Idalina Sachango, Nivedita Vatsa, Dharini Ramaswamy, Akila Forde

CAPP 30254: Text Analysis of COVID-19 Mask-Related Tweets

Objective

Given the loud and polarized opinions around mask-wearing, ranging from necessity to curb the COVID-19 pandemic and discussions about civil liberties, especially with the striking of the transportation mask mandate in April 2022, we decided to employ textual analysis techniques to analyze sentiment towards masks. We closely followed the work of Sander et al. and Wang et al., to understand the process of constructing a sentiment analysis algorithm. Blei et al. informed our understanding of the Latent Dirichlet Allocation model that we discuss in later sections.

Data Overview

Text from tweets form the basis of our analysis. To obtain the data, we first downloaded a series of tweet IDs associated with COVID-19. These tweet IDs are published by the Panacea Lab at Georgia State University. We then used the Twitter API and the `twarc` Python library to obtain the full text of the tweets. This process is known as rehydration.
There exist millions of tweets about the pandemic. So to narrow our search, we made a few decisions:

- We decided to focus on certain key dates instead of sampling all available dates. Our selection of dates is based on the timing of known historical events such as the start of the CDC mask mandate, Donald Trump contracting COVID-19, the introduction of the vaccine, President Biden's election / inauguration, etc.
- We decided to filter for tweets that discuss masking. Currently, our identifying keywords are "mask", "face covering", "mouth covering", "N95", "face diaper", and other similar terms.
- We focus only on English tweets and ignore retweets. Information supplied by the Panacea Lab allows for this type of filtering.

As a result of this filtering process, we went from 2+ million tweets and ultimately acquired a sample of approximately 84,252
tweets, ranging from April 2020-April 2022.

Exploratory Data Analysis

Our exploratory analysis focused on word and hashtag frequencies. For words, we stratified the tweets by "events" (e.g., onset of the pandemic, elections, etc.) and found that different words were more salient. This was a promising finding and led us to believe that topic modeling would produce meaningful results. Please see the appendix for our word clouds.

We view hashtags as a crude tag for topics. Analyzing the most frequent hashtags shows a general pro-mask sentiment, but also shows that masking is intertwined with political views too. Again, please see the appendix.

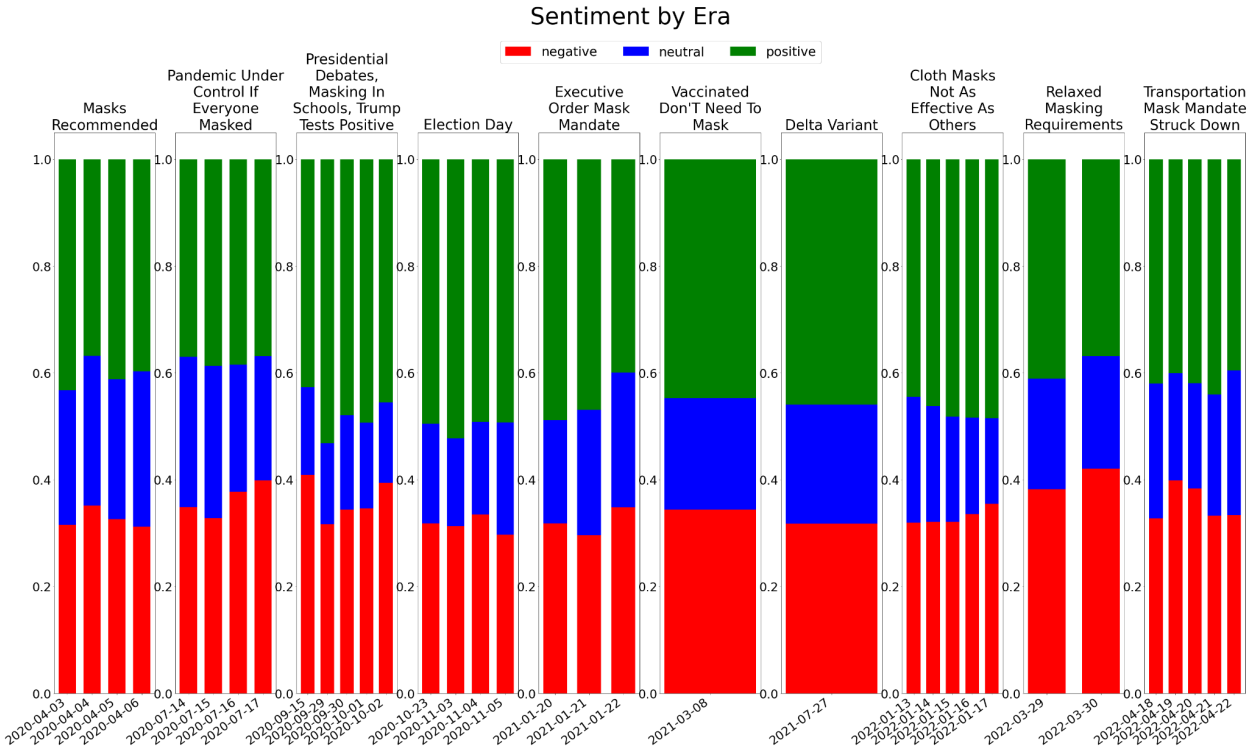
- Hashtags are a shortcut to identifying topics and sentiment; reveal a pro-mask stance peppered with political opinions

Sentiment analysis using VADER

One of the approaches we took in terms of analyzing the tweets was the use of VADER (Valence Aware Dictionary and sEntiment Reasoner), a lexicon and rule-based tool attuned to sentiments expressed in social media. This allowed us to categorize tweets by tonality, emoji-use and punctuation. Tweets were categorized into the following bins based on their compound score as calculated by **nltk.vaderSentiment**'s polarity score function as below:

- positive sentiment: compound score >= 0.05
- neutral sentiment: (compound score > -0.05) and (compound score < 0.05)
- negative sentiment: compound score <= -0.05

Given the snapshot nature of our data, we categorized specific dates into eras as seen in the figure below, to illustrate the movement of sentiment around certain events. By and large, there are marginal shifts in sentiment over time, with positive tweets on average accounting for a little less than 50% of all tweets. The largest positive shares of tweets coinciding with politically-related events such as the Presidential Debates, Election Day and the Executive Order Mask Mandate.



However, VADER calculates tweet sentiment as a heuristic score of tonality, context is not taken into account so positive tweets may not be exactly positively referencing masks, the same with negative tweets. For example the following text was categorized as positive in reference to Donald Trump's COVID-19 Diagnosis in October 2020, despite being sarcastic and ironic in nature: "I hope he wore "the biggest mask" he's ever seen these past few days. #TrumpHasCovid #BidenHarris2020 #COVID19 #WearAMask." While VADER may not be too insightful on opinions, it does inform our further analysis on topic modeling, by understanding which topics are spoken with more positive or more negative language.

Data pre-processing: cleaning and tokenizing tweets

To process the tweets, we converted all text to lowercase, removed websites (indicated by "https" or "www"), removed Twitter mentions (beginning with "@"), and other special characters. Note that we chose not to remove hashtags (beginning with "#") because we felt that these tokens would be informative (e.g., "stayhome" and "covidisnotover"). The data cleaning process was iterative and after our first pass, we noticed that words with numbers (e.g., "aa642a") were increasing the size of the word list, without adding value. So we chose to remove all words with a digit, but allowed some exceptions such as "n95" and "sarscov2," among others.

We then remove stopwords, many of which were included in the **`nltk`** library. After a few iterations, we also removed "amp" which often appears in tweets to indicate ampersands. Finally, to create a bag of words, we tokenized the text (*i.e.*, extracted individual words) and lemmatized each token (*i.e.*, converted them to their root form). In our initial pass, we used the **`nltk`** library (**`WordNet`**) to lemmatize, but after some trial and error, we decided to use **`SpaCy`**. Even though it took much longer to run, we found that it performed much better. See examples below:

keyword:	**masks**	**masking**	**masked**	**quarantined**	**distancing**	**distanced**
`nltk`	mask	masking	masked	quarantined	distancing	distanced
`SpaCy`	mask	mask	mask	quarantine	distance	distance

K-means

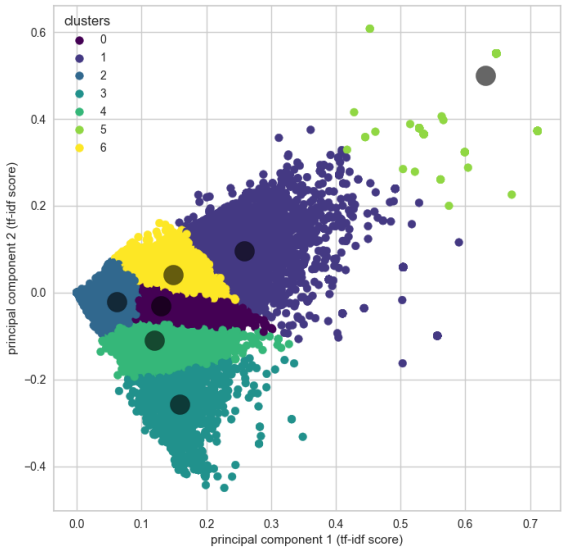
Our first attempt at building a topic model was to pass our processed data into a K-means model which we further use to validate our Latent Dirichlet Model (LDA) by comparing the initial results from K-means to the results from LDA. [[why maybe?]]

For this topic modeling algorithm, we began by finding the tf-idf scores of each word in the corpus. We used **`sklearn`**'s **`TfidfVectorizer`** to do this; it takes the entire corpus of tweets and computes the tf-idf scores of each word per document and vectorizes it with the result being *n* feature vectors corresponding to the *n* distinct words in the corpus. Each entry in these feature vectors corresponds to an observation, or tweet, in our dataset. The interpretation being that each feature has a tf-idf score calculated for each tweet in our set. The result of vectorizing our corpus gave us a matrix in high dimensional space - 73395 feature vectors corresponding to

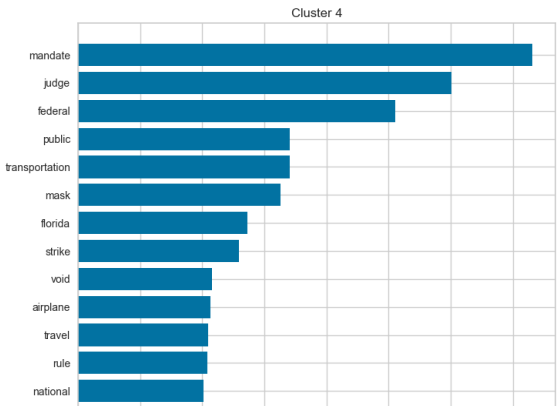
73395 distinct words. We applied a dimensionality reduction algorithm to take this matrix and bring it down to 2-dimensional space. This allowed us to easily interpret the findings of our K-means model.

Next, we attempted to find the optimal number of clusters. We defined a function that takes as input a range of cluster numbers and our dimension-reduced matrix. For every number within that range, we initialize a k-means model with our data and compute its loss. We graph the loss associated with each cluster and pick the cluster that is associated with the local minima that corresponds to the 'kink' or 'elbow' in the curve of the graph. Because the cluster number associated with this minima was sensitive to change in our algorithm (expanding the range increased the optimal cluster number determined) we cross-checked with **`KElbowVisualizer`** from the **`yellowbrick`** package. The **`KElbowVisualizer`** consistently determined that the optimal number of clusters for our model was 7. Because it consistently predicted, we proceeded with this model's decision.

Using the **`kmeans`** class from **`sklearn`**'s **cluster** module, we fit our data and predicted the cluster label of each tweet in our corpus. By tf-idf score and color, here is the output of our model expressed visually:
Grabbing subsets of our tweet corpus associated with the clusters estimated from our k-means model, we were able to plot the top 15 words for each of our 7 clusters. The clusters were able to



pick out distinct words that hinted at what the topics output from our LDA model might be. Below is one example that seems to be associated with transportation and travel which was also identified in our LDA model.



Latent Dirichlet Allocation

Next, we decided to use an LDA approach to better tease out the topics in our corpus of tweets. Unlike k-means clustering which assigns each tweet to a single cluster, LDA assigns each tweet to a distribution of topics. We felt that this would be a useful technique for identifying different and frequently overlapping attitudes.

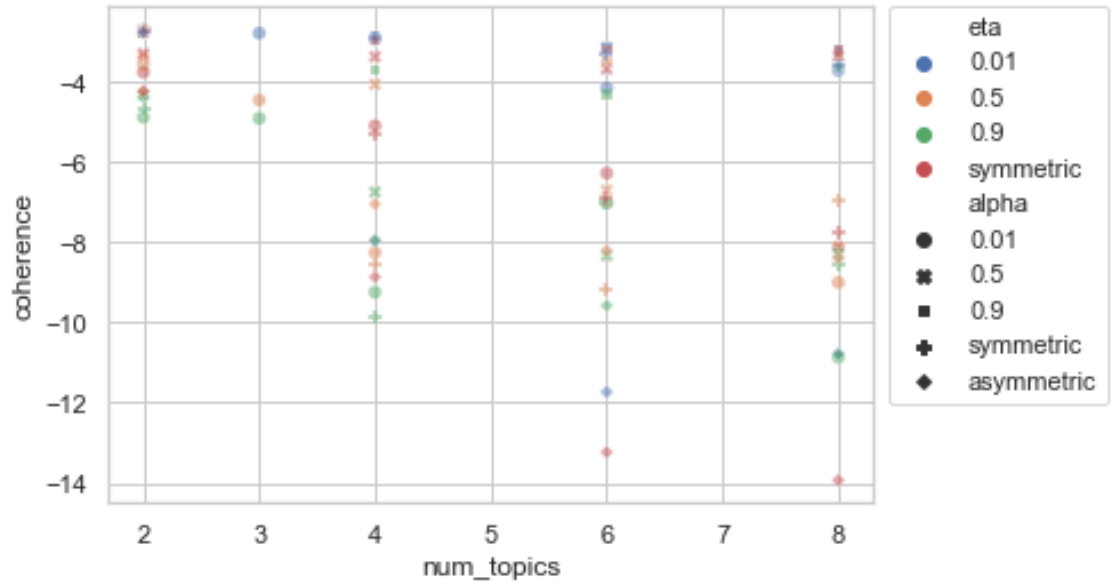
Similar to the k-means method, the researcher must identify a suitable number of topics. If k topics are specified, then the result of an LDA technique gives: (a) k topics defined by topic keywords and a corresponding weight, and (b) for each topic and each tweet, it generates a weight showing the association between the tweet and that particular topic.

In preparation for estimating the LDA model, first we converted each tweet into a "bag of words" (or tokens) using the pre-processing steps above. In our case, our tokens included bi-grams as they helped us cover a wider range of topics. For example, safety advisories frequently included the phrase "stay home" while political comments mentioned "white house." After a first pass, we decided to include tf-idf measures in our bag of words instead of absolute counts because this would allow us to assign lower weights to frequently occurring tokens. We also constructed a dictionary mapping each token to a unique identifier.

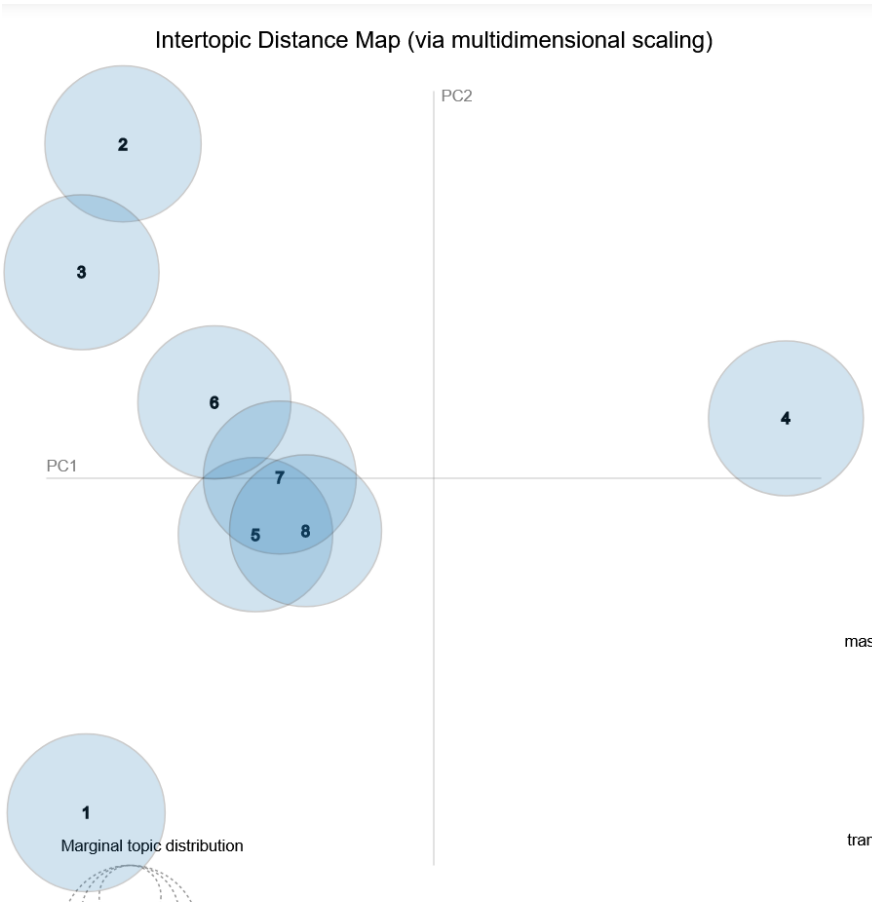
We first conducted our LDA analysis on the entire corpus of tweets. LDA models are defined by numerous hyperparameters. For this project, we focus on three: the number of topics (k), our prior assumption about the concentration of topics per tweet (α), and our prior assumption about the concentration of words per topic (η). We estimated 80+ combinations of these hyperparameters. Ideally, we would have done a grid search on not only these three parameters but several others as well (*e.g.*, the number of passes for model generation, number of documents in each training chunk, the extent to which extreme values ought to be removed, minimum count for bigrams, inclusion of 3+ grams, etc.). Due to limited computational resources, we focused on a set of 80 models, which took 15+ hours to run.

For the purposes of evaluating each model, we first generated a "coherence" measure. Specifically, we used the UMass coherence measure which captures how frequently the words within a given topic co-occur. Coherence scores do not strictly indicate the best models, but they help us select the better performing ones. UMass coherence scores are negative and values closer to zero are considered more coherent. The chart below shows each model by coherence score.[1] We manually evaluated a collection of models that had "higher" coherence scores.

[1] The 83 combinations of hyperparameters are all combinations of (4 topic counts $\times$ 5 values of α $\times$ 4 values of η). The reason for this "odd" number is that three of these models are related to when k = 3 (where α = 0.01 and η = 0.01, 0.05, 0.09). We stopped the operation early on k = 3 because in our first pass alone, we could see that higher topic counts led to more interpretable results. Since the models were estimated and ready to evaluate, we included them in this chart.



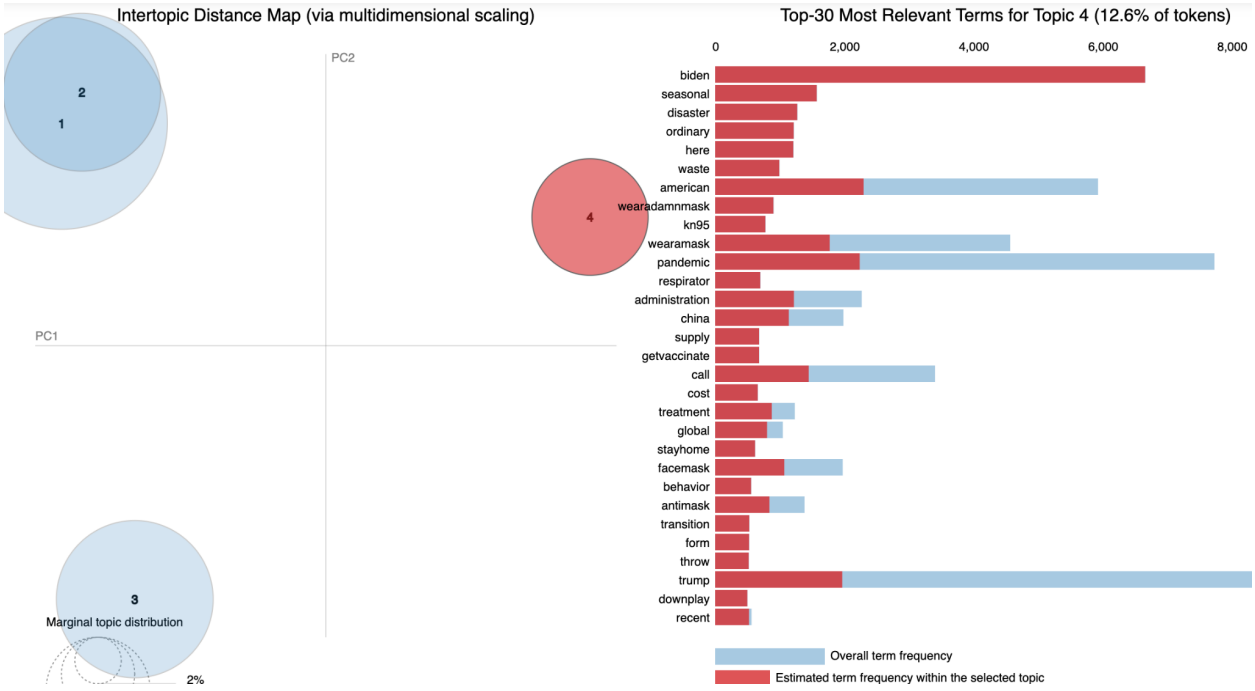
For each topic within a mode's result set, we printed a series of tweets for which that topic was "dominant" (had the highest weight). After reading these tweets, we tried to determine an overarching "theme" with the help of topic keywords (also supplied by the model). For smaller values of *k*, we were unable to identify a clear topic. We ultimately settled on a model with 8 topics. Reviewing some of these tweets, manually, we can see that some topics are well-defined. For example, topic 1 features tweets of political interest. This includes commentary on President Trump and other politicians as well as public health policy (e.g., school closures). In the appendix, we offer a closer look at each topic, including top keywords and example tweets. Some topics are less clearly defined. From the chart above, we can see that topics 5, 7, 8 show considerable overlap. For example, the top 20 keywords for these topics (appendix) have overlapping keywords. That said, if we focus on tweets with higher topics, we can see that topic 5 highlights masking skepticism, topic 7 mentions masking on transportation, and topic 8 engages with public impatience over masking (*e.g.*, COVID-19 is not over yet).



Topic modeling: Latent dirichlet allocation

In addition to running an LDA model on the overall set of cleaned tweets, we also conducted an LDA model on subsets of tweets that were classified with positive and negative sentiments using VADER. Because the data processing steps differed for the VADER and LDA models, we merged the subset of positive/negative sentiment tweets with the cleaned dataset prepared for the LDA model on tweet id. Unlike the overall LDA model, we omitted extremes by filtering tokens that appeared in less than 15 documents, more than 80% of the documents, and kept the 100,000 most frequently occurring tokens. Additionally, we ran a TF-IDF model on our corpus. For each bag of words in the corpus, we kept the tokens whose TF-IDF scores were above a threshold of 0.2. Due to this being a smaller subset, we ran LDA models with just four topics.

At first glance, it is difficult to easily glean insight and identify the topic categories from the output generated by the LDA model. However, when visualizing the topic categorization on an intertopic distance map below, we can better understand how topics were split. Within the negative sentiment tweets, we can see that in Topic 4, terms like “biden”, “trump”, “pandemic”, and “american” occurred most frequently. We can infer that this topic relates to government response and sentiment around both presidents’ pandemic-related policies.



Limitations & Further Work

We have identified three major limitations of our current work. First, the decision to choose Twitter as our primary data source comes with the repercussions of having a sample of individuals who are non-representative of the average US resident. According to Pew Research, Twitter users are young and skew to the political left. Additionally, 80 percent of all posts are driven by only 10 percent of Twitter users. Further research may include identifying sources that give us a more varied sample of individuals with different political leanings and more varied participation patterns.

Additionally, stratifying by sentiment might not be as informative as we would have hoped. Many tweets in our corpus were related to news, which can be tonally positive even if they're intended to be ideologically neutral. Stratification by sentiment is also sensitive to false negative classification and false positive classification, which further work must be mindful of.

Finally, instead of our snapshot approach, a larger and more chronologically continuous set of tweets would have helped explain changing sentiment over time.

An added technical limitation is that we had insufficient computing resources to conduct a thorough grid search and arrive at the strongest model.

Conclusion

Through completing this project, we learned the following:

- Importance of using multiple methods to validate choice of hyperparameters (coherence scores, investigating grid search, human judgment)
- Interpretation of models like LDA can be challenging and need adequate time
- When models are computationally expensive and struggle to provide results efficiently, we need to scale down our parameters or dataset so that we can move forward with interpretations

Lastly, our data suggests that sentiment towards masks in the time periods we studied were largely pro-mask tended to express concern over public and personal safety, voice political frustration, and seek information about masking protocols. Further testing should be done on validating our LDA model output by evaluating our model's performance on some secondary task - like document classification or information retrieval - or by estimating the probability of unseen documents given some training data.

References

Adam Hughs & Steven Wojcik. (2019). "Sizing Up Twitter Users." *Pew Research*, https://pewrsr.ch/38Y86Hb

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. (2003). "Latent dirichlet allocation." J. Mach. Learn. Res. 3, null (3/1/2003), 993–1022.

Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. (2009). "Evaluation methods for topic models. In Proceedings of the 26th Annual International Conference on Machine Learning (ICML '09)". *Association for Computing Machinery*, New York, NY, USA, 1105–1112. https://doi.org/10.1145/1553374.1553515

Hutto, C.J. & Gilbert, Eric. (2015). "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Proceedings of the 8th International Conference on Weblogs and Social Media," ICWSM 2014.

Lyu JC, Luli GK (2021). "Understanding the Public Discussion About the Centers for Disease Control and Prevention During the COVID-19 Pandemic Using Twitter Data: Text Mining Analysis" Study J Med Internet Res;23(2):e25108

Sanders, A. C., White, R. C., Severson, L. S., Ma, R., McQueen, R., Alcântara Paulo, H. C., Zhang, Y., Erickson, J. S., & Bennett, K. P. (2021). "Unmasking the conversation on masks: Natural language processing for topical sentiment analysis of COVID-19 Twitter discourse." *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, *2021*, 555–564.

Wang, H., Li, Y., Hutch, M., Naidech, A., & Luo, Y. (2021).“Using Tweets to Understand How COVID-19-Related Health Beliefs Are Affected in the Age of Social Media: Twitter Data Analysis Study.” *Journal of Medical Internet Research*, *23*(2), e26302. https://doi.org/10.2196/26302

Xue, J., Chen, J., Hu, R., Chen, C., Zheng, C., Su, Y., Zhu, T., “Twitter Discussions and Emotions About the COVID-19 Pandemic: Machine Learning Approach,” *Journal of Medical Internet Research*, *22*(11), e20550. https://www.jmir.org/2020/11/e20550

Yadav, Kajal (2021). “Text Clustering using K-means: Complete guide on a theoretical and practical understanding of K-means algorithm.” *Towards Data Science*, https://bit.ly/3xoB2QP

Appendix

Figure 1: Word Cloud by Selected Event

Figure 2: Top Hashtags

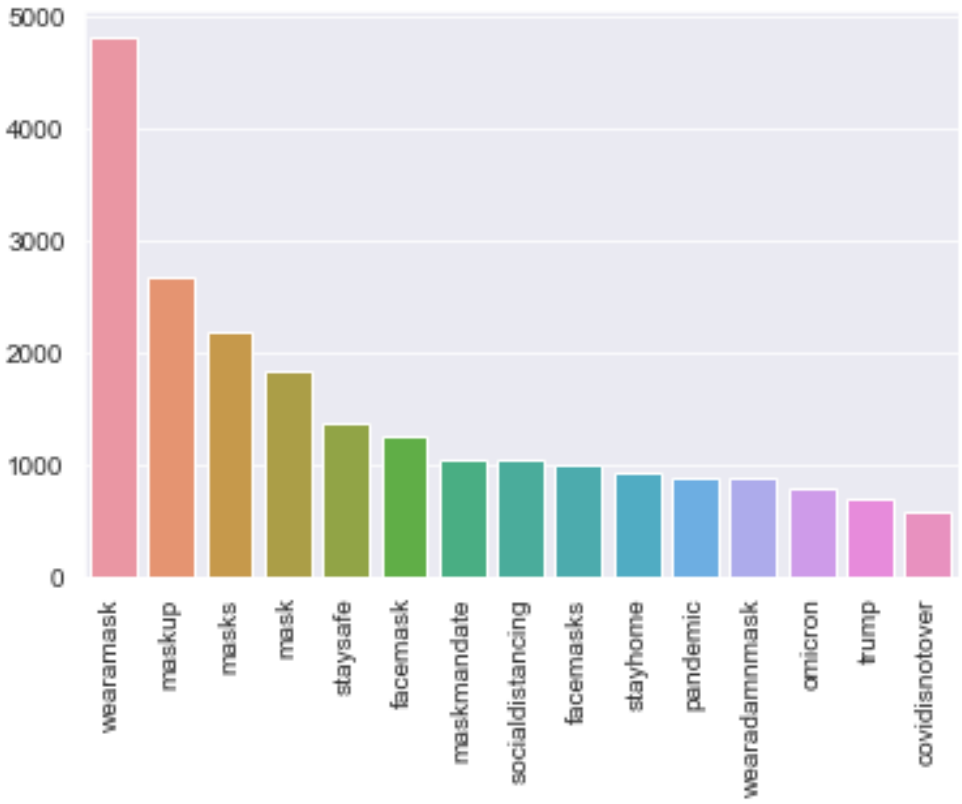


Table 1: Top 20 Topic Keywords

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
trump	face	virus	please	do	do	mandate	vaccine
testing	coronavirus	new	distance	wear	you	public	continue
school	see	covid	hand	work	people	cdc	everyone
president	covering	wearamask	spread	people	go	we	covid19

vaccination	respiratory	facemask	keep	say	wear	judge	back
first	cloth	case	stay	right	without	rule	order
positive	year	number	safe	think	know	travel	well
test	virus	update	social	protection	n95	federal	protect
would	kid	recommend	maskup	give	get	require	get
two	face_covering	live	help	want	make	health	free
crowd	shop	doctor	vaccinate	science	person	maskmandate	way
due	lift	hope	other	around	change	end	new
wear	life	coronavirus	prevent	die	good	plane	hospital
day	wear	covid19	take	one	mask	airline	follow
staff	cause	surge	protect	I	coronavirus	transportation	high
business	mean	child	wash	mask	nose	long	even
put	force	tell	social_distance	do_work	never	say	find
covid19	covid19	news	study	refuse	leave	drop	try
worn	infection	week	wear	variant	thing	mandatory	flu
american	million	city	cover	read	lie	rise	wear

Table 2: Manual Review and Labeling of Topics

LDA: k = 8, α = 0.9, η = symmetric
TOPIC 1: Politics

· COVID-19 Masks Are a Crime Against Humanity and Child Abuse · He's qualified to be a political hack, and might have gotten away with just dissing vapers & playing the kid card for #BigTobacco. But now that he's lied to the world about the value of wearing masks during this #COVID19 pandemic, he is totally screwed. Nobody will trust him now. · @DrThomasPaul Cuomo is the culprit. While having Covid19, he went outside to play in the park with son and didn't wear a mask😡. He has caused so many high Covid19 cases. He definitely needs contact tracing!! All of CNN should be quarantined!! · President Biden is putting into play his national COVID-19 strategy to ramp up vaccinations and testing, increase the use of masks and reopen schools and businesses
TOPIC 2: Directives on Masking at Sites of Commerce
· BBC News - Coronavirus: Face coverings in England's shops to be compulsory from 24 July · Delta Air Lines initially called covid-19 an ‘ordinary seasonal virus’ as mask mandate was lifted · We are relieved to see US #maskmandate lift to facilitate global travel as #COVID19 has transitioned to an ORDINARY SEASONAL VIRUS · Coronavirus: Face masks to become mandatory in shops
TOPIC 3: Tracking & Monitoring COVID-19 Spread
· Dr. Fauci is warning America again, into a #Covid19 ‘risk period’ My doctor team, recommend professional mask for you to buy channels, hope the virus will be defeated! Keep safe and social, keep away from viruses! · Google search “ushoparea “ · America100,028 died from #COVID19 1,711,185 have contracted virus. My doctor team recommends a professional mask purchase channel for you, hoping that the virus will be defeated. Stay safe and social, stay away from viruses! · Google search "ushoparea"

· Kemp bans cities, counties from mandating masks | If this isn't political malpractice I don't know what is. Whatever process is appropriate for removing a governor in GA should be under way now.
TOPIC 4: Directives on Personal Hygiene and Distancing
· We are seeing widespread community transmission of COVID-19 all across Gallatin County. Please do your part to slow the spread. Wear a mask, avoid gatherings, wash hands and sanitize! · Covid-19 is real you guys. Please wear masks, social distance and wash your hands. We need to start saving each other. Let us be responsible and respectful · Wear a mask, wash your hands and stay 6 feet apart.
TOPIC 5: Mask Discourse on Skepticism (?)
· Masks don't work! So get over it! Covid-19 biggest conspiracy since the dems killed their own , Kennedy! If your so worried get jabbed and don't complain when you get some incurable disease. Had it no big deal. Natural immunity only way to go. · @BashirAhmaad Buhari didn't wear mask for months. Those around him wore wear masks and I thought the guy had ODESHI against Covid-19 · Coronavirus, 5G, anti-mask and other conspiracy theories – an essential read for rebutti... https://t.co/tFVkK5MoUo
TOPIC 6: Mask Discourse on Compliance
· Hospital boss who blamed Covid-19 outbreak on staff pictured without mask · Every time you put on a mask, never forget who screwed the world 誰是讓世界戴上口罩的兇手 #china #who #xijinping #tedrosadhanom #wuhanvirus #murderer #中國 #武漢肺炎 #習近平 #世衛 #殺人犯
TOPIC 7: Masking on Transportation

· The US Justice Department to appeal a ruling that ended a gov't order that required travellers to wear masks on public transportation across the US due to COVID-19 · Federal judge appointed by Trump strikes down the Center for Disease Control's Covid-19 mask mandate for public transportation. One judge, apparently smarter than the entire CDC says “Wearing a mask cleans nothing." I wonder how many deaths #JudgeKathrynKimballMizelle will cause. · Lyft follows Uber’s lead and removes its mask mandate: Despite the ongoing COVID-19 pandemic, face masks are no longer required for Lyft riders and drivers. Yesterday, Florida federal judge Kathryn Kimball Mizelle voided a federal mask mandate on… https://t.co/ujHn8j0C54 https://t.co/TwnEd6CxiE
TOPIC 8: COVID-19 is back (or not over yet)
· Covid-19 is back in Europe .Unsafe, protect yourself. In order to enable everyone to buy high-quality masks,we continue to try new ways to buy,including buying Amazon. Finally,my hospital team and I found that the following purchase method is the best · Parts Of Idaho Repeal Mask Mandates Even Though Hospitals Full Of COVID-19 Patients · Mask rules get tighter in Europe in winter's COVID-19 wave (from @AP)