Unmasking the conversation on masks: Natural language processing for topical sentiment analysis of COVID-19 Twitter discourse

**Abraham C. Sanders, Rachael C. White, Lauren S. Severson,
Rufeng Ma, MS, Richard McQueen, BS, Haniel C. Alcântara Paulo,
Yucheng Zhang, John S. Erickson, PhD, Kristin P. Bennett, PhD,
Rensselaer Polytechnic Institute, Troy, New York**

Abstract *In this exploratory study, we scrutinize a database of over 1 million tweets collected across the first five months of 2020 to draw conclusions about public attitudes towards the preventative measure of mask usage during the COVID-19 pandemic. In recent months, a body of literature has emerged to suggest the robustness of trends in online activity as proxies for the epidemiological and sociological impact of COVID-19. We employ natural language processing, clustering and sentiment analysis techniques to organize tweets relating to mask-wearing into high-level themes, then relay narratives for individual clusters through automatic text summarization. We find that topic clustering and visualization based on mask-related Twitter data offers revealing insights into societal perceptions of COVID-19 and techniques for its prevention. We observe that the volume and polarity of mask related tweets has greatly increased. Importantly, the analysis pipeline presented can be leveraged by the health community for the assessment of public response to health interventions in the ongoing global health crisis.*

1 Introduction

Social media provides a rich corpus of text characterizing a real-time view of daily happenings and current events within our communities. As such, it has potential utility for individuals and entities wishing to keep their finger on the pulse of both social and public health issues. Mask-wearing during the COVID-19 pandemic falls into both categories, as the consensus in the scientific community that wearing masks is key to controlling the spread of the SARS-CoV-2 virus[1] has been met with a non-negligible element of resistance to wearing masks within the population for various sociopolitical reasons. Research avenues investigating this mask usage discrepancy are increasingly relevant in light of both the evolution of the coronavirus into a border-independent global crisis and the extent to which public perceptions of the virus have changed over time.

Background and Related Works: In the pandemic-era reality that has evolved over the first half of 2020, social distancing has become the necessary norm, and it is known that social media and similar methods of online exchange are playing a bigger role than ever in keeping people connected and 3 informed.[2] One account suggests that social media platforms have seen as much as a 61% usage spike since the onset of the pandemic.[3] The social implications arising from a mass shift to virtual connectivity have been well-documented, especially regarding the dissemination of information about infection events by these means and the resulting influence on public perceptions. Importantly, Sebastian et al. have shown that the impact of locally spreading awareness is amplified if the social network of potential infection events and the network over which individuals communicate overlap, with more pronounced amplification for networks having high levels clustering.[4] This finding lends key support to the central assumptions of the analysis we present here.

In keeping with the stimulation of social media activity observed to accompany disease outbreak events, a body of literature has emerged over the past decade that looks specifically at how trends in online activity and discourse can help inform epidemiological models.[5] In conjunction, a suite of programming frameworks and models drawing on data harvested from Twitter have been developed to answer specific research questions about viral trends.[6,7] However, we observe that within this class of models, more temporally- and geospatially-comprehensive analyses of themes in the conversation about COVID-19 and its prevention are less frequent.

Major Contributions: This analysis aims to provide insight into the broadscale conversation surrounding mask-wearing that has been evolving on Twitter since March of 2020, when infection rates initially spiked in the United States, Europe, and other regions throughout the world. To this end, we develop a novel pipeline employing state-of-the-art natural language processing (NLP) techniques in order to systematically characterize Twitter discourse about and public attitudes towards the topic of mask usage during the COVID-19 pandemic. Specifically, we collect and

analyze a comprehensive sample of coronavirus-related tweets textually specific to mask-wearing. We employ clustering techniques to organize these tweets into fifteen high-level themes and fifteen specific topics within each theme, then perform sentiment analysis on the entire corpus, and also on each theme and topic, across a five-month period. We then apply an abstractive text summarization model using NLP to automatically interpret and describe the subject of the conversation occurring within each theme and topic cluster. We use data visualization and statistical analyses to examine trends in sentiments and divisiveness of the clusters.

Our pipeline is distinct from others recently developed for COVID-19-related information characterization. While other works have primarily drawn from unfiltered Twitter corpora or honed in on manually-annotated datasets specific to a particular hypothesis, we elect to study a compromise of the two approaches by refining an index of strictly tweets related to both COVID-19 and masks based on text-based keyword identification. With this semi-selective approach, we highlight the thematic trends that manifest organically in the tweets we have collected, while also ensuring that the global English-speaking conversation surrounding mask-usage during the pandemic is represented.

We find two central, co-occurring trends in the English-speaking Twitterverse by means of the presented pipeline. First, Twitter discourse surrounding mask-wearing within our curated dataset is concluded to grow consistently polarized over time, irrespective of the high-level topic with which it is associated. Moreover, we find evidence to suggest that sentimentality related to masks and mask-use as expressed on Twitter grew increasingly negative over the first five months of 2020. Cumulatively, we concur that a qualitative, semantic Twitter-based analysis pipeline is capable of revealing striking insights into deep-rooted channels in public reactions and responses to the pandemic. We hope that the methods developed here can evolve into tools to help provide rapid real-time assessment of public health measures to inform future interventions.

2 Methods

2.1 Data Collection

We used the Twitter streaming API[8] to collect 189,958,459 original tweets filtered by keywords loosely associated to COVID-19 [1] over a five month period beginning on March 17th, 2020 and ending on July 27th, 2020. Retweets during this time period were discarded, however the original tweets being referenced were collected. Twitter's API provides access to a representative random sample of approximately 1% of the data in near real time, and it has been shown that samples of tweets obtained via the API reflect the general user content generation patterns of the complete Twittersphere accurately.[9] We stored all collected tweets in Elasticsearch[10] indices for efficient search and retrieval. Using Elasticsearch, we further filtered our corpus of collected tweets by the criteria that a tweet must include at least one keyword indicating it is strongly associated to COVID-19 and at least one keyword indicating it is strongly associated to mask-wearing. This filter yielded a corpus of 1,013,039 tweets which we used for our analysis. We have made the collected corpus of tweets and the full source code for the data collection and analysis pipeline publicly available at `https://github.com/TheRensselaerIDEA/COVID-masks-nlp`. In compliance with the Twitter content redistribution policy[2], we only provide the tweet IDs corresponding to the collected tweet text used in this work.

Table 1: Filter criteria we used to identify tweets that are related to both COVID-19 and mask-wearing. A tweet must contain at least one keyphrase in both categories to be included.

Keyphrases related to COVID-19	Keyphrases related to mask-wearing
"ncov", "sars-cov-2", "covid", "covd", "covid19", "corona", "virus", "coronavirus", "koronavirus", "wuhancoronavirus", "kungflu", "epidemic", "pandemic", "quarantine", "lockdown", "flatten the curve", "flattenthecurve", "cdc"	"mask", "wearamask", "masking", "N95", "face cover", "face covering", "face covered", "mouth cover", "mouth covering", "mouth covered", "nose cover", "nose covering", "nose covered", "cover your face", "coveryourface"

[1] In addition to explicit COVID-19 keywords such as "coronavirus", we include keywords such as "school" and "cancelled" in order to include tweets about a wider array of topics impacted by the pandemic.

[2] Policy can be found at `https://developer.twitter.com/en/developer-terms/agreement-and-policy`

2.2 Analysis Pipeline

We develop an analysis pipeline to extract, label, summarize, and present the themes, topics and sentiment present in our tweet corpus using state-of-the-art natural language processing tools. While we use it here for analysis of our corpus pertaining to mask-wearing, our methods can be applied to any dataset of text documents. We have included an online supplement [3] containing additional details on implementation decisions and software packages used.

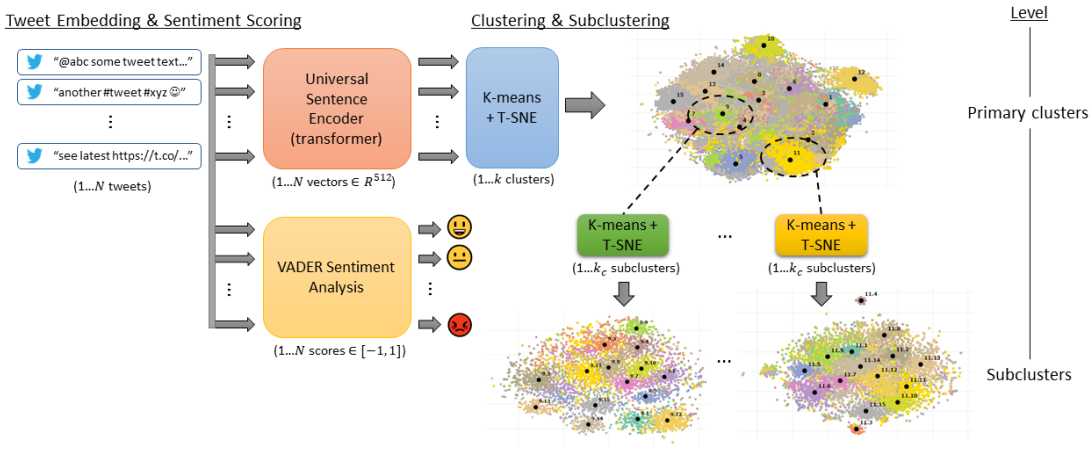


Figure 1: k-means is used to cluster the tweets their embedding space. A two-level cluster hierarchy is created by applying k-means again to each cluster.

Step 1: Retrieval & Sampling: The first step in the analysis pipeline is the retrieval of a representative random sample of tweets from the corpus. We chose $N = 100000$ as our sample size, and restricted sampled tweets to those created within the range of March 1st, 2020 to August 1st, 2020 - a sample space of 1,012,815 tweets. Once retrieved, all tweets are cleaned by removing URLs and non-punctuation characters and then normalizing all whitespace character sequences to single spaces.

Step 2: Embedding & Sentiment Scoring: After retrieving and cleaning the sample, each tweet is embedded into a 512-dimensional vector space using the transformer[11] implementation of Google's Universal Sentence Encoder.[12] The vector that represents each tweet is given by the sum of the contextual word representations at each position of the transformer encoder output. Semantically similar tweets are grouped together in the resulting embedding space, where cosine similarity provides a metric of how close two tweets are in meaning.

To assess tweet sentiment, each tweet is also scored using the VADER algorithm - a social-media-centric, lexicon-based sentiment characterization approach.[13] VADER provides a compound polarity score between -1 and 1 where -1 is the most negative and 1 is the most positive. We use the authors' recommended threshold of ± 0.05 to discretize the score where $s \leq -0.05$ is negative, $-0.05 < s < 0.05$ is neutral, and $s \geq 0.05$ is positive.

Step 3: Clustering & Subclustering: Next, we apply k-means in the embedding space to create a two-level cluster hierarchy - the corpus is grouped into k primary clusters and each primary cluster is then grouped into k_c subclusters. We interpret the primary clusters as representing high-level discussion themes and the subclusters as specific topics within each theme. We re-order the cluster numbers 1 through k and subcluster numbers 1 through k_c by average sentiment score, with 1 being the most negative. To select the optimal number of primary clusters and subclusters, we performed a computational study of the k-means objective function across a range of choices for k and k_c. As documented in our supplement, we selected $k = 15$ and $k_c = 15$ since these values provided a good balance between cluster quality and avoidance of topical redundancy.

We then use t-Distributed Stochastic Neighbor Embedding (t-SNE)[14] to project the clustered embedding space into two dimensions for presentation. In Figure 1, the cluster and subcluster scatterplots use coordinates in $\mathbb{R}^2$ given by t-SNE.

[3] Available at `https://therensselaeridea.github.io/COVID-masks-nlp/paper_supplement.pdf`

The primary cluster plot is color coded by cluster assignment and the subcluster plots are color coded by subcluster assignment. The black points represent the cluster and subcluster centers.

Step 4: Cluster & Subcluster Labeling: We find keywords that both describe and differentiate the discussion within each cluster and subcluster, and use these keywords as labels. We compute relative frequencies for words across each cluster, ignoring stopwords and non-alphanumeric characters. Using the relative frequencies, we score each word according to its contribution to the Kullback-Leibler divergence between the word distribution of the cluster and the word distribution of the entire corpus sample: $\text{score}(w) = KL(W_S||W_C) = P(W_S = w) \log \frac{P(W_S = w)}{P(W_C = w)}$. Here, W_C is the word probability distribution for the corpus sample and W_S is the word probability distribution for the sub-sample (cluster). Subclusters are labeled in the same manner, with the parent cluster taking the place of the corpus sample. Additional illustration of the labeling method is included in the supplement.

A single label representing the corpus sample is computed using the 8 words with the highest overall frequencies. For each cluster and subcluster, we then select the 3 words with the highest scores and concatenate them to create theme and topic labels respectively. To avoid reuse of keywords across labels, cluster labels cannot contain keywords that exist in the corpus sample label, and subcluster labels can not contain keywords that exist in the parent cluster label.

Step 5: Cluster & Subcluster Summarization: To augment human interpretations of each cluster and subcluster,

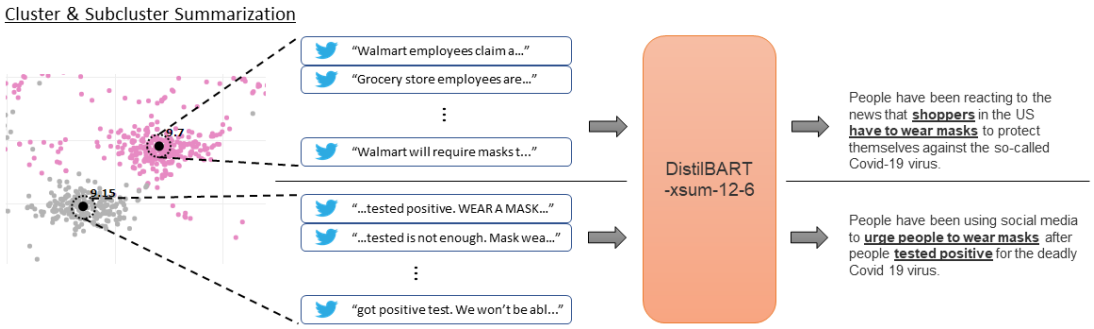


Figure 2: The tweets embedded nearest the subcluster center (shown as a black dot) are used to create the input "article" for DistilBART to summarize.

we generate summaries using DistilBART, an abstractive summarization model from the HuggingFace Transformers[15] package based on Facebook's BART[16] model. While the labels provide a quick description of the type of discussion happening within a cluster or subcluster, the one-to-three sentence summary produced by this process conveys this information in a much more meaningful way. We use a DistilBART instance fine-tuned on the extreme summarization (xsum) task[17] which aims to generate concise summaries of articles without relying on extractive summarization strategies. For each subcluster, we generate the input "article" for DistilBART to summarize by concatenating the text of 20 tweets which are embedded nearest to the subcluster center. For each cluster, we generate the input by concatenating all of the model-generated summaries of its subclusters.

2.3 Sentiment Analysis

Divisiveness in Sentiment: In order to better understand the sentiment profile of the tweet clusters, we developed a divisiveness score to assess the present level of polarization in tweet sentiment. The score is given by a real number such that polarized samples with mostly positive and negative sentiment and little neutral sentiment are given a score greater than zero, while samples with consensus, where most sentiment is unimodally concentrated on a single category, are given a score lesser than zero. Otherwise, in the case where sentiment is uniformly distributed across categories, samples have a score equal to zero.

The score itself is based on the Sarle's Bimodality Coefficient[18] (BC) with an added correction through a weighted

average with the BC of the uniform distribution and then a logit transformation. This weighting counterbalances the large variance of the BC, based on the skewness and kurtosis, for small samples[19], so that such samples with little information are considered to still have uniformly polarized sentiment.

3 Results

We examine the tweet volume and sentiment concerning masks from March to July 2020 for the entire sample. Table 2 has sentiment average, overall divisiveness, and trends in divisiveness for each cluster for tweets from March to July 2020. Cluster interpretations in Section 4 further clarify the nature of the mask discourse.

Figure 3a shows the number of negative (red), neutral (yellow) and positive tweets (blue) per week. Clearly both the volume and polarity of the discussion have dramatically increased starting in mid-June. Figure 3b shows the labels provided by the keyword analysis for each cluster, ordered from most negative sentiment to most positive sentiment. Figure 3c shows the weekly counts for tweets by sentiment for each week. Clusters 1-3 are the most negative clusters, which, as later detailed in Section 4, respectively discuss the topics of Donald Trump, individuals not wearing masks, and government mask and social distancing mandates.

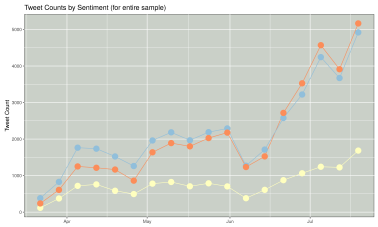
Cluster Divisiveness: To characterize the polarization of each topic cluster and the changes in polarization over time, we perform global and per-week analyses of the divisiveness scores for all clusters. For each cluster we compute divisiveness for each week, then run a linear regression of divisiveness against time; the results are shown in Table 2.

We see that for all clusters, except for Clusters 12 and 14, the confidence intervals for the slope of the fitted lines are entirely positive, indicating an increasing trend in divisiveness over time. However, no clusters display particularly steep trends, with the most significant one being Cluster 13 with a slope equivalent to only 0.0649% of the overall divisiveness score. All clusters are shown to be divisive, however, Clusters 6 and 13 possess the lowest divisiveness scores, while Clusters 2, 3 and 15 are shown to be the most divisive. Cluster 15 in particular is found to have the greatest divisiveness score, however, this result likely comes from a known fault of Sarle's BC when handling heavily skewed distributions.[18] In this case, the divisiveness score is likely incorrectly inflated due to the cluster distribution being heavily skewed towards positive sentiment, shown in Figure 3c. Clusters 2 and 3 then evidently come out to be the most polarizing out of all clusters presented, both also having comparatively large values for the fitted regression line slope with 95% certainty of increasing sentiment divisiveness.

Table 2: Average sentiment scores, divisiveness scores, and regression line slopes with 95% confidence intervals, and qualitative descriptions of time series trends. Clusters are listed in order of increasing sentiment score.

Cluster	Mean Sentiment	Sentiment 95% CI	Divisiveness Score	Divisiveness LR Slope	Divisiveness LR Slope 95% CI	Trend in Divisiveness Over Time
1	-0.1645	(-0.1767, -0.1522)	1.7472	0.0434	(0.0129 , 0.0740)	Increasing
2	-0.1147	(-0.1263, -0.1031)	2.3017	0.0935	(0.0642 , 0.1227)	Increasing
3	-0.0942	(-0.1071, -0.0811)	2.2086	0.0868	(0.0579 , 0.1157)	Increasing
4	-0.0546	(-0.0657, -0.0434)	2.1962	0.0905	(0.0627 , 0.1184)	Increasing
5	-0.0469	(-0.0589, -0.0347)	1.5292	0.0436	(0.0205, 0.0667)	Increasing
6	-0.0391	(-0.0500, -0.0281)	0.7651	0.0278	(0.0135, 00422)	Increasing
7	-0.0364	(-0.0503, -0.0224)	1.3233	0.0783	(0.0592, 0.0975)	Increasing
8	0.0272	(0.0132, 0.0411)	1.3727	0.0394	(0.0143, 0.0644)	Increasing
9	0.0365	(0.0218, 0.0510)	1.9079	0.0466	(0.0210, 0.0726)	Increasing
10	0.0387	(0.0221, 0.0551)	1.4149	0.0437	(0.0250, 0.0629)	Increasing
11	0.0394	(0.0286, 0.0502)	1.7917	0.0508	(0.0215, 0.0800)	Increasing
12	0.0607	(0.0221, 0.0551)	1.2747	0.0118	(-0.0179 , 0.0416)	*Inconclusive*
13	0.0693	(0.0584, 0.0801)	0.5094	0.0331	(0.0187, 0.04744)	Increasing
14	0.3042	(0.2934, 0.3151)	0.8411	0.0153	(-0.0048, 0.0354)	*Inconclusive*
15	0.3399	(0.3272, 0.3527)	2.4018	0.0694	(0.0242,0.1146)	Increasing

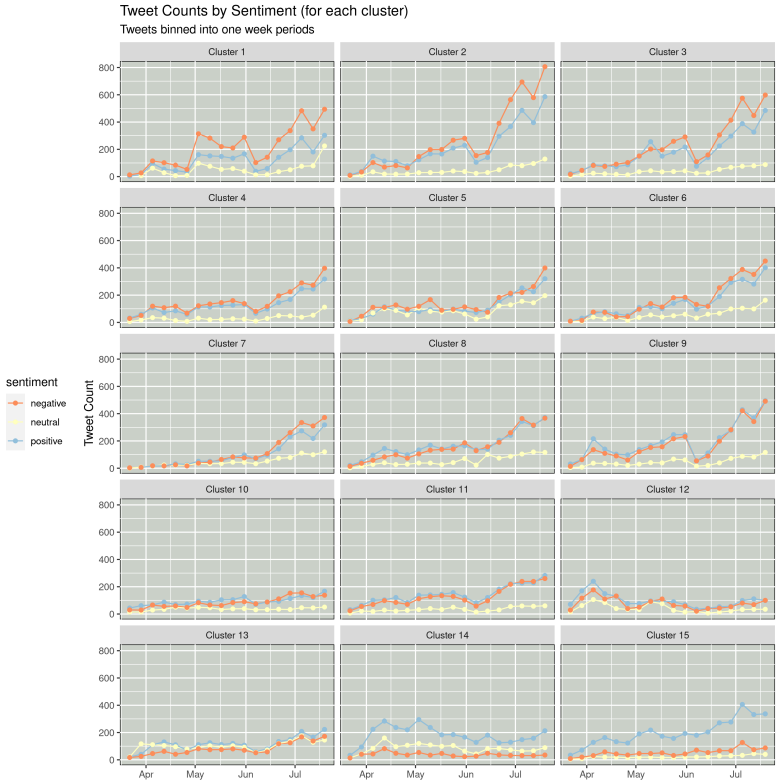
Variance in Sentiment Over Time: A one-way ANOVA was conducted for differences in mask-related sentiment across five consecutive months of early 2020 (March through July). The omnibus analysis of variance in sentiment was performed on the basis of observed normality of residuals, and with the caveat that a Breusch-Pagan test pointed to heterogeneity of variance between months. Caveat considered, the ANOVA test result indicated with significance



(a) Weekly tweet count by discretized sentiment for the entire sample

1. trump / president / realdonaldtrump
2. vaccine / flu / stop
3. lockdown / social / distancing
4. cdc / public / risk
5. pandemic / global / middle
6. coronavirus / mandate / governor
7. 19 / fucking / real
8. wearamask / maskwearing / masking
9. quarantine / gloves / store
10. corona / hero / gagging
11. droplets / spread / stop
12. n95 / surgical / microns
13. coronavirus / face / make
14. face / facemask / coronavirus
15. hands / wash / stay

(b) Top keywords by cluster



(c) Weekly tweet count by discretized sentiment for each cluster

Figure 4: Sentiment over time, for the entire tweet corpus and for each cluster

($p < 10^{-16}$) the presence of at least one distinct difference in sentiment among the five pandemic months analyzed. A subsequent Bonferroni-corrected pairwise t-test further confirmed statistically significant differences in mean sentiment score across all months studied ($p < 0.001$). In light of this finding, we followed up with a Dunnett's Correction test, a two-sided test for any difference, which compares the value of the response variable for each group to a selected control response value.[20] We chose the mean sentiment from March, the earliest period in the pandemic's development for which we had substantial tweet volume, as the baseline. The results concurred, at $\alpha = .05$, that the mean sentiment scores computed for the months of April (4), May (5), June (6), and July (7) all differed significantly from that of March (3), at $p = 0.0143$ for April and $p < 0.001$ for all other months. We further elected to re-run the Dunnett Contrasts with the alternative hypothesis that the mean sentiment for each month was *less* than the mean sentiment for March. This test assessed the null hypothesis that there was either an increase or stagnation in mean sentiment between the month of March and each respective other month. We observed that this null hypothesis was soundly rejected for each month. Cumulatively, we find significant evidence to suggest that the mean sentiment score related to masks and mask-use, as expressed within our curated tweet dataset, exhibited an overall decrease over the first five months of 2020.

4 Cluster Interpretations

In this section, we select five clusters found to be particularly striking in content. We have made available an interactive document containing the full listing of all clusters, subclusters, and automatically-generated summaries.[4].

We order the clusters by increasing overall sentiment score, report on the trends in our internally-defined sentiment and divisiveness metrics, and include the automatically-generated summary for each. We then provide manual annotations of the prominent themes that arise, as derived from a method of inspecting small samples of tweets lying near each of

[4]Our interactive cluster notebook can be found at `https://therensselaeridea.github.io/COVID-masks-nlp/analysis/twitter.html`

Figure 5: Using the interactive R Notebook to browse and summarize clusters. Cluster 1 is summarized as, "People have been reacting to news that President Donald Trump has refused to wear a face mask in public to protect himself from the deadly coronavirus pandemic."

the fifteen subcluster centers within each cluster. We see that support for mask wearing and cluster sentiment do not necessarily correspond.

Cluster 1: trump / president / realdonaldtrump **(Overall Sentiment : -0.1645 ; Divisiveness : 1.7472)**

DistilBart summary: *People have been reacting to news that President Donald Trump has refused to wear a face mask in public to protect himself from the deadly coronavirus pandemic.*

Interpretation: This cluster (shown in Figure 5) features Twitter users expressing a spectrum of attitudes towards U.S. president, Donald Trump. Opinions specifically revolve around Trump's handling of the COVID-19 pandemic in the United States. Distinctly, there exists an evident theme of frustration arising from observations that Trump has refused to wear a mask in public appearances, despite statements from public health officials encouraging the action. It should be noted that, in complement, a sizeable discussion thread of a more positive and supporting nature also exists concerning President Trump. A major theme observed here among the pro-Trump tweets is the impression that the media is biased against the president, and that this in turn fosters a public motive to exaggerate the virus. The anti-Trump tweets in this cluster are mostly focused on the president's long refusal to wear a face mask, although this finding is predictable given the nature of the data set from which the tweets are drawn.

Cluster 2: vaccine / flu / stop **(Overall Sentiment : -0.1147 ; Divisiveness : 2.3017)**

DistilBART summary: *Following the news that people in the US are being urged to wear face-covering masks to prevent the spread of a new virus that has killed more than 4,000 people in China.*

Interpretation: Cluster 2, "vaccine / flu / stop", is a grim cluster in terms of its overall sentiment, and is distinctly polemical in its semantics. It is found that the majority of tweets sampled from this cluster are pro-mask tweets complaining about individuals who don't wear masks. The dominant attitude towards masks observed among the tweets sampled for inspection is positive, despite the overall negative sentimentality computed for the cluster as a whole. In contrast with the more semantically upbeat "face / hands / stay" cluster (Cluster 15), this aggregation contains an apparent host of tweets related to death and dying. The social nature of disease is a major motif (i.e. "Your actions affect all of us.")

Cluster 3: lockdown / social / distancing **(Overall Sentiment : -0.0942 ; Divisiveness 2.3017)**

DistilBART summary: *Following the news that the US government has ordered people to wear face masks in public to prevent the spread of the deadly Covid-19 coronavirus, people across the world have been reacting to the news on social media.*

Interpretation: Cluster 3 gives an indication of the societal turbulence relating to and arising from mask mandates, social distancing enforcement, and similar lockdown-related occurrences globally. Paradoxically, the overall average sentiment of -0.0941 computed for this cluster is borderline neutral. Individual topics manifesting in this representation are observed to vary greatly, but the concerns represented in the tweets sampled appear to be, at minimum, tangentially centered around the themes of imprisonment, isolation, and quarantine. A strong racial emphasis is evident, with discourse notably focusing around protests of the #BlackLivesMatter movement, an international phenomenon co-occurring with the coronavirus pandemic mid-year. Several subclusters of Cluster 3 entertain conversations about international responses to the virus, notably around the idea that mask-wearing to prevent the spread of disease agents is a long-standing cultural norm in some regions. In keeping with the slightly negative overall computed sentiment for this cluster, many of the tweets seem to carry a sarcastic tone and a strong indication of resentment towards perceived hypocrisy surrounding mask-usage.

Cluster 12: n95 / surgical / microns **(Overall Sentiment : 0.0693 ; Divisiveness : 1.2747)**

DistilBART summary: *News that a shortage of N95 respirator masks in the US is causing a worldwide shortage has been shared on social media.*

Interpretation: Discourse within Cluster 12 focuses on information about N95 masks and related forms of personal protective equipment (P.P.E.). The evolution of the conversation around the accessibility of medical resources over the timeline of the pandemic is clearly represented. One notable stream of discussion points to the presence of a debate over how useful cloth masks are as guards against infectious agents in comparison to surgical masks. The shortage of respirators experienced by the medical community in the United States is also referenced, as is the concept that a change in tonality and meaning surrounding the suggested usage of N95 masks was observed from the U.S. CDC shortly after the pandemic infiltrated U.S. borders.

Cluster 15: hand / wash / stay **(Overall Sentiment : 0.3399 ; Divisiveness : 2.4018)**

DistilBART summary: *Social media users have been sharing their tips and advice on how to prevent the spread of the deadly coronavirus.*

Interpretation: Our most positive cluster overall, "hand / wash / stay" is composed of distinct thrusts of tweets sharing tips on prevention measures for stopping the spread of COVID-19, as well as helpful tips for self-protection from the virus. There appears to be highly positive sentiment expressed towards masks and other PPE in general, and well-meaning admonitions such as "Wash your hands and socially distance!" are frequent. In contrast to other clusters we have explored, the Cluster 15 tweets surveyed contain comparatively little in the way of aggressive, sarcastic or antagonistic semantic content. As such, this cluster may be interpreted to be an echo of the official messaging of the CDC and similar organizations.

5 Discussion

The objective of our analysis framework was to study the distribution of global mask-related social media discourse, the specific topics within this distribution, their sentimentality trends and how the latter have changed over time. In comparison to the reliable but low-context official sources of COVID-19 infection and death rate data, the accessibility and sheer quantity of organic discourse played out over Twitter make this platform an invaluable source of dynamic information on public perception of masks and mask usage during the coronavirus pandemic. The cumulative results of our pipeline point to the existence of two central, co-occurring trends in the English-speaking Twitterverse: consistently polarized Twitter discourse surrounding mask-wearing, and an accompanying overall increase in negative sentimentality. Further investigation is needed to explore whether these two factors are independent.

While mask-wearing is inherently a health-related issue, the politicization of mask-wearing on Twitter is exposed in this investigation. Regarding the cluster found to focus on US President Donald J. Trump, the mere fact that the president of the United States holds such bearing in the global Twitter conversation about mask-wearing amidst COVID-19

is intriguing. This finding speaks to the degree to which sociopolitical dynamics hold sway over the public perception of epidemiological crises like the pandemic. It suggests that such high-profile influences can play an important role (be it positive or negative) in the spreading of awareness about medical prevention techniques. The topic-sensitivity of the clustering approach we develop also opens doors for new health-related insights regarding COVID-19's trajectory and impact. Given the fact that public awareness of an infectious disease outbreak, if disseminated rapidly over a highly-connected social network, can significantly lower the infection rate of the disease,[4] our semantic clustering on mask usage could potentially inform existing governmental or institutional frameworks for promoting prevention-oriented conversation, and thus reduce the likelihood of outbreak incidents.

While our pipeline is effective there are many opportunities for improvement, an open question arising from this research is that of how well VADER-computed sentiment estimations reflect public opinion in a semantic sense. In this work we leverage lexicon-based sentiment analysis as a proxy for human attitudes and emotions, but we plan to further refine this approach to ensure more accurate and detailed sentiment representation, e.g. comprehension of sarcasm expressed towards a particular topic.

Two important limitations of our summarization method should be noted. First, the BART-based decoder is a generative language model which creates summaries autoregressively by repeatedly sampling from next-word probability distributions over an entire vocabulary. For this reason, the output summaries are prone to factual inaccuracy in a manner which extractive summarization approaches are not. Second, large or irregularly shaped subclusters may be poorly represented by the tweets immediately surrounding the subcluster center. In these situations the generated summary may not be applicable to the entire subcluster. We accept these as limitations of the system and advise readers to regard the summaries as context clues rather than as given facts.

6 Conclusion

In light of both the evolution of the virus into a global crisis and the extent to which the implications of the virus have changed in the public eye over time, semantic analyses of the character we present are increasingly relevant as sources of information to the medical research community for a host of health-related considerations. As we see, mining Twitter data allows for rapid summarization of population opinions about empirically-supported disease prevention measures. Understanding the sentiments and major trends that manifest can help inform future public health interventions and messages to increase effectiveness. Overall, from an analytical perspective, we find that thematic clustering and visualization based on mask-related Twitter data can offer distinct and unique insights on the societal perceptions of COVID-19, complementary to findings from more traditional epidemiological data sources. With the aid of abstractive visualizations like the clustering techniques presented, acute estimations of what individuals are actually saying and feeling amidst the viral destruction can be made. As future work, we hope to further evolve this pipeline into a valuable tool that enables health providers and policy makers to assess in real-time the public response to public interventions in the ongoing global health crisis.

Acknowledgements

This study was supported by the Rensselaer Institute for Data Exploration and Applications, the Data INCITE Lab, and a grant from the United Health Foundation.

References

1. D. K. Chu, E. A. Akl, S. Duda, K. Solo, S. Yaacoub, H. J. Schünemann, A. El-harakeh, A. Bognanni, T. Lotfi, M. Loeb *et al.*, "Physical distancing, face masks, and eye protection to prevent person-to-person transmission of SARS-CoV-2 and COVID-19: a systematic review and meta-analysis," *The Lancet*, 2020.

2. T. Nabity-Grover, C. M. Cheung, and J. B. Thatcher, "Inside out and outside in: How the COVID-19 pandemic affects self-disclosure on social media," *International Journal of Information Management*, p. 102188, 2020. [Online]. Available: https://bit.ly/2YzkzIG

3. R. Holmes, "Is COVID-19 Social Media's Levelling Up Moment?" Apr 2020. [Online]. Available: https://bit.ly/2QnYFDX

4. F. Sebastian, G. Erez, W. Chris, J. Vincent A. A., and G. Bryan, "The spread of awareness and its impact on epidemic outbreaks." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 16, p. 6872, 2009. [Online]. Available: https://www.pnas.org/content/106/16/6872

5. N. E. Kogan, L. Clemente, P. Liautaud, J. Kaashoek, N. B. Link, A. T. Nguyen, F. S. Lu, P. Huybers, B. Resch, C. Havas *et al.*, "An Early Warning Approach to Monitor COVID-19 Activity with Multiple Digital Traces in Near Real-Time," *arXiv preprint arXiv:2007.00756*, 2020. [Online]. Available: https://arxiv.org/abs/2007.00756

6. E. Dong, H. Du, and L. Gardner, "An interactive web-based dashboard to track COVID-19 in real time," *The Lancet. Infectious Diseases*, vol. 20, pp. 533 – 534, 2020.

7. S. Zong, A. Baheti, W. Xu, and A. Ritter, "Extracting COVID-19 Events from Twitter," *arXiv preprint arXiv:2006.02567*, 2020. [Online]. Available: https://arxiv.org/abs/2006.02567

8. "Consuming streaming data — twitter developer." [Online]. Available: https://bit.ly/32xxtbf

9. Y. Wang, J. Callan, and B. Zheng, "Should we use the sample? Analyzing datasets sampled from Twitter's stream API," *ACM Transactions on the Web (TWEB)*, vol. 9, no. 3, pp. 1–23, 2015. [Online]. Available: https://dl.acm.org/doi/10.1145/2746366

10. "Elasticsearch: The official distributed search analytics engine." [Online]. Available: https://bit.ly/3lqUa9F

11. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

12. D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar *et al.*, "Universal Sentence Encoder," *arXiv preprint arXiv:1803.11175*, 2018. [Online]. Available: https://arxiv.org/abs/1803.11175

13. C. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text," in *ICWSM*, 2014.

14. L. van der Maaten and G. Hinton, "Visualizing high-dimensional data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

15. T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "HuggingFace's Transformers: State-of-the-art Natural Language Processing," *ArXiv*, vol. abs/1910.03771, 2019.

16. M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

17. S. Narayan, S. B. Cohen, and M. Lapata, "Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization," *arXiv preprint arXiv:1808.08745*, 2018.

18. R. Pfister, K. A. Schwarz, M. Janczyk, R. Dale, and J. Freeman, "Good things peak in pairs: a note on the bimodality coefficient," *Frontiers in Psychology*, vol. 4, p. 700, 2013.

19. D. B. Wright and J. A. Herrington, "Problematic standard errors and confidence intervals for skewness and kurtosis," *Behavior Research Methods*, vol. 43, no. 1, pp. 8–17, 2011.

20. S. Lee and D. K. Lee, "What is the proper way to apply the multiple comparison test?" *Korean Journal of Anesthesiology*, vol. 71, no. 5, p. 353, 2018.