracking covid-19 discourse on twitter in north america: Infodemiology study using topic modeling and aspect-based sentiment analysis. *Journal of medical Internet research 23*, 2 (2021), e25431.

[9] LAMSAL, R. Coronavirus (covid-19) tweets dataset, 2020.

[10] MEYEROWITZ, E. A., VANNIER, A. G., FRIESEN, M. G., SCHOENFELD, S., GELFAND, J. A., CALLAHAN, M. V., KIM, A. Y., REEVES, P. M., AND POZNANSKY, M. C. Rethinking the role of hydroxychloroquine in the treatment of covid-19. *The FASEB Journal 34*, 5 (2020), 6027–6037.

[11] MIKOLOV, T., CHEN, K., CORRADO, G., AND DEAN, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[12] PASHCHENKO, D., RAZOVA, E., KOTELNIKOVA, A., VYCHEGZHANIN, S., AND KOTELNIKOV, E. Interpretation of language models attention matrices in texts sentiment analysis. In *2022 VIII International Conference on Information Technology and Nanotechnology (ITNT)* (2022), IEEE, pp. 1–4.

[13] PASHCHENKO, D., RAZOVA, E., KOTELNIKOVA, A., VYCHEGZHANIN, S., AND KOTELNIKOV, E. Interpretation of language models attention matrices in texts sentiment analysis. *VIII International Conference on Information Technology and Nanotechnology* (2022).

[14] PENNINGTON, J., SOCHER, R., AND MANNING, C. D. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (2014), pp. 1532–1543.

[15] POOLSAWAD, N., KAMBHAMPATI, C., AND CLELAND, J. Balancing class for performance of classification with a clinical dataset. In *proceedings of the World Congress on Engineering* (2014), vol. 1, pp. 1–6.

[16] SANH, V., DEBUT, L., CHAUMOND, J., AND WOLF, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[17] SERRANO, S., AND SMITH, N. A. Is attention interpretable? *arXiv preprint arXiv:1906.03731* (2019).

[18] STROEBE, W., VANDELLEN, M. R., ABAKOUMKIN, G., LEMAY JR, E. P., SCHIAVONE, W. M., AGOSTINI, M., BÉLANGER, J. J., GÜTZKOW, B., KREIENKAMP, J., REITSEMA, A. M., ET AL. Politicization of covid-19 health-protective behaviors in the united states: Longitudinal and cross-national evidence. *PloS one 16*, 10 (2021), e0256740.

[19] THE NOW, D. Hydrator. https://github.com/docnow/hydrator, 2020.

[20] VASHISHTH, S., UPADHYAY, S., TOMAR, G. S., AND FARUQUI, M. Attention interpretability across nlp tasks. *arXiv preprint arXiv:1909.11218* (2019).

[21] VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, Ł., AND POLOSUKHIN, I. Attention is all you need. *Advances in neural information processing systems 30* (2017).

[22] WISESTY, U. N., RISMALA, R., MUNGGANA, W., AND PURWARIANTI, A. Comparative study of covid-19 tweets sentiment classification methods. In *2021 9th International Conference on Information and Communication Technology (ICoICT)* (2021), IEEE, pp. 588–593.

Appendix

Below are the full lists of polarizing pandemic-related terms identified from the first layer of our transformer model, with min-occurrence thresholds of 11 and 88.

Table 7: Transformer Layer 0 Polarizing Terms With Above-Average Occurrence (Full)

Term	Variance
hydroxychloroquine	4.69E-05
sanitize	2.59E-05
disinfectant	2.46E-05
sanitizer	2.36E-05
remdesivir	2.34E-05
quarantine	2.33E-05
unvaccinated	2.22E-05
irresponsible	2.07E-05
incompetence	1.98E-05
disinformation	1.84E-05
astrazeneca	1.81E-05
ivermectin	1.77E-05
exercise	1.64E-05
asymptomatic	1.59E-05
insensitive	1.53E-05
incompetent	1.50E-05
hesitancy	1.47E-05
symptom	1.44E-05
precaution	1.36E-05
curfew	1.32E-05
infect	1.32E-05
fauci	1.29E-05
nhi	1.23E-05
mismanage	1.22E-05
mismanagement	1.20E-05
allocate	1.19E-05
ventilator	1.18E-05
pharma	1.16E-05
inject	1.16E-05
whatsapp	1.12E-05
sputnik	1.11E-05
covidiot	1.09E-05
hypocrisy	1.05E-05
vaccination	1.04E-05
positivity	1.03E-05
pfizer	1.02E-05

Table 8: Transformer Layer 0 Polarizing Terms With 1 Standard Deviation Above Average Occurrence (Full)

Term	Variance
quarantine	2.33E-05
symptom	1.44E-05
infect	1.32E-05
vaccination	1.04E-05
pfizer	1.02E-05
distancing	8.64E-06
pandemic	8.17E-06
vaccinate	7.29E-06
covid19	6.52E-06
modi	6.21E-06
immunity	5.88E-06
covid	4.08E-06
lockdown	3.84E-06
ban	3.55E-06
biden	3.51E-06
fall	3.24E-06
coronavirus	3.07E-06
politic	2.81E-06
rally	2.51E-06
order	2.23E-06
price	1.99E-06
flu	1.99E-06
die	1.94E-06
vaccine	1.85E-06
government	1.77E-06
election	1.51E-06
wave	1.48E-06

Impact Statement

During periods of national crisis, such as the ongoing COVID-19 pandemic, delivering accurate and helpful information to large groups of people of different backgrounds and identities remains a complex process where the ramifications of accidental misinterpretation or misinformation can lead to serious medical or social harm. As society has increasingly shifted communication to online social media platforms such as Twitter, special attention has centered on language choice in tweets as a driving force in shaping public sentiment. Our project model discovered examples of such polarizing words within COVID-19 related tweets that were observed to prominently spark both positive, approving reactions as well as negative, disparaging reactions to tweets containing those words.

Abuse of incendiary terms like those extracted from our project carries potential risks by driving audience groups away from crucially important medical information or updates, especially in environments like Twitter where instant emotional reactions inspire echo chamber-like groups that lead to the distortion of original facts and spread of misinformation (indeed, "disinformation" is one of our most controversial words from Table 7!). Thus, we hope that our work with identifying key polarizing terms will inspire informative COVID-19 tweets to be drafted with more neutral-sentiment language so that important news is delivered to the public without sparking hateful, racist, or nonfactual propaganda in response. While some terms like "quarantine" or "pandemic" are likely unavoidable when discussing in the context of COVID-19, we strongly urge from our findings that terms like "covidiot" are unnecessary and should be removed from tweet language in order to minimize hateful discourse, further disunity and distrust within the country, and possible reactionary misinformation.

Our model is also limited by the own bias of its data source. We obtain the corpus texts to train the transformer model directly from raw COVID-19 tweets published by actual users who are effectively anonymous, so the personal bias of those users is reflected in the words we characterize as polarizing as well. For example, while we conclude that layer 0 of the transformer reasonably labeled noun phrases and COVID-19 terms as controversial, other layers of the network displayed unrelated terms like "indian", "kenyan", and "hk" (Hong Kong), whose controversy status may be driven by racist sentiment in some Twitter communities. By directly using public sentiment as a gauge for word controversy, we also inevitably capture the racism and prejudice within Twitter platforms as well and risk labeling innocuous words as controversial simply because there are groups of users that react irrationally negatively (or positively) to those words. To avoid perpetuating harmful discrimination in our model, we encourage future work on this project to incorporate de-biasing on the resulting polarized words before utilizing the results to inform actual COVID-19 communication.

Ultimately, as modern society is already sharply polarized along political party lines, de-polarizing news language and tweets about a national pandemic like COVID-19 ensures that vital information reaches the largest audience possible. Thus, we believe that our research has potential for positive societal impact by explicitly suggesting extreme sentiment words to replace with more impartial phrasing when announcing research-backed guidelines and information to benefit public health. However, extra care should be exercised to ensure that our polarization results are not conversely used to exacerbate existing divisions, or drive distrust in information from reputable sources.