A Dataset of Covid-19 Tweets Surrounding Face Masks for Sentiment Analysis

Jae Ro, James Howe
Georgia Institute of Technology

Abstract

It is no understatement that the 2019 novel coronavirus (COVID-19) has not only taken its toll on the health of the American public, but it has also caused a major social shift toward online interactions as well. Accordingly, one such outlet for these online interactions as well as the spread of information regarding this pandemic has been Twitter. The goal of this study is to expand upon the research done by Chen et. al in gathering Twitter data related to COVID-19 by more specifically analyzing COVID-19 Tweets surrounding the topic of face masks. Are face masks helpful in preventing one from catching the virus? Should face masks be worn by non-healthcare professionals? Are face masks going to help decrease the spread of this virus on a public health level? These types of questions have been floating around throughout the past few months. As a result, through this study we attempt to
1) Create a Twitter dataset containing tweets related to face masks and COVID-19, 2) Observe high-level trends in tweets within this dataset, and 3) Analyze public sentiment about face masks (and its potential shift) amidst this pandemic over time.

I. Introduction

The 2019 novel coronavirus (COVID-19), had an unprecedented global impact. The virus had been declared to be a global health emergency as of January 30, 2020 by the WHO, and at the time of this writing, the virus has significantly affected people from around the world. Many people from countries in East Asia have long been accustomed to the regular use of face masks, but until now this has not been the case in countries like the US or the UK. As information and public perception of the situation has changed over time, so too have people's opinions of the matter, including views on face-masks. Especially at a time when people are discouraged from going-out in public, opinions are being expressed via social media outlets like Twitter and Facebook. This paper will discuss a study that has been conducted to analyze the discourse surrounding mask-use on Twitter across time starting from January 21, 2020 until April 10, 2021.

This study will analyze the volume of both mask-related tweets as well as COVID-19 tweets at large within a random sample between the months of January and April. This analysis will include discussion of the percentage of COVID-19 tweets that discuss masks over time, and the sentiment expressed in tweets towards masks over time. In order to collect the data from Twitter necessary for doing this study, we decided to make use of the COVID-19-TweetIDs dataset proposed by Chen et al. This dataset is part of an ongoing project that started on January 28, 2020, and is set out to construct a largely comprehensive dataset of all TweetIDs that correspond to Tweets about the virus. While the project started on January 28, data was retro-actively collected from one week

prior, making the earliest Tweets in the dataset dated back to January 21. To date, the dataset consists of 94,671,486 tweets spanning across multiple languages, and is stored as IDs. These IDs each map to specific posts on Twitter, and require the use of a Twitter API key to hydrate IDs into tweets stored as jsonl files. The IDs themselves have been captured from all hours of the day, and are spread across 1,880 text files.

Due to processing and memory limitations, it was decided to take a random sample of approximately 10% of the total COVID-19-TweetIDs dataset. Sampling was conducted by iterating over each of the 1,880 text files, randomly shuffling the IDs present in the given file, and overwriting the file with only the floor of the first 10% of IDs. The resulting sample dataset used in this study consisted of approximately 9,400,000 tweets which were hydrated from their respective IDs. It should be noted that since no information is known about a given tweet until it is hydrated, additional data points were later filtered during pre-processing. Still, since all processes were handled consistently, and since the data was sampled randomly, we believe the final version of our sampled dataset is representative of the original 94,671,486 tweets.

Prior to performing computation on the tweets, the data was pre-processed. Any tweets not written in English were removed from the set so as to reduce variability. We believe that if all tweets in our dataset are in the same language, it will lead to better models. In theory if the same word is written differently in different languages, it can lead to the model having duplicate word embedding vectors, and since the two languages would have minimal overlap, the two words would almost never be seen together and the corresponding embeddings would be very different. The majority of the tweets in the data were written in English, comprising about 66.66% of the whole. The second most common language in the dataset was Spanish, comprising 10.93%. All other languages in the data were each less than 4% of the whole. With English being the most prevalent language in the set, it was determined that the data should be constrained to only include tweets in English. Any cases of a tweet appearing more than once in the dataset from the same user was removed, and all tweets were set to lowercase. We also removed numbers and most punctuation marks, as these would not contribute to sentiment overall.

Since our sampled and pre-processed version of the COVID-19-TweetIDs dataset was not labelled with sentiment values, we were presented with a couple different options of how to proceed. The initial intention was to utilize an unsupervised k-means clustering algorithm to see if the data could naturally cluster into groupings based on sentiment or some other metrics. However, we instead opted to train a supervised learning model using a separate labelled dataset of tweets. By training the model on such a dataset, we would hope that the model could generalize to accurately classify the sentiment of the COVID-19 tweets. To train this model, we leveraged the Twitter US Airline Sentiment dataset that consists of 14,500 tweets labelled for sentiment. The data in this dataset was labelled as either positive, negative, or neutral. For the sake of keeping our analysis binary, we re-labelled the data as being either negative or not-negative, by merging the positive and neutral classes together. We believed this dataset would help to produce a strong model as the types of tweets present in the dataset appear to use very common and generalizable language. We predicted that if a model can accurately predict the sentiment

of this dataset within some degree of confidence, then it would generalize to be able to accurately predict the content of our COVID-19 dataset.

II. Methodology

As mentioned previously, the initial dataset of Tweet IDs consisted of over 94 million tweet IDs that, when hydrated, would contain raw data in several different languages. Due to hardware limitations, 10% of this dataset was sampled prior to hydration via a Python script. The script would shuffle the IDs present in each of the 1880 text files in the IDs dataset, then overwrite each file with the first 10% of the respective IDs. Sampling in this way ensured that tweets were randomly sampled out of the total 94 million while still maintaining the represpective volume of tweets per day. Note that while approximately 10% of the IDs were included in our sample, this would later be reduced further in volume during later steps to clean the data

As is with any practical machine learning or natural language processing problem, clean data that can be used right out of the box is often sparse and hard to come by. Furthermore, given that this is currently a relevant and ongoing pandemic, the dataset compiled by Emily Chen et.al was raw Twitter data made up of various different languages, emojis, links, etc. In light of this, much of our effort was put into cleaning this raw twitter Covid-19 dataset and creating a cleaned twitter dataset surrounding the topics of both Covid-19 and face masks that can be used for various sentiment or text classification models.

To start the processing of the data we chose to limit the data to only include tweets that contained the word “mask”. At this point we made the assumption that Covid-19 related tweets that contained the word “mask” were surrounding the topic of face-masks and their relationship to this virus. Upon completion of this filtering, we found that about 2-3% of the sampled Covid-19 tweets contained the word ‘mask’. It should be noted that though it was reported that 66% of the entire dataset was reported as being in English by Emily Chen et.al, we were not able to confidently maintain this proportion. As a result, we moved to filter only tweets composed primarily of English from our Covid-19 Mask data. This task proved to be more challenging than originally planned due to the small amount of text that was often found in a given tweet (ex. “hi @so&so” → classified as “Swedish” by various language detection libraries). Therefore, to accomplish this task, we used an ensemble approach with python’s polyglot and langdetect modules. If both of these language detectors were not able to detect english with the text of a given tweet, then the tweet was left out. Moving forward, the next step was to remove all unnecessary characters and tokens that would not significantly contribute to the overall sentiment of a given tweet. Tweets were cleaned to remove punctuation (ex. @mentions, and #hashtags) while keeping the content following these symbols. Urls were then parsed and taken out using regular expressions. The tweets that contained contractions were then expanded (‘can’t’ → ‘can not’). Accordingly, texting slang was then expanded as well to capture the meaning and sentiment within the text (ex. ‘rly’ → ‘really’). Finally, emojis were parsed and replaced in the tweet with the text equivalent definition (ex. 😀 → ‘smiling face with open mouth’) and all characters within the tweet were converted to lowercase. Due to the

large amount of data and the lack of available free computational resources, the data was cleaned and processed in batches by month. After all processing was complete the combined sampled and cleaned data was made up of 124,277 instances; about 1.3% of the 10% sampled Covid-tweets (number does not account for tweets within this sample that were non-english).

In addition, in order to build our sentiment prediction model, we needed a labeled sentiment data set to train our model on. The data set we decided to use was the *Twitter U.S. Airline Review* dataset taken from Appens Open Source Datasets (used in a 2015 Kaggle competition for sentiment classification). These tweets were then cleaned using the same approach mentioned above in order to allow some form of translation between it and our Covid-19 mask dataset. This dataset was originally comprised of 14,500 tweets that were labeled as either Negative, Positive, or Neutral in sentiment. After running this data through our processing pipeline, we were left with a cleaned dataset of 13,200 data points. Accordingly, we made the decision to turn this multi-class classification problem into a binary classification. As a result, the labels were renamed as either “Negative” or “Non-Negative” (made up of both positive and neutral sentiment labels). This decision was made with the hypothesis that a classifier that can predict negative sentiment with high accuracy on a data set taken from a different context (airline reviews) would be able to perform better on our covid-19 mask twitter dataset than a classifier that was trained to classify negative, positive, and neutral sentiment.

Once the preprocessing and cleaning of the datasets was completed, the next phase in our study involved performing exploratory descriptive analysis of our Covid-19 Mask Twitter dataset. We collected high-level observations about the mask data such as the total volume of tweets/day that mentioned masks, most frequently used words in the mask data, most retweeted tweets mentioned masks, and the content of tweets themselves (further discussed in the analysis section of this paper).

Finishing up the exploratory descriptive analysis, we moved to training our sentiment predictor model. After experimenting with various word/document embedding generators and supervised learning models, we chose to build our predictor using an LSTM (Long-term-Short-term Feed-orward Recurrent Neural Network) that took input in the form of a fixed-length, padded TF-IDF encoded vector and output in the form of a fixed-length vector representing the possible sentiment labels. This model was trained on 80% of the airline review twitter data and then validated on the remaining 20% of the data. Our LSTM network consisted of an embedding layer with a vocab size of 4000, an embedding dimension of 128, a one-dimensional spatial dropout layer with a dropout rate of 40%, and LSTM with 20% dropout rate and a dense softmax activation layer with 2 classes (Negative and Non-Negative). This model was then compiled with a loss metric of Categorical_crossentropy, ADAM optimizer, and performance metric of accuracy. This model was then trained over 10 epochs with a batch size of 32. After fine-tuning our hyperparameters to what you see above, the model was then used to experiment on our Covid-19 Mask Twitter dataset. The initial goal was to see if our model that was trained on sentiment labeled Airline Twitter Reviews would be able to make meaningful predictions on the sentiment of Covid-19 tweets surrounding the topic of face masks. The Negative or Non-Negative sentiments that we were aiming to predict were in perspective of topics

such as whether masks can actually help everyday citizens in this pandemic or not, whether they should be used by non-health professionals, whether they can prevent the spread of the virus, etc. Finally, after our model output its sentiment predictions, the results were qualitatively and quantitatively analyzed according to correctness of sentiment and proportion of negative to non-negative sentiments. Due to the fact that this data set was not originally labeled, but rather formed through this study, percent accuracy of our model on this data set was unable to be calculated. The results of this study were then compiled and discussed in the following sections.

III. Results

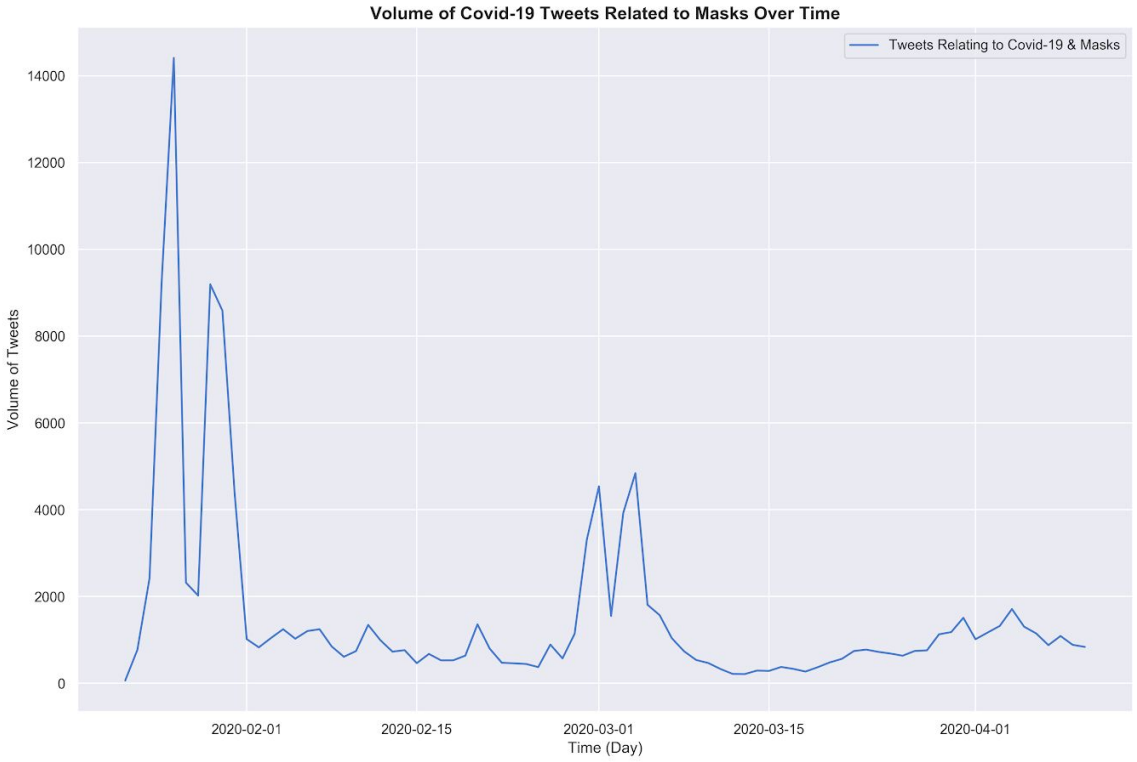

Figure 1: Volume of Covid-10 Tweets Related to Masks Over Time

The figure above illustrates the volume of COVID-19 tweets in the sample that mentioned masks that were posted each day from January 21 to April 10.

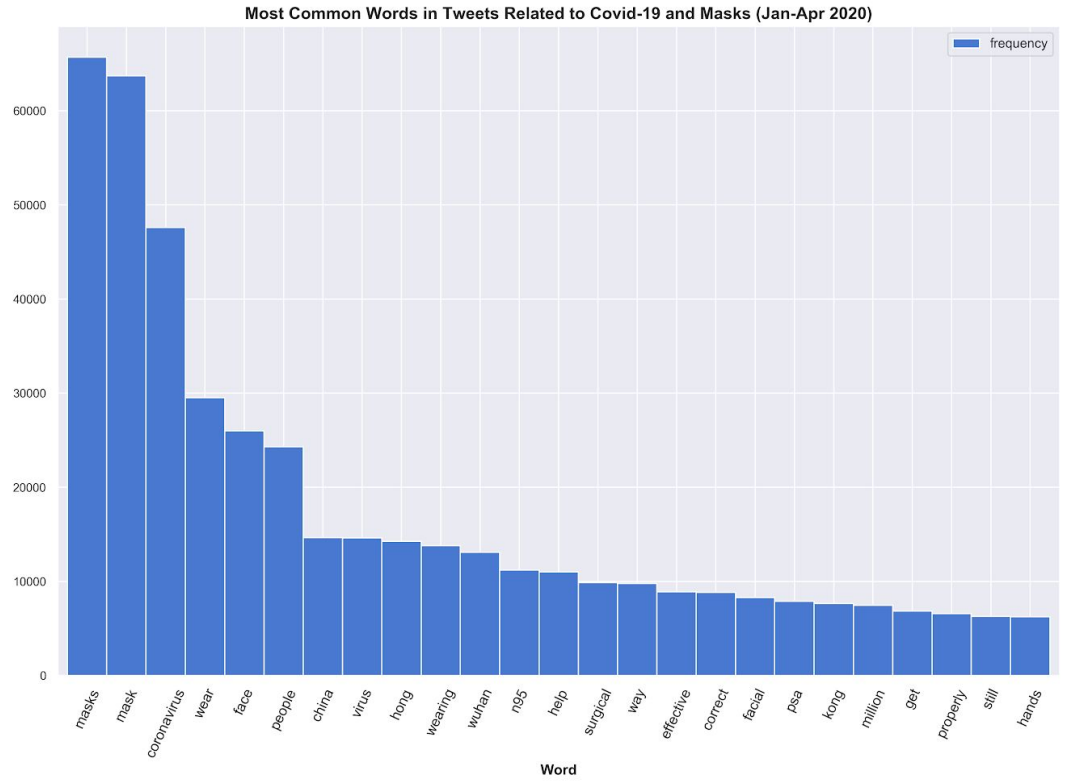

Figure 2: Most Common Words in Tweets Related to Covid-19 and Masks (Jan-Apr 2020)

The figure above illustrates the 25 most frequently used words in our sample of COVID-19 tweets that mention masks. Note that stop-words were excluded when finding these word-frequencies.

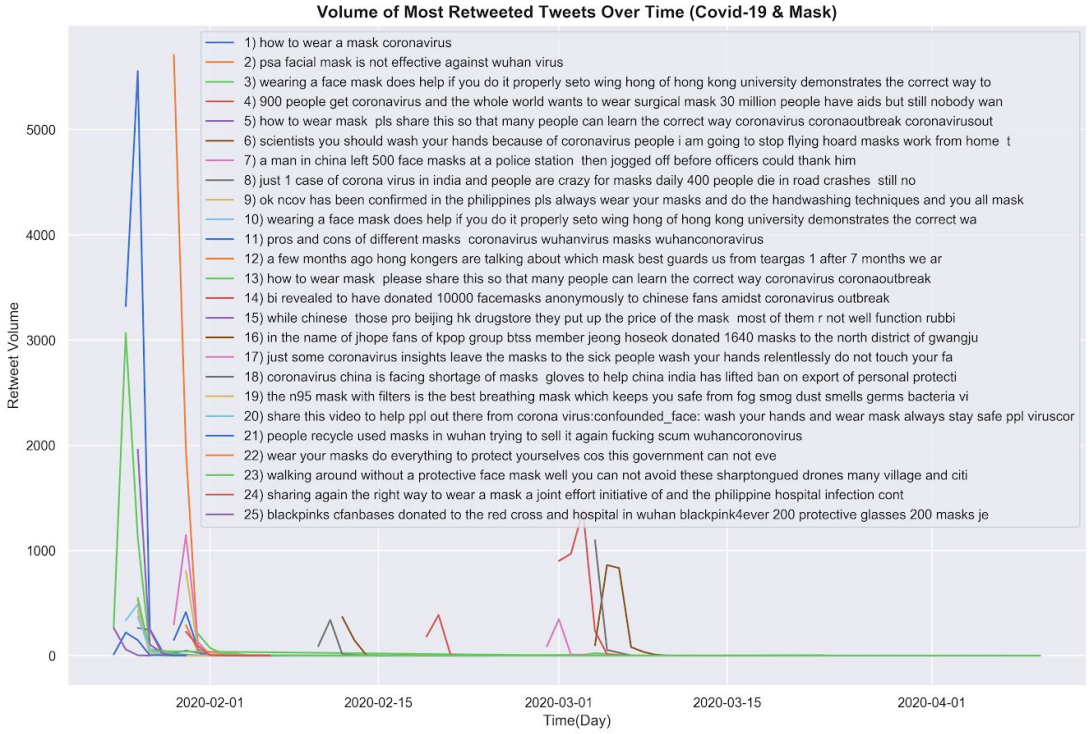

Figure 3: Volume of Most Retweeted Tweets Over Time (Covid-19 & Mask)

The figure above illustrates the 25 most retweeted COVID-19 tweets in our sample between January 21 and April 10 that mention masks.

The LSTM model that had been trained on the Twitter US Airline Sentiment dataset had achieved an in-sample accuracy of approximately 96.35%, and an out-of-sample accuracy of approximately 81.60%. The following table illustrates some of the predictions made by the model with the out-of-sample Airline Sentiment data.

predictedLabel	actual	generatedCleanTweets
Non-Negative	Non-Negative	thank you excited to be working with you guys
Negative	Negative	thank you for letting me luggage is still in denver but i am in phl neveragain disappointed
Negative	Negative	for our delays i am out of more money because of you
Non-Negative	Non-Negative	so the fares i see for flights in fall are the lowest they will be
Negative	Negative	understand weather is an issue but on time cancelled flighted reinstated cancelled flighted with...
Negative	Negative	why can not i find a cheap flight from dc to st louis the prices went up like crazy for april
Negative	Negative	service today missed my connecting flight then the customer service desk was terrible to me real...
Negative	Negative	status shows delayed it was just waiting for takeoff so did it depart it certainly did not at
Negative	Negative	i booked it on us airways site do not see a cancelled flight link
Negative	Negative	what is happening with the flight from fll to sfo why the delay and the reroute
Non-Negative	Negative	thank you for 7 hrs at terminal d in dulles airport
Non-Negative	Negative	make this delay go away maybe upgrade me and seats we are headed to columbus
Non-Negative	Non-Negative	awesome deals for only 39 each way
Negative	Negative	we waited 40 min for our bags after a 45 min flight
Negative	Non-Negative	have clients with an 11 hr layover at iah during the day will they have to claim recheck luggage...
Negative	Negative	again horrible service again attitude when asked for information again you make me not want to f...
Negative	Negative	if a business decision is made that and possibly causes lost customers is it a good business dec...
Negative	Negative	the departure time keeps getting late flightr i will be lucky if i am home by 3am
Negative	Negative	you have got to be kidding me no info on from dca to and your phone line tells me to call back l...
Negative	Negative	that is understandable my issue is with a new flight without the personnel to do changed my plan...
Non-Negative	Non-Negative	may i have my companion pass please

Table 1: LSTM Model Predictions of Out-of-Sample Airline Review Tweets Compared to Actual Labels

The following table shows 5 examples predictions made by the trained model on pre-processed mask-related COVID-19 tweets.

to wear or not to wear mask in china either way cancer is waiting chinapneumonia wuhancoronavirus ...	Non-Negative
after tons of criticisms cathaypacific finally announce that staff can wear masks banning masks was really crazy oh gosh who made that silly decision before wuhanpneumonia wuhancoronavirus chinavirus chinesepneumonia chinavirus sars2	Negative
to wear or not to wear mask in china either way cancer is waiting chinapneumonia wuhancoronavirus ...	Non-Negative
flight attendants demanding to wear face masks on all flights worldwide as wuhancoronavirus spreads hongkong	Negative
to wear or not to wear mask in china either way cancer is waiting chinapneumonia wuhancoronavirus ...	Non-Negative

Table 2: LSTM Model Predictions on Sample of Mask-Related COVID-19 Tweets

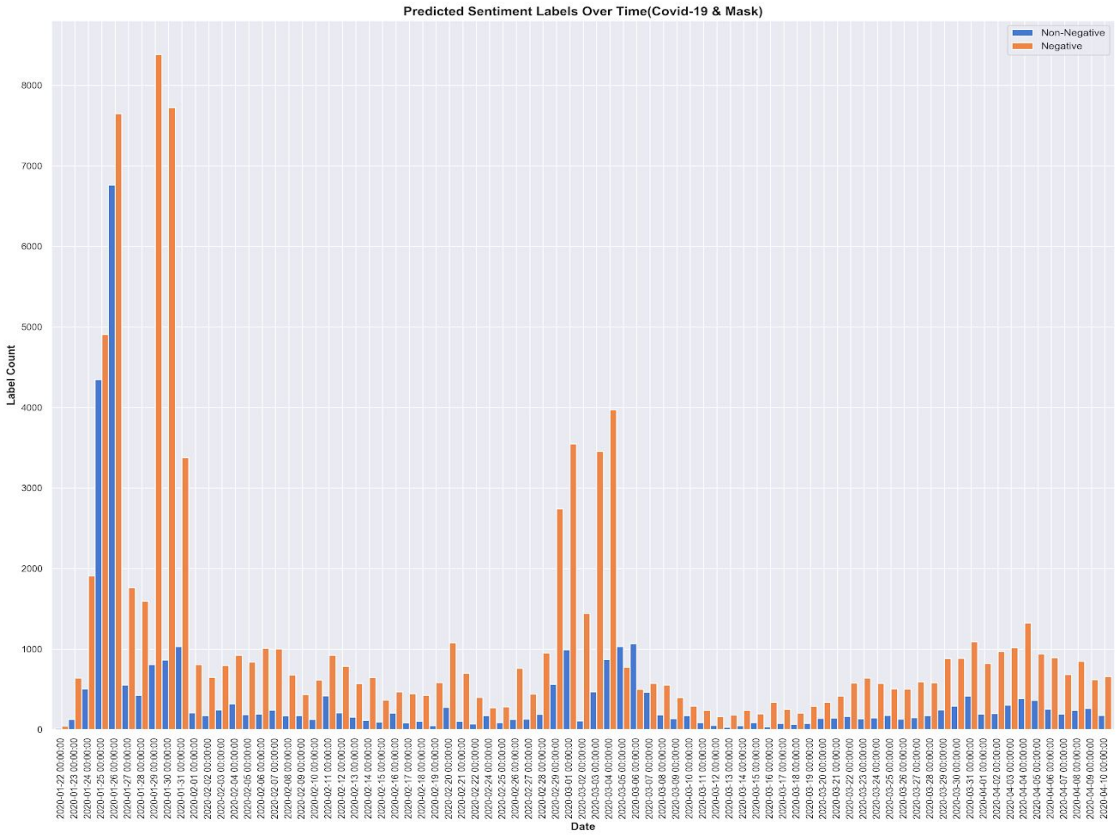

Figure 4: Predicted Sentiment Labels Over Time (Covid-19 & Mask)

The figure above shows the volume of tweets labeled by the model as either not-negative (shown in blue) or negative (shown in orange) when predicting sentiment on the mask-related COVID-19 data.

IV. Discussion

From Figure 1 we can see that the most volume of mask-related tweets occurred during late January, coinciding with the declaration of the virus as a global health emergency. Within our sample, the day with the greatest number of mask-related tweets saw over 14,000 tweets. Since our sample of the data after sampling and cleaning consisted of 124,277 tweets, this means that approximately 11% of the observed mask tweets occured on this single day. Figure 1 also shows that a small spike in mask commentary occurred in early March.

While it had been observed that the overall volume of COVID-19 related tweets drastically increased from January to February, and then even more so in March, we have not observed the same trend with respect to mask-related tweets specifically. We had expected that these numbers would generally scale together with the idea that people would have more to say about masks as concerns surrounding the virus rose, but the data would suggest otherwise. It is possible that further analysis beyond the topic of masks could reveal the main topic or topics of discussion during this time interval.

Figure 2 generally shows that Zipf's law is followed with respect to the 25 most frequent words in our sample. Again, note that filtered stop words were not included in this analysis of word frequencies. Considering that this sample had been obtained by filtering

the COVID-19 sample for mentions about masks, it is expected that we see the most frequently occuring word(s) is/are “mask(s)”. Likewise, the next most commonly occuring word is “coronavirus” which can be explained by the general fact that this dataset is about all tweets discussing the virus. With little exception, the top 25 occuring words represent words relating to “China” and wearing facial masks.

Figure 3 examines the 25 most retweeted tweets present in our sample. Note that the peaks in volumes seen in Figure 1 are consistent with the trends in peaks reached by the most retweeted tweets. This is likely explained by overlaps between these two plots. That is, the peaks in volumes of mask-tweets are most likely the result of retweets of the same tweets.

The most retweeted mask-related tweet is fittingly about how to wear a face mask. Considering that the peak of this retweet occured around the time that the virus was declared to be a global health emergency, it makes sense that many people would be interested in wanting to know how to properly wear masks. The second and third most retweeted tweets are interesting in that they express opposing thoughts and opinions regarding the use of face masks. When the coronavirus was initially beginning to be discussed publicly around the world, the American CDC had advised Americans only to wear facial masks if they were sick or if they work in medical environments. This recommendation was later changed, but at the time, some believe this advised practice from the CDC led to people questioning whether or not face masks are effective at preventing infection from COVID-19. This could explain why we see discourse surrounding this topic in the most retweeted posts.

As discussed in previous sections, an LSTM model had been trained using a separate labelled dataset of tweets focused on airline reviews. Across trials, this model achieved 81%-82% out-of sample accuracy. On a class-by-class basis in the out-of-sample data, the model classified 73.84% of the non-negative tweets correctly and 85.33% of the negative tweets correctly. While the model generally performed well overall, this does reveal that it was more effective at accurately recognizing negative tweets than others. Some of the predictions made by the model with this dataset are present in Table 1. This subset of predictions generally matches the actual labels, but there are some exceptions. One of these exceptions occurs with the cleaned tweet “thank you for 7 hrs at terminal dulles airport.” This example shows how the model struggles with cases of sarcasm, as this message sounds like it could be positive, but it actually implies a negativity due to a long flight layover.

Since our COVID-19 dataset did not already have sentiment labels, we attempted to employ transfer learning with the LSTM model from the airline reviews to the tweets about COVID-19. In the future, we plan to experiment with other techniques to classify the tweets, and this will likely include manually labelling a portion of the COVID-19 tweets with sentiment labels. However, since we had not done that for the experiment at hand, we assumed with some degree of confidence below 81% that the LSTM model could predict sentiment with the COVID-19 data. Some of these predictions can be seen in Table 2. This table shows a couple negative tweets that the model seems to have labelled corrected, but it also shows a retweeted tweet that is more ambiguous. The latter tweet, “to wear or not to wear mask in china either way cancer is waiting...”, actually reveals a

nuance to be considered in future work. The model has labelled this tweet as being non-negative. In a way this is true, and in a way this it is not. We perceive this to be a negative tweet overall, but it worth noting that the tweet is neutral toward masks themselves. Future work will refine the sentiment analysis conducted to be more precise in how sentiment is assessed.

Nevertheless, we will be assuming that the predicted sentiment values for the mask-related COVID-19 tweets to be accurate overall with approximately 81% confidence. Although, we could increase this confidence upon labeling the data manually and then evaluating the accuracy. With that said, Figure 4 conveys that each day in the study observed more negative tweets than not, with very few exceptions. This suggests that the overall discourse surrounding masks has been negative. As a disclaimer, it is possible that the greater number of negative predictions is related to the bias of the model discussed above to better classify negative data.

V. Future Research

Through this study, we tested various machine learning models in attempting to find a model that would be appropriate for predicting the sentiment of tweets regarding COVID-19 and face masks. From word/document embedding models such as word2vec and doc2vec to traditional RNNs and then finally to LSTMs, this study has progressed through the history of NLP models. Furthermore, we have been able to understand the strengths of each of these models as well as the major shortcomings that have led to the newer models. As a result, one area of future research would be to use BERT (Bidirectional Encoder Representations from Transformers) to predict the sentiment of our COVID-19 Facemask Twitter dataset instead of our LSTM-based model. We believe that this switch from a unidirectional feed-forward neural network (relevant only to the context of its training data) to a bi-directional model that is pre trained on the Wikipedia corpus will allow for a much higher Transfer Learning rate, and as a result will be able to formulate more accurate sentiment predictions. Furthermore, through this study we found that the Tweets that were in support of the use of face masks in preventing the spread of COVID-19 were classified as Negative in sentiment because they were criticizing those who didn't wear these masks outside. As a result, we propose to expand our study to analyze sentiment of support For or Against the use of face masks throughout the Tweet history of this pandemic instead of just analyzing general Negative and Non-Negative sentiment

VI. Conclusion

This study has examined a small portion of the full COVID-19 TweetIDs dataset. By sampling said dataset and pre-processing the sample to include a cleaner version of the selected tweets, we have created a more approachable dataset that can be further explored for insights regarding mask-related Twitter discourse during the global pandemic. The methodology present here can be applied to other subtopics of COVID-19 to explore other areas. We have assessed general trends of mask-related tweets, and have now recognized several ways to further this work. As this research is continued, the

groundwork set forth in this paper will serve as a baseline to better understand public sentiment during this global pandemic.

VII. Bibliography

https://www.kaggle.com/crowdflower/twitter-airline-sentiment
https://www.aclweb.org/anthology/W16-4903.pdf
https://www.bioinf.jku.at/publications/older/2604.pdf
https://arxiv.org/pdf/2003.13907.pdf
https://arxiv.org/abs/1607.05368
https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005110
https://arxiv.org/pdf/2003.10359.pdf
https://dl.acm.org/doi/abs/10.1145/2766462.2767830
https://arxiv.org/pdf/1810.04805.pdf
https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewPaper/2857