Dimensionality Reduction Technique Like TSNe, better or not?

Exploratory Analysis of Covid-19 Tweets using Topic Modeling, UMAP, and DiGraphs ?

Catherine Ordun
Department of Information Systems
University of Maryland
Baltimore County
Booz Allen Hamilton
Baltimore, Maryland
cordun1@umbc.edu

Sanjay Purushotham
Department of Information Systems
University of Maryland
Baltimore County
Baltimore, Maryland
psanjay@umbc.edu

Edward Raff
Department of Computer Science
University of Maryland
Baltimore County
Booz Allen Hamilton
Baltimore, Maryland
eraff1@umbc.edu

***Abstract*—This paper illustrates five different techniques to assess the distinctiveness of topics, key terms and features, speed of information dissemination, and network behaviors for Covid19 tweets. First, we use pattern matching and second, topic modeling through Latent Dirichlet Allocation (LDA) to generate twenty different topics that discuss case spread, healthcare workers, and personal protective equipment (PPE). One topic specific to U.S. cases would start to uptick immediately after live White House Coronavirus Task Force briefings, implying that many Twitter users are paying attention to government announcements. We contribute machine learning methods not previously reported in the Covid19 Twitter literature. This includes our third method, Uniform Manifold Approximation and Projection (UMAP), that identifies unique clustering-behavior of distinct topics to improve our understanding of important themes in the corpus and help assess the quality of generated topics. Fourth, we calculated retweeting times to understand how fast information about Covid19 propagates on Twitter. Our analysis indicates that the median retweeting time of Covid19 for a sample corpus in March 2020 was 2.87 hours, approximately 50 minutes faster than repostings from Chinese social media about H7N9 in March 2013. Lastly, we sought to understand retweet cascades, by visualizing the connections of users over time from fast to slow retweeting. As the time to retweet increases, the density of connections also increase where in our sample, we found distinct users dominating the attention of Covid19 retweeters. One of the simplest highlights of this analysis is that early-stage descriptive methods like regular expressions can successfully identify high-level themes which were consistently verified as important through every subsequent analysis.**

***Index Terms*—covid, umap, lda, twitter, coronavirus**

I. Introduction

Monitoring public conversations on Twitter about healthcare and policy issues, provides one barometer of American and global sentiment about Covid19. This is particularly valuable as the situation with Covid19 changes every day and is unpredictable during these unprecedented times. Twitter has been used as an early warning notifier, emergency communication channel, public perception monitor, and proxy public health surveillance data source in a variety of disaster and disease outbreaks from hurricanes[1], terrorist bombings [2], tsunamis [3], earthquakes [4], seasonal influenza [5], Swine flu [6], and Ebola [7]. In this paper, we conduct an exploratory analysis of topics and network dynamics of Covid19 tweets.

Since January 2020, there have been a growing number of papers that analyze Twitter activity during the Covid19 pandemic in the United States. We provide a sample of papers published since January 1, 2020 in Table I. Chen, et al. analyzed the frequency of 22 different keywords such as "Coronavirus", "Corona", "CDC", "Wuhan", "Sinophobia", and "Covid-19" analyzed across 50 million tweets from January 22, 2020 to March 16, 2020[8]. Thelwall also published an analysis of topics for English-language tweets from March 10-29, 2020.[9]. Singh et al. [10] analyzed distribution of languages and propogation of myths, Sharma et al. [11] implemented sentiment modeling to understand perception of public policy, and Cinelli et al.[12] compared Twitter against other social media platforms to model information spread.

Our contributions are applying machine learning methods not previously analyzed on Covid19 Twitter data, mainly Uniform Manifold Approximation and Projection (UMAP) to visualize LDA generated topics and directed graph visualizations of Covid19 retweet cascades. Topics generated by LDA can be difficult to interpret and while there exist coherence values [22] that are intended to score the interpretability of topics, they continue to be difficult to interpret and are subjective. As a result, we apply UMAP, a dimensionality reduction algorithm and visualization tool that "clusters" documents by topic. Vectorizing the tweets using term-frequency inverse-document-frequency (TF-IDF) and plotting a UMAP visualization with the assigned topics from LDA allowed us to identify strongly localized and distinct topics. We then visualized "retweet cascades", which describes how a social media network propagates information [23], through the use of graph models to understand how dense networks become over time and which users dominate the Covid19 conversations.

In our retweeting time analysis, we found that the median time for Covid19 messages to be retweeted is approximately 50 minutes faster than H7N9 messages during a March 2013 outbreak in China, possibly indicating the global nature, volume, and intensity of the Covid19 pandemic. Our keyword analysis and topic modeling were also rigorously explored, where we found that specific topics were triggered to uptick by Live White House Briefings, implying that Covid19 Twitter

I have this data

Reference 8,9,10,11,12 look interesting, will explore them later

LDA is an unsupervised learning technique because we don't know how many topics exist in the data.

LDA Models generate highest scoring keywords for "n" fixed number of topics(user-defined) from the data.

This "n" is arbitrary since we don't have any measure to assess the right value.

These "n" topics are not names, they are just words attached together in "n" different batches.

So What's the Problem?
The problem is we don't have any absolute measure to assess the quality of "Topics" generated by the different models. (Keep in the mind when I say topic I mean different group of keywords labelled as a "Topic".

Here Coherence Value comes into play
It acts as an absolute measure to assess the quality of "Topics" generated by different LDA Models. and helps us fix the "n" (number of topics) in LDA Model.

TABLE I: Papers published on Covid19 Twitter Analysis since January 2020

Author	Number Tweets	Time Period	Keywords	Feature Analysis	Geospatial	Topic Modeling	Sentiment	Transmission	Network Models	UMAP
Jahanbin [13], et al.	364,080	Dec. 31 2019 - Feb. 6 2020			x					
Banda, et al.[14]	30,990,645	Jan. 1 - Apr 4, 2020	x							
Medford, et al. [15]	126,049	Jan. 14 - Jan. 28, 2020	x	x		x	x			
Singh, et al.[10]	2,792,513	Jan. 16, 2020 - Mar. 15, 2020	x	x	x	x				
Lopez, et al. [16]	6,468,526	Jan. 22 - Mar. 13, 2020	x	x	x					
Cinelli, et al. [12]	1,187,482	Jan. 27 - Feb. 14, 2020		x		x		x		
Kouzy, et al. [17]	673	Feb 27, 2020	x	x						
Alshaabi, et al. [18]	Unknown	Mar. 1 - Mar 21, 2020	x	x						
Sharma, et al. [11]	30,800,000	Mar. 1, 2020 - Mar. 30, 2020	x	x	x	x	x	x	x	
Chen, et al. [8]	8,919,411	Mar. 5, 2020 - Mar. 12, 2020	x							
Schild [19]	222,212,841	Nov. 1, 2019 - Mar. 22, 2020	x	x		x			x	
Yang, et al.[20]	Unknown	Mar. 9, 2020 - Mar. 29, 2020	x						x	
Ours	**23,830,322**	**Mar. 24 - Apr. 9, 2020**	**x**	**x**		**x**			**x**	**x**
Yasin-Kabir, et al.[21]	100,000,000	Mar. 5, 2020 - Apr. 24, 2020	x	x	x		x			

TABLE II: Average Frequency of Keyword Tweets by Minute

Corpus	**bed**	**hospital**	**mask**	**icu**	**help**	**nurse**	**doctors**	**vent**	**test_pos**	**serious_cond**	**exposure**	**cough**	**fever**
3/24/2020	3.341	30.068	**38.295**	3.159	2.591	4.886	8.455	25.977	0.636	0.023	0.250	0.409	0.023
3/25/2020	3.117	33.021	**38.734**	2.819	3.181	3.745	8.064	24.691	1.298	0.043	0.277	0.372	0.106
3/28/2020	1.819	30.648	**34.352**	1.714	2.362	4.800	8.486	38.790	0.962	0.019	0.181	0.181	0.029
3/30/2020	2.783	40.957	**53.796**	2.311	3.287	6.996	13.009	24.887	1.111	0.025	0.215	0.296	0.043
3/31/2020	2.109	30.673	**72.877**	1.447	3.677	5.633	10.410	17.995	1.020	0.014	0.152	0.494	0.147
4/2/2020	2.065	29.410	**84.467**	1.474	3.164	6.147	10.450	23.424	0.814	0.018	0.192	0.357	0.045
4/5/2020	2.218	31.812	**62.786**	2.493	3.039	5.798	10.735	17.909	1.026	0.014	0.175	0.309	0.052
Mean	2.493	32.370	**55.044**	2.203	3.043	5.429	9.944	24.811	0.981	0.022	0.206	0.345	0.064

users are highly attuned to government broadcasts. We think this is important because it highlights how other researchers have identified that government agencies play a critical role in sharing information via Twitter to improve situational awareness and disaster response [24]. Our LDA models confirm that topics detected by Thelwall et al. [9] and Sharma et al. [11], who analyzed Twitter during a similar period of time, were also identified in our dataset which emphasized healthcare providers, personal protective equipment such as masks and ventilators, and cases of death.

A. Research Questions

This paper studies five research questions:

1) What high-level trends can be inferred from Covid19 tweets?
2) Are there any events that lead to spikes in Covid19 Twitter activity?
3) Which topics are distinct from each other?
4) How does the speed of retweeting in Covid19 compare to other emergencies, and especially similar infectious disease outbreaks?
5) How do Covid19 networks behave as information spreads?

The paper begins with Data Collection, followed by the five stages of our analysis: Keyword Trend Analysis, Topic Modeling, UMAP, Time-to-Retweet Analysis, and Network Analysis. Our methods and results are explained in each section. The paper concludes with limitations of our analysis. The Appendix provides additional graphs as supporting evidence.

II. Data Collection

Similar to researchers in Table I, we collected Twitter data by leveraging the free Streaming API. From March 24, 2020 to April 9, 2020, we collected 23,830,322 (173 GB) tweets. Note, in this paper, we refer to the Twitter data interchangeably as both "dataset" and "corpora" and refer to the posts as "tweets". Our dataset is a collection of tweets from different time periods shown in Table V. Using the Twitter API through tweepy, a Python Twitter mining and authentication API, we first queried the Twitter track on twelve query terms to capture a healthcare-focused dataset: 'ICU beds', 'ppe', 'masks', 'long hours', 'deaths', 'hospitalized', 'cases', 'ventilators', 'respiratory', 'hospitals', '#covid', and '#coronavirus'. For the keyword analysis, topic modeling, and UMAP tasks, we analyzed non-retweets that brought the corpus down to 5,506,223 tweets. In the Time-to-Retweet and Network Analysis, we included retweets but selected a sample out of the larger 23.8 million corpus of 736,561 tweets. Our preprocessing steps are described in the Data Analysis section that follows.

III. Keyword Trend Analysis

Prior to applying keyword analysis, we first had to preprocess the corpus on the "text" field. First, we removed retweets using regular expressions, in order to focus the text on original tweets and authorship, as opposed to retweets that can inflate the number of messages in the corpus. We use no-retweeted corpora for both the keyword trend analysis and the topic modeling and UMAP analyses. Further we formatted datetime to UTC format, removed digits, short words less than 3 characters, extended the NLTK stopwords list to also exclude "coronavirus", "covid19", "19", "covid", removed "https:" hyperlinks, removed "@" signs for usernames, removed non-Latin characters such as Arabic or Chinese characters, and implemented lower-casing, stemming, and tokenization. Finally, using regular expressions, we extracted tweets that

Author applied a known technique to trending data.
This is respectful because no one else applied UMAP.

Twitter data especially from the Chen source is heavy, as of May end I have 80 million tweets roughly 300GB.

After Cleaning the data it's only 35GB.

These keywords were tracked.

Tweets extracted via API have RT in front of them if they are retweet.

1. Removed Retweets
2. Used non retweets corpora for keywords trend analysis, topic modelling and UMAP analysis.
3. Converted to UTC Format.
4. Removed digits, len(words) < 3.
5. Added to the NLTK Stopwords.
6. Removed urls, @, non-Latin Chars
7. Implemented Lower casing, stemming and tokenization

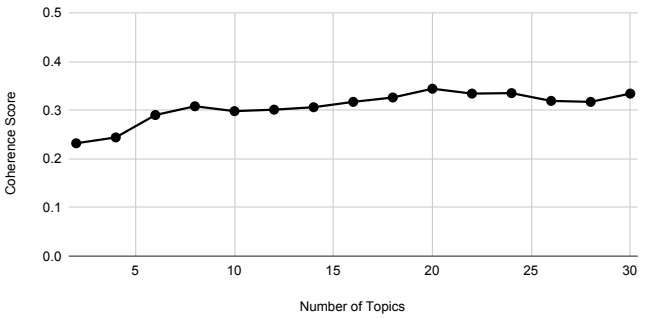


Fig. 1: Coherence Scores by Number of Topics

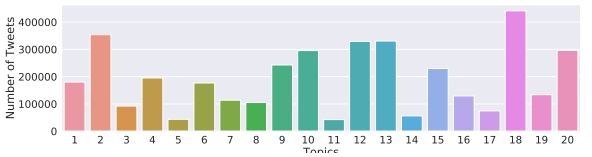


Fig. 2: Distribution of 20 Topics in the Corpora

contained the following thirteen single terms: 'bed', 'hospital', 'mask', 'icu', 'help', 'nurse', 'doctors', 'vent', 'test_pos', 'serious_cond', 'exposure', 'cough', and 'fever', in order to gain insights about currently trending public concerns. We present values of the raw counts of the tweets in the Appendix under Table VI and the frequencies of tweets per minute here in Table II.

The greatest rate of tweets occurred for the tweets consisting of the term "mask" (mean 55.044) in Table II, followed by "hospital" (mean 32.370) and "vent" (mean 24.811). Tweets of less than 1.0 mean tweets per minute, came from groups about testing positive, being in serious condition, exposure, cough, and fever. This may indicate that people are discussing the issues around Covid19 more frequently than symptoms and health conditions in this dataset. We will later find out that several themes consistent with these keyword findings are mentioned in topic modeling to include personal protective equipment (PPE) like ventilators and masks, and healthcare workers like nurses and doctors.

This roughly makes sense because people were panicking about the shortage of masks, hospital beds, vent (ventilators). Every country battling COVID faced shortage of necessary equipments.

IV. Topic Modeling

LDA are mixture models, meaning that documents can belong to multiple topics and membership is fractional [25]. Further, each topic is a mixture of words, where words can be shared among topics. This allows for a "fuzzy" form of unsupervised clustering where a single document can belong to multiple topics, each with an associated probability. LDA is a bag of words model where each vector is a count of terms. LDA requires the number of topics to be specified. Similar to methods described by Syed et al. [26], we ran 15 different LDA experiments varying the number of topics from 2 to 30, and selected the model with the highest coherence value score. We selected the LDA model that generated 20 topics, with a medium coherence value score of 0.344. Roder et al. [22] developed the coherence value as a metric that calculates the agreement of a set of pairs and word subsets and their associated word probabilities into a single score. In general, topics are interpreted as being coherent if all or most of terms are related.

Explained this on the Page-1 Left Side.

Our final model generated 20 topics using the default parameters of the Gensim LDA MultiCore model [1] with an overall coherence score of 0.428 after modifying the chunksize to 50,000. The topics are provided in Figure 2 and include the terms generated and each topic's coherence score measuring interpretability. Similar to the high-level trends inferred from extracting keywords, themes about PPE and healthcare workers dominate the nature of topics. The terms generated also indicate emerging words in public conversation including "hydroxychloroquine" and "asymptomatic".

The logic behind mentioned all the hyper-parameters inside LDA Model is a way to reproduce results. For example chunk size can vary from 1 to 1 billion so there is no way we can fix that on our own. We need author to provide it to us and they have done a great job.

Our results also show four topics that are in non-English languages. In our preprocessing, we removed non-Latin characters in order to filter out a high volume of Arabic and Chinese characters. In Twitter there exists a Tweet object metadata field of "lang" for language to filter tweets by a specific language like English ("eng"). However, we decided not to filter against the "lang" element because upon observation, approximately 2.5% of the dataset consisted of an "undefined" language tag, meaning that no language was indicated. Although it appears to be a small fraction, removing even the "undefined" tweets would have removed several thousand tweets. Some of these tweets that are tagged as "undefined" are in English but contain hashtags, emojis, and Arabic characters. As a result, we did not filter out for English language, leading our topics to be a mix of English, Spanish, Italian, French, and Portuguese. Although this introduced challenges in interpretation, we feel it demonstrates the global nature of worldwide conversations about Covid19 occurring on Twitter. This is consistent with what Singh et al. Singh et al. [10] reported as a variety of languages in Covid19 tweets upon analyzing over 2 million tweets. As a result, we labeled the four topics by the language of the terms in the respective topics: "Spanish" (Topic 1), "Portuguese" (Topic 14), "Italian" (Topic 16) and "French" (Topic 19). We used Google Translate to infer the language of the terms.

When examining the distribution of the 20 topics across the corpora in Figure 2, Topics 18 ("potus"), 12 ("case.death.new"), 13 ("mask.ppe.ventil"), and 2 ("like.look.work") were the top five in the entire corpora. For each plot, we labeled each topic with the first three terms of each topic for interpretability. In our trend analysis, we summed the number of tweets per minute, and then applied a moving weighted average of 60 minutes for topics March 24 - March 28, and 60 minutes for topics March 30 to April 8th. We provided two different plots in order to visualize smaller time frames such as March 24 of 44 minutes compared to

[1] https://radimrehurek.com/gensim/models/ldamulticore.html

TABLE III: 20 Topics Generated from LDA Model

Topic	C_V	Terms	Language
1	0.922	de, la, el, en, que, lo, por, del, para, se, es, con, un, al, est, una, su, ms, caso, todo	Spanish
2	0.241	like, look, work, dont, amp, peopl, time, read, support, respiratori, great, death, us, case, hospit, listen, im, presid, agre, way	English
3	0.222	hospit, realli, patient, johnson, bori, oh, shit, amp, peopl, make, death, e, blood, like, call, treat, human, trial, guy	English
4	0.171	china, thank, lockdown, viru, latest, corona, pandem, covid2019, us, lie, hai, ye, stayhom, trump, daili, way, social, quarantin, help, 5g	English
5	0.363	case, spread, help, slow, risk, symptom, daili, mask, identifi, sooner, asymptomat, us, test, market, selfreport, de, 2, 9, question, commun	English
6	0.413	day, case, week, news, ago, state, health, two, month, death, last, 15, us, delhi, hospit, one, 2, new, said, lockdown	English
7	0.287	test, case, hospit, posit, corona, dr, viru, kit, patient, ppe, doctor, data, govern, work, de, say, vaccin, death, drug, amp	English
8	0.173	die, world, peopl, case, us, death, der, tell, und, flu, corona, da, im, never, cant, fr, thousand, africa, help, ist	English
9	0.413	mask, face, wear, make, one, public, protect, cdc, peopl, dont, n95, recommend, us, viru, love, cloth, new, 0, trump, work	English
10	0.440	mask, home, stay, peopl, pleas, ppe, hospit, help, work, wear, amp, like, worker, care, nurs, safe, sure, dont, doctor, hand	English
11	0.296	hospit, nurs, le, case, de, ppe, work, new, doctor, go, pay, help, let, one, live, us, local, time, staff, lockdown	English
12	0.572	case, death, new, report, total, confirm, day, posit, number, york, us, state, 1, today, 2, 3, updat, test, peopl, rise	English
13	0.483	mask, ppe, ventil, hospit, medic, trump, suppli, donat, us, need, worker, state, china, n95, million, use, help, order, equip, amp	English
14	0.713	de, que, e, em, da, per, el, com, la, para, um, se, os, le, na, un, mai, brasil, dia, del	Portuguese
15	0.490	case, death, number, total, countri, updat, time, india, confirm, recov, china, corona, hour, last, us, news, peopl, new, activ, hospit	English
16	0.582	di, il, e, la, na, per, che, non, sa, al, si, un, da, del, ng, ang, le, ha, con, het	Italian
17	0.247	great, god, news, sad, shame, ppe, bless, hydroxychloroquin, hospit, de, death, ventil, stori, die, amp, hear, man, case, hong, holi	English
18	0.329	trump, peopl, death, american, live, stop, amp, us, let, hospit, time, viru, caus, like, one, dont, true, go, kill, media	English
19	0.904	de, le la, en, et, du, pour, un, pa, que, il, ce, au, qui, confin, dan, une, est, cest, sur	French
20	0.293	hospit, im, peopl, still, govern, dont, thing, amp, death, fuck, one, work, job, state, money, model, us, start, happen, ive	English

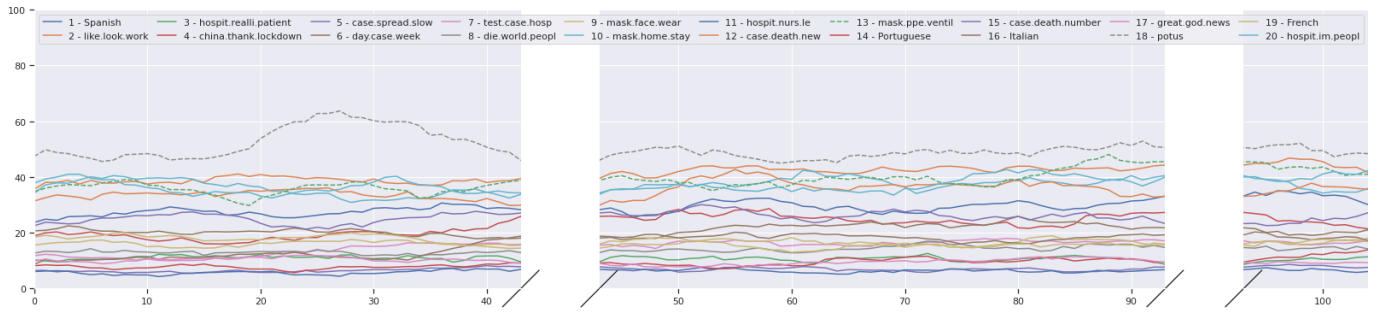


Fig. 3: Trend of Topics over Time from March 24 to March 28, 2020

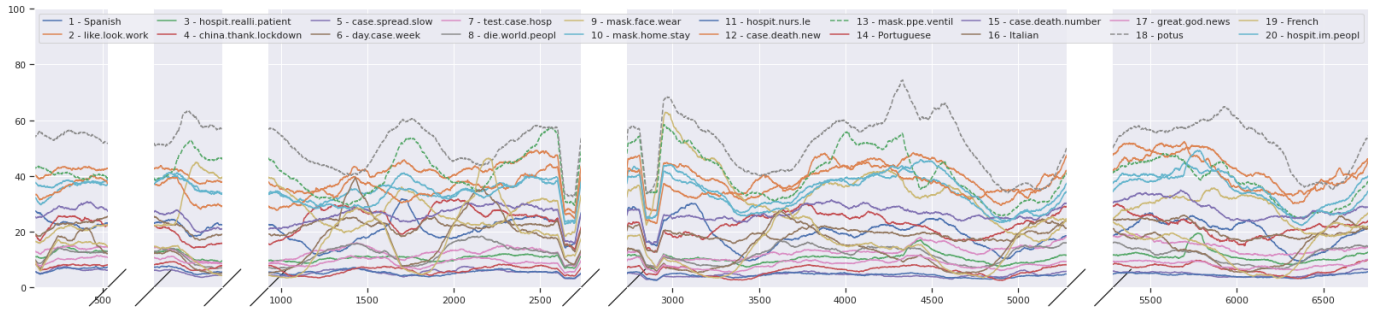


Fig. 4: Trend of Topics over Time from March 30 to April 8, 2020

longer time frames of 1477 for April 8. The results plotted in figures Figure 3 and Figure 4 show similar trends on a time-series basis per minute across the entire corpora of 5,506,223 tweets. These plots are in a style of "broken axes" [2] to indicate that the corpora are not continuous periods of time, but discrete time frames, which we selected to plot on one axis for convenience and legibility. We direct the reader to Table V for reference on the start and end datetimes, which are in UTC format, so please adjust accordingly for time zone.

The x-axis denotes the number of minutes, where the entire corpora is 8463 total minutes of tweets. Figure 3 shows that for the corpora of March 24, 25, and 28, the topics (denoted in hash-marked lines) focused on Topic 18 "potus" and Topic 13 "mask.ppe.ventil" trended greatest. For the later time periods of March 30, March 31, April 4, 5 and 8 in Figure 4, Topic 18 "potus" and Topic 13 "mask.ppe.ventil" (also in hash-marked lines) continued to trended high. It is also interesting that Topic 18 was never replaced as the top trending topic, across a span of 17 days (April 8, 2020 also includes early hours of April 9 2020 EST), potentially as this may have been a proxy for active government listening. The time series would temporally decrease in frequency during overnight hours, between the

[2]https://github.com/bendichter/brokenaxes

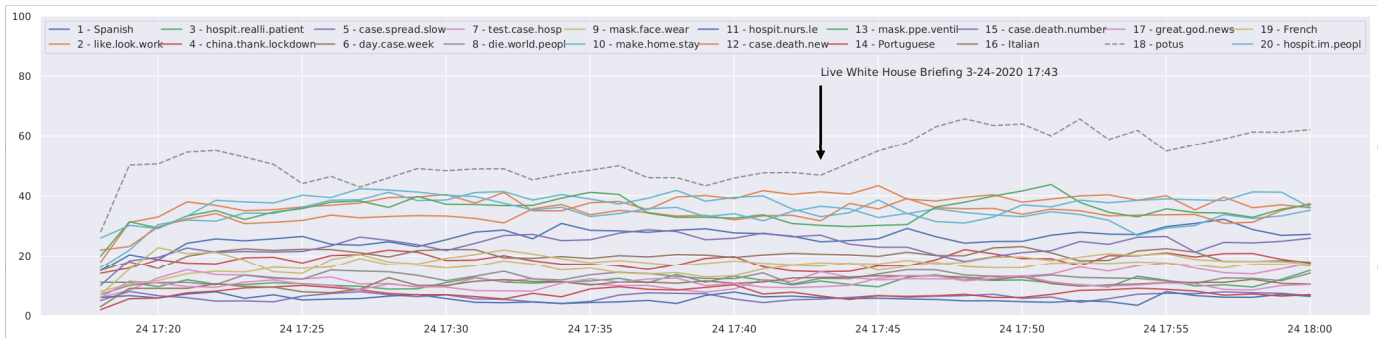


Fig. 5: March 24 5:17 PM to 6:00 PM EST Topics Time Series

hours of midnight and 6:00 AM EST. But when examining the trend of the Topic 18 "potus" topic, we found that several live press briefings with the Coronavirus Task Force from @WhiteHouse would stimulate a spike in the Topic 18 topic 60 tweets per minute.

- March 24, 2020, LIVE: Press Briefing with Coronavirus Task Force at 5:43 PM EST
- April 3, 2020, LIVE: Press Briefing with Coronavirus Task Force at 5:24 PM EST followed by a retweet from @WhiteHouse "Coronavirus—and we salute the great medical professionals on the front lines." at 5:59 PM EST
- April 4, 2020, LIVE: Press Briefing with Coronavirus Task Force at 4:13 PM EST
- April 5, 2020, LIVE: Press Briefing with Coronavirus Task Force at 6:53 PM EST
- April 6, 2020, LIVE: Press Briefing with Coronavirus Task Force at 5:41 PM EST
- April 8, 2020: LIVE: Press Briefing with Coronavirus Task Force at 5:46 PM EST

We applied change point detection in the time series of tweets per minute for Topic 18 in the datasets March 24, 2020, April 3 - 4, 2020, April 5 - 6, 2020, and April 8, 2020, to identify whether the live press briefings coincided with inflections in time. Using the ruptures Python package [27] containing a variety of change point detection methods, we used binary segmentation [28], a standard method for change point detection. Given a sequence of data $y_{1:n} = (y_1, ..., y_n)$ the model will have m changepoints with their positions $\tau_{1:m} = (\tau_1, ..., \tau_m)$. Each changepoint position is an integer between 1 and $n-1$. The m changepoints split the time series data into $m+1$ segments, with the ith segment containing $y_(\tau_{i-1}+1) : \tau_i$. Changepoints are identified by minimizing a cost function, C for a given segment, where $\beta f(m)$ is a penalty to prevent overfitting.

$$\sum_{i=1}^{m+1} [C(y_(\tau_{i-1}+1) : \tau_i)] + \beta f(m)$$

where twice the negative log-likelihood is a commonly used cost function.

Binary segmentation detects multiple changepoints across the time series by repeatedly testing on different subsets of the sequence. It checks to see if a τ exists that satisfies:

$$C(y_{1:\tau} + C(y_{(\tau+1):n}) + \beta < C(y_{1:n})$$

If not, then no changepoint is detected and the method stops. But if a changepoint is detected, the data are split into two segments consisting of the time series before (Figure 7 blue) and after (Figure 7 pink) the changepoint. We can clearly see in Figure 7 that the timing of the White House briefing indicates a changepoint in time, giving us the intuition that this briefing influenced an uptick in the the number of tweets. We provide additional examples in the Appendix.

Our topic findings are consistent with the published analyses on Covid19 and Twitter, such as [10] who found major themes of healthcare and illness and international dialogue, as we noticed in our four non-English topics. They are also similar to by Thelwall et al. [9] who manually reviewed tweets from a corpus of 12 million tweets occurring earlier and overlapping our dataset (March 10 - 29). Similar topics from their findings to ours includes "lockdown life", "politics", "safety messages", "people with COVID-19", "support for key workers", "work", and "COVID-19 facts/news".

Further, our dataset of Covid19 tweets from March 24 to April 8, 2020 occurred during a month of exponential case growth. By the end of our data collection period, the number of cases had increased by 7 times to 427,460 cases on April 8, 2020 [29]. The key topics we identified using our multiple methods were representative of the public conversations being had in news outlets during March and April, including:

- CDC allowing private companies to make tests (March 3, 2020)
- President Trump declaring Covid19 a national emergency (March 13, 2020)
- CDC advising against social gatherings of more than 50 people (March 15, 2020)[30]
- CDC issuing (March 17, 2020) Strategies for Optimizing the Supply of Facemasks
- President Trump mentioning hydroxychloroquine as a potential Covid19 treatment.[31]

Change Point Detection detects abrupt shifts in time series trends. (i.e. shifts in a time series' instantaneous velocity), that can be easily identified via the human eye, but are harder to pinpoint using traditional statistical approaches.

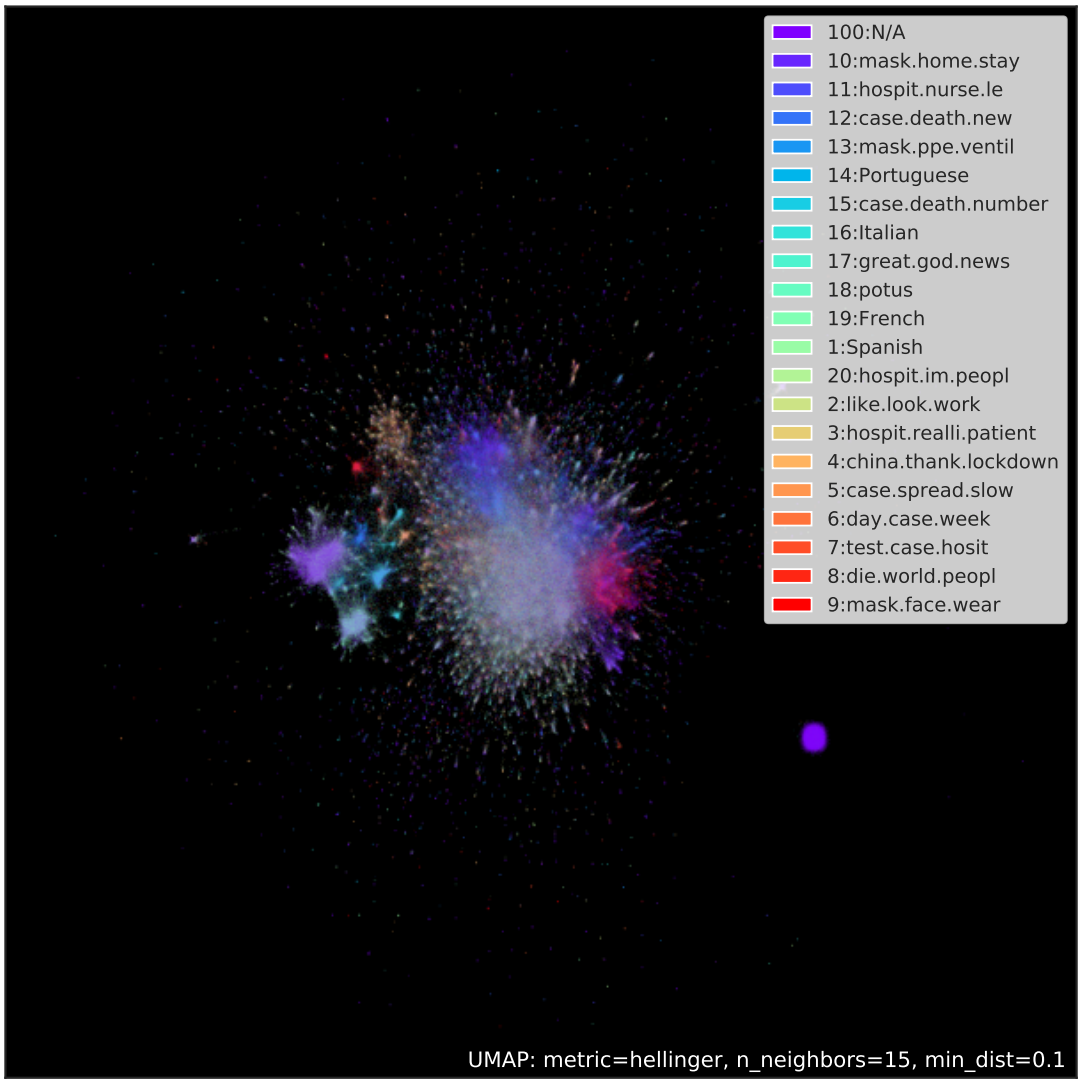


Fig. 6: Visualization of One Million Tweets with Topic Labels

- The HHS Assistant Secretary for Health and U.S. Surgeon General issuing a letter to the healthcare community to optimize ventilator use (March 31, 2020)
- The White House issues a Memorandum on Order Under the Defense Production Act Regarding the Purchase of Ventilators (April 2, 2020)[32]
- CDC issuing guidance on wearing facial coverings (April 3, 2020) [33]

V. Uniform Manifold Approximation and Projection

Term-frequency inverse-document-frequency (TF-IDF)[34] is a weight that signifies how valuable a term is within a document in a corpus, and can be calculated at the n-gram level. TF-IDF has been widely applied for feature extraction on tweets used for text classification [35] [36], analyzing sentiment [37], and for text matching in political rumor detection [23] With TF-IDF, unique words carry greater information and value than common, high frequency words across the corpus. TF-IDF can be calculated as follows:

$$w_{i,j} = tf_{i,j} \times \log \frac{N}{df_i}$$

Where i is the term, j is the document, and N is the total number of documents in the corpus. TF-IDF calculates the term frequency $tf_{i,j}$ multiplied by the log of the inverse document frequency $\frac{N}{df_i}$. The term frequency $tf_{i,j}$ is calculated as the frequency of i in j divided by all terms i in given j. The inverse document frequency is $\frac{N}{df_i}$ is the log of the total number of documents j in the corpus divided by the number of documents j containing term, i.

Using the Scikit-Learn implementation of TfidfVectorizer and setting max_features to 10000, we transformed our corpus of 5,506,223 tweets into a $\mathbb{R}^{n \times k}$ sparse dimensional matrix of shape (5506223, 10000). Note, prior to fitting the vectorizer, our corpus of tweets was pre-processed during the keyword

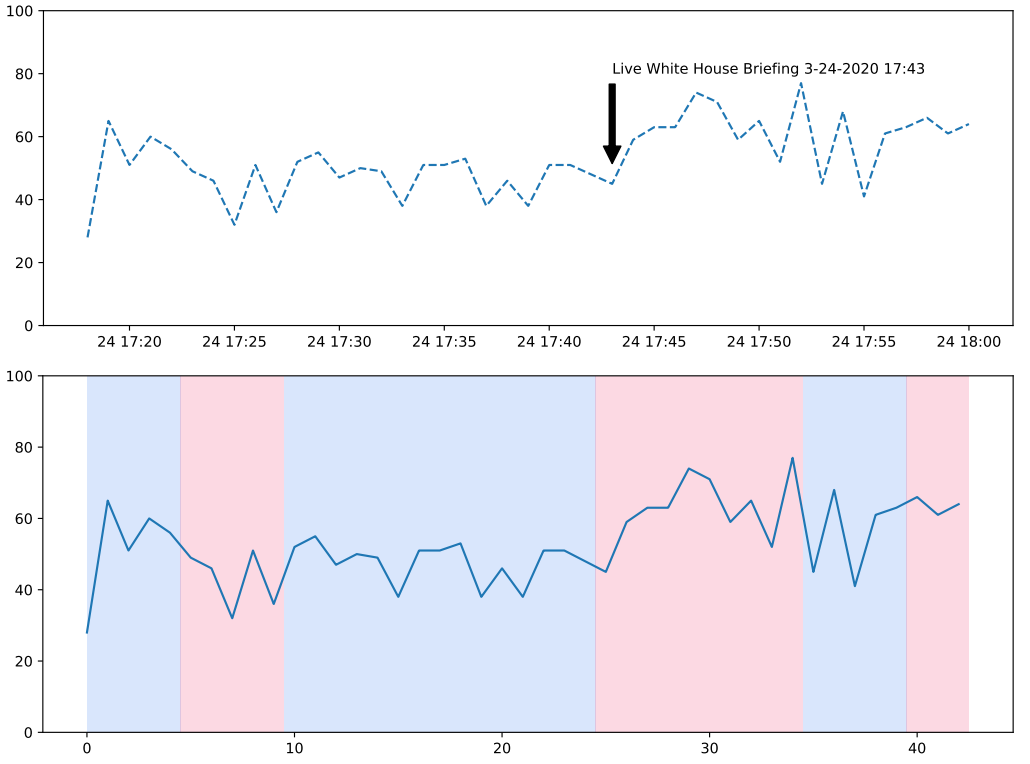


Fig. 7: Change Point Detection using Binary Segmentation for March 24, 2020

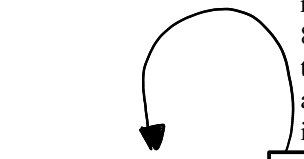
analysis stage. We chose to visualize how the 20 topics grouped together using Uniform Manifold Approximation and Projection (UMAP) [38]. UMAP is a dimension reduction algorithm that finds a low dimensional representation of data with similar topological properties as the high dimensional space. It measures the local distance of points across a neighborhood graph of the high dimensional data, capturing what is called a fuzzy topological representation of the data. Optimization is then used to find the closest fuzzy topological structure by first approximating nearest neighbors using the Nearest-Neighbor-Descent algorithm and then minimizing local distances of the approximate topology using stochastic gradient descent [39]. When compared to t-Distributed Stochastic Neighbor Embedding (t-SNE), UMAP has been observed to be faster [40] with clearer separation of groups.

Due to compute limitations in fitting the entire high dimensional vector of nearly 5.5M records, we randomly sampled one million records. We created an embedding of the vectors along two components to fit the UMAP model with the Hellinger metric which compares distances between probability distributions, as follows:

$$h(P,Q) = \frac{1}{\sqrt{2}} \cdot \left\| \left(\sqrt{P} - \sqrt{Q}\right) \right\|_2$$

We visualized the word vectors with their respective labels, which were the assigned topics generated from the LDA model. We used the default parameters of n_neighbors = 15 and min_dist = 0.1. Figure 6 presents the visualization of the TF-IDF word vectors for each of the 1 million tweets with their labeled topics. UMAP is supposed to preserve local and global structure of data, unlike t-SNE that separates groups but does not preserve global structure. As a result, UMAP visualizations intend to allow the reader to interpret distances between groups as meaningful. In Figure 6 each topic is color-coded by its respective topic.

The UMAP plots appear to provide further evidence of the quality and number of topics generated. Our observations is that many of these topic "clusters" appear to have a single dominant color indicating distinct grouping. There is strong local clustering for topics that were also prominent in the keyword analysis and topic modeling time series plots. A very distinct and separated mass of purple tweets represents the "100: N/A" topic which is an undefined topic. This means that the LDA model outputted equal scores across all 20 topics for any single tweet. As a result, we could not assign a topic to these tweets because they all had uniform scores. But this visualization informs us that the contents of these tweets were uniquely distinct from the others. Examples of tweets in this "100: N/A" cateogry include "See, #Democrats are always guilty of whatever", "Why are people still getting in cruise ships?!?", "Thank you Mike you are always helping others and sponsoring Anchors media shows.", "We cannot let this woman's brave and courageous actions go to waste! #ChinaLiedPeopleDied #Chinaneedstopay", "I wish people in this country would just stay the hell home instead of GOING TO THE BEACH". Other observations reveal that the mask-



related topic 10 in purple, and potentially a combination of 8 and 9 in red are distinct from the mass of noisy topics in the center of the plot. We can also see distinct separation of aqua-colored topic 18 "potus" and potentially topics 5 and 6 in yellow.

We refer the reader to other examples where UMAP has been leveraged for Twitter analysis, to include Darwish et al. [41] for identifying clusters of Twitter users with controversial topic similarity, Vargas [42] for event detection, political polarization by Darwish et al. [41] and estimating political leaning of users by [43].

VI. Time-to-Retweet Analysis

Retweeting is a special activity reserved for Twitter where any user can "retweet" messages which allows them to disseminate their messages rapidly to their followers. Further, a highly retweeted tweet might signal that an issue has attracted attention in the highly competitive Twitter environment, and may give insight about issues that resonate with the public [44]. Whereas in the first three analyses we used no retweets, in the time-series and network modeling that follows, we exclusively use retweets. We began by measuring time-to-retweet. Wang et al. [1] calls this "response time" and used it to measure response efficiency and speed of information dissemination during Hurricane Sandy. Wang analyzed 986,579 tweets and found that 67% of re-tweets occur within 1 h [1]. We researched how fast other users retweet in emergency situations, such as what Spiro [45] reported for natural disasters, and how Earle [46] reported as 19 seconds for retweeting about an earthquake.

We extracted metadata from our corpora for the Tweet, User, and Entities objects. For reference, we direct the reader to the Twitter Developer guide that provides a detailed overview of each object[47]. Due to compute limitations, we selected a sample that consisted of 736,561 tweets that included retweets from the corpora of March 24 - 28, 2020. However, since we were only focused on retweets, out of the corpus of 736,561 tweets, we reduced it to 567,909 (77%) that were only retweets. The metadata we used for both our Time-to-Retweet and Directed Graph analyses in the next section, included:

1) Created_at (string) - UTC time when this Tweet was created.
2) Text (string) - The actual UTF-8 text of the status update. See twitter-text for details on what characters are currently considered valid.
3) From the User object, the id_str (string) - The string representation of the unique identifier for this User.
4) From the retweeted_status object (Tweet) - the created_at UTC time when the Retweet was created.
5) From the retweeted_status object (Tweet) - the id_str which is the unique identifier for the retweeting user.

We used the corpus of retweets and analyzed the time between the tweet created_at and the retweeted created_at.

$$time\_to\_rt = rt\_object - tw\_object$$

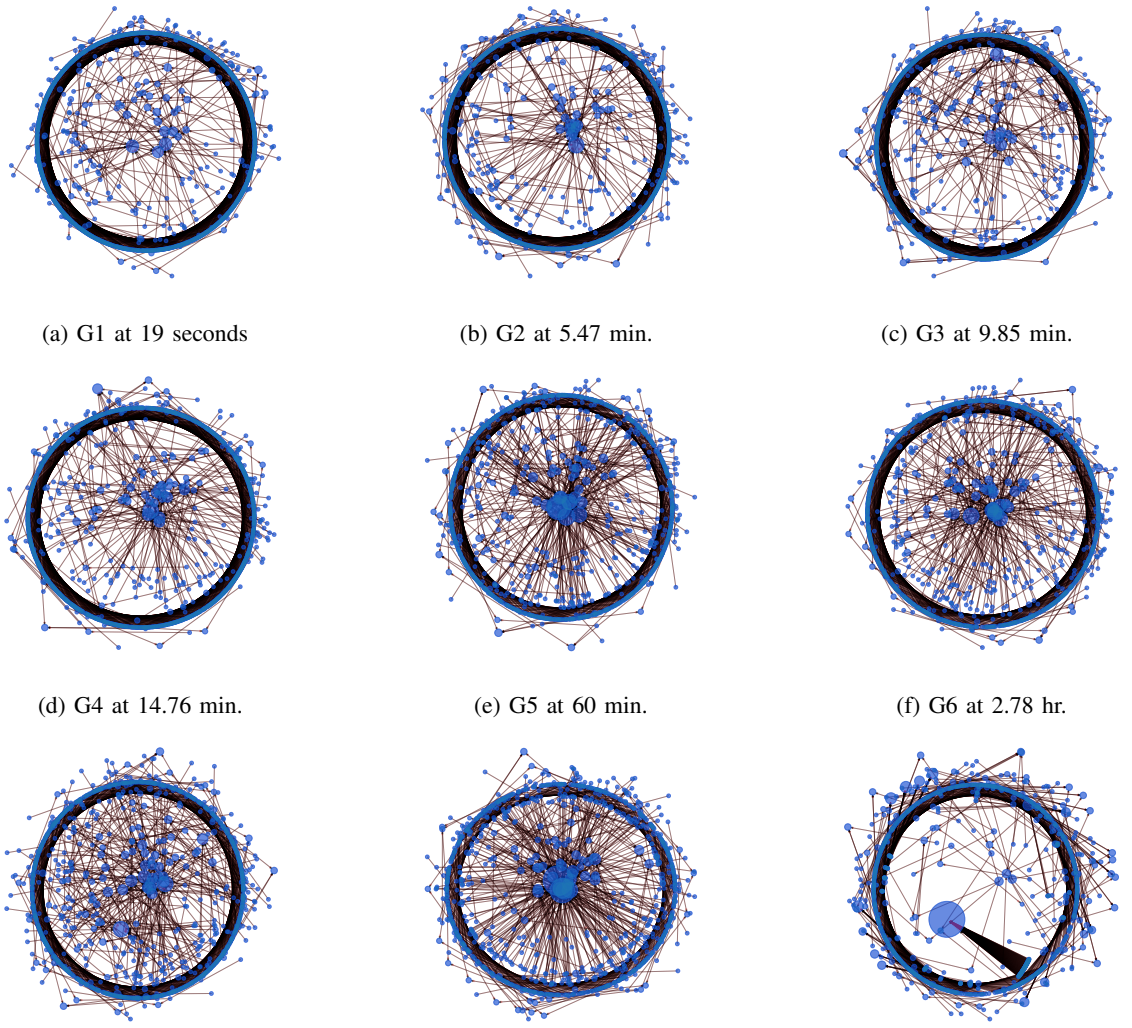
Here, the rt_object is the datetime in UTC format for when the message that was retweeted was originally posted. The tw_object is the datetime in UTC format when the current tweet was posted. As a result, the datetime for the rt_object is older than the datetime for the current tweet. This measures the time it took for the author of the current tweet to retweet the originating message. This is similar to Kuang et al. [48] who defined response time of the retweet to be the time difference between the time of the first retweet and that of the origin tweet. Further, Spiro et al. [45] calls these "waiting times". The median time-to-retweet for our corpus was 2.87 hours meaning that half of the tweets occurred within this time (less than what Wang reported as 1.0 hour), and the mean was 12.3 hours. Figure 9 shows the histogram of the number of tweets by their time to retweet in seconds and Figure 10 shows it in hours.

Further, we found that compared to the 2013 Avian Influenza outbreak (H7N9) in China described by Zhang et al. [49] Covid19 retweeters sent more messages earlier than H7N9. Zhang analyzed the log distribution of 61,024 H7N9-related posts during April 2013 and plotted reposting time of messages on Sina Weibo, a Chinese Twitter-like platform and one of the largest microblogging sites in China Figure 12. Zhang found that H7N9 reposting occurred with a median time of 222 minutes (i.e. 3.7 hours) and a mean of 8520 minutes (i.e. 142 hours). Compared to Zhang's study, we found our median retweet time to be 2.87 hours, about 50 minutes faster than the reposting time during H7N9 of 3.7 hours. When comparing Figure 11 and Figure 12, it appears that Covid19 retweeting does now completely slow down until 2.78 hours later (10^4 seconds). For H7N9 it appears to slow down much earlier by 10 seconds.

Unfortunately few studies appear to document retweeting times during infectious disease outbreaks which made it hard to compare how Covid19 retweeting behavior against similar situations. Further, the H7N9 outbreak in China occurred seven years ago and may not be a comparable set of data for numerous reasons. Chinese social media may not represent similar behaviors with American Twitter and this analysis does not take into account multiple factors that imply retweeting behavior to include the context, the user's position, and the time the tweet was posted [44].

A. TF-IDF Message and User Description Features of Rapid Retweeters

We also analyzed what rapid retweeters, or those retweeting messages even faster than the median, in less than 10,000 seconds were saying. In Figure 21 we plotted the top 50 TF-IDF features by their scores for the text of the retweets. It is intuitive to see that URLs are being retweeted quickly by the presence of "https" in the body of the retweeted text. This is also consistent with studies by Suh et al. [50] who indicated that tweets with URLs were a significant factor impacting retweetability. We found terms that were frequently mentioned during the early-stage keyword analysis and topic modeling mentioned again: "cases", "ventilators", "hospitals", "deaths",



(a) G1 at 19 seconds (b) G2 at 5.47 min. (c) G3 at 9.85 min.

(d) G4 at 14.76 min. (e) G5 at 60 min. (f) G6 at 2.78 hr.

(g) G7 at 3.7 hr. (h) G8 at 24 hr. (i) G9 at 168 hr.

Fig. 8: Directed Graphs of Covid19 Retweeting Activity at Nine Different Points in Time (G1 - G9) between March 24 - March 28th using the Kamada Kawai Layout

"masks", "test", "american", "cuomo", "york", "president", "china", and "news". When analyzing the descriptions of the users who were retweeted in Figure 21, we ran the TF-IDF vectorizer on bigrams in order to elicit more interpretable terms. User accounts whose tweets were rapidly retweeted, appeared to describe themselves as political, news-related, or some form of social media account, all of which are difficult to verify as real or fake.

TABLE IV: Statistics about Each Network Community

Graphs	Ranking Speed	Time Point	Density	Nodes	1st	2nd	3rd
G1	1	19 sec	0.000428	1278	11	11	9
G2	2	328 sec (5.47 min)	0.000449	1248	17	8	8
G3	3	591 sec (9.85 min)	0.000450	1247	13	12	9
G4	4	885.6 sec (14.76 min)	0.000460	1234	17	10	10
G5	5	3600 sec (60 min)	0.000567	1110	41	27	20
G6	6	10000 sec (2.78 hrs)	0.000538	1139	18	15	15
G7	7	13,320 sec (3.7 hrs)	0.000540	1138	17	17	11
G8	8	86,400 sec (24 hrs)	0.000685	1005	63	43	26
G9	9	604,800 sec (1 week)	0.000598	1067	92	9	9

VII. Network Modeling

We analyzed the network dynamics of nine different time periods within the March 24 - 28, 2020 Covid19 dataset, and visualized them based on their speed of retweeting. These types of graphs have been referred to as "retweet cascades" which describes how a social media network propagates information [23]. Similar methods have been applied for visualizing rumor propogation by Jin et al. [23] We wanted to analyze how Covid19 retweeting behaves at different time points. We used published disaster retweeting times to serve as benchmarks for selecting time periods. As a result, the graphs in Figure 8 are plotted by retweeting time of known benchmarks - the median time to retweet after an earthquake which implies rapid notification, the median time to retweet after a funnel

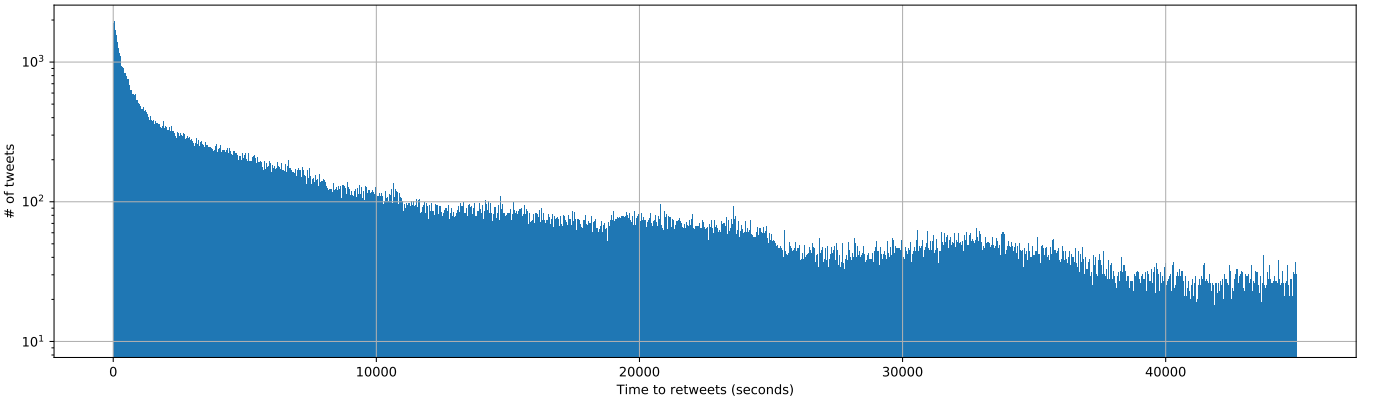


Fig. 9: Seconds to Retweet, March 24 - 28th Corpora

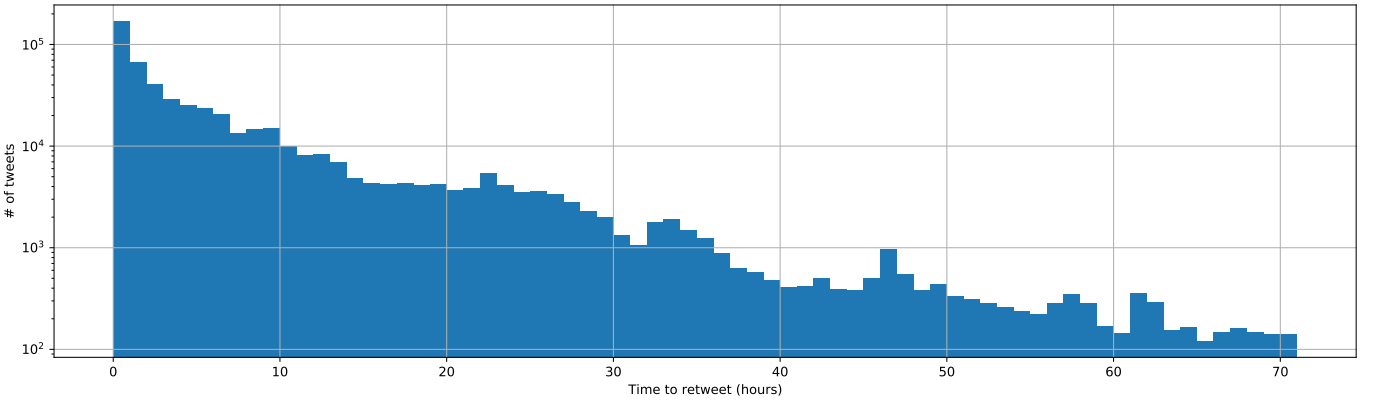


Fig. 10: Hours to Retweet, March 24 - 28th Corpora

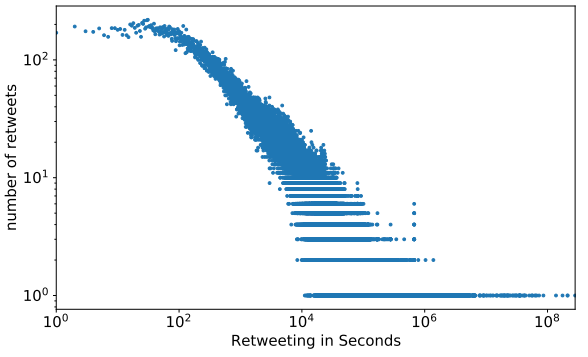


Fig. 11: Log Distribution of Covid19 Retweets from March 24 - 28, 2020

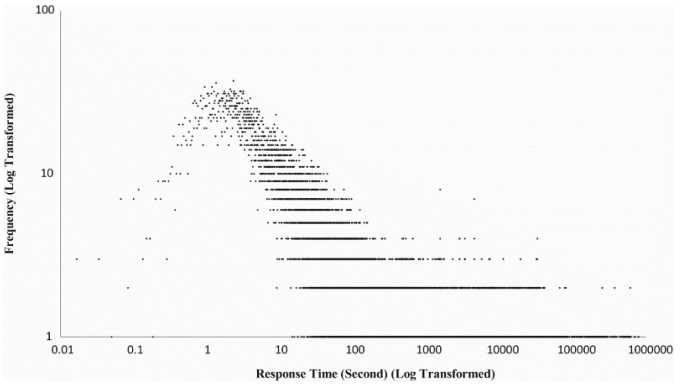


Fig. 12: Log Distribution of H7N9-related messages on Sina Weibo, March 2013

cloud has been seen, all the way to a one-day or 24 hour time period. We did this to visualize a retweet cascade of fast to slow information propogation. We used median retweeting times published Spiro et al. [45] for the time it took users to retweet messages based on hazardous keywords like "Funnel Cloud", "Aftershock", and "Mudslide". We also used the H7N9 reposting time which Zhang et al. [49] published of 3.7 hours.

We generated a Directed Graph for each of the nine time periods, where the network consisted of a source which was the author of the tweet (User object, the id_str) and a target which was the original retweeter shown in Table IV. The goal was to analyze how connections change as the retweeting speed increases. The nine networks are visualized in Figure 8. Graphs were plotted using networkx and drawn using the Kamada Kawai Layout[51], a force-directed algorithm. We

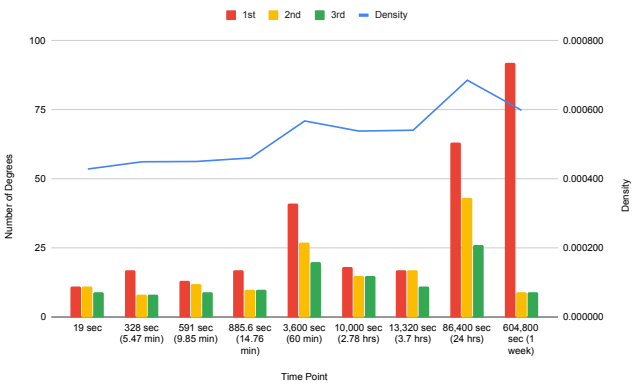


Fig. 13: Increasing Density and Degree for Top 3 Users

modeled 700 users for each graph. We found that more nodes became too difficult to interpret. The size of the node indicates the number of degrees, or users that it is connected to. It can mean that the node has been retweeted by others several times. Or, it can also mean that the node itself has been retweeted by others several times.

The density of each network increases over time shown in Figure 8 and Figure 13. Very rapid retweeters, in the time it takes to retweet after an earthquake, start off with a sparse network with a few nodes in the center being the focus of retweets in Figure 8a. By the time we reach Figure 8d, the retweeted users are much more clustered in the center and there are more connections and activity. The top retweeted user in our median time network Figure 8g, was a news network and tweeted *"The team took less than a week to take the ventilator from the drawing board to working prototype, so that it can"*. By 24 hours out in Figure 8h, we see a concentrated set of users being retweeted and by Figure 8i, one account appears to dominate the space being retweeted 92 times. This account was retweeting the following message several times *"She was doing #chemotherapy couldn't leave the house because of the threat of #coronavirus so her line sisters..."*. In addition, the number of nodes generally decreased from 1278 in "earthquake" time to 1067 in one week, and the density also generally increased, shown in Table IV.

These retweet cascade graphs provide only an exploratory analysis. Network structures like these have been used to predict virality of messages, for example memes over time as the message is diffused across networks [52]. But, analyzing them further could enable 1) an improved understanding about how Covid19 information diffusion is different than other outbreaks, or global events, 2) How information is transmitted differently from region to region across the world, and 3) What users and messages are being concentrated on over time. This would support strategies to improve government communications, emergency messaging, dispelling medical rumors, and tailoring public health announcements.

VIII. Limitations

There are several limitations with this study. First, our dataset is discontinuous and trends seen in Figure 3 and Figure 4 where there is an interruption in time should be taken with caution. Although there appears to be a trend between one discrete time and another, without the missing data, it is impossible to confirm this as a trend. As a result, it would be valuable to apply these techniques on a larger and continuous corpus without any time breaks. We aim to repeat the methods in this study on a longer continuous stream of Twitter data in the near future.

Next, the corpus we analyzed was already pre-filtered with thirteen "track" terms from the Twitter Streaming API that focused the dataset towards healthcare related concerns. This may be the reason why the high level keywords extracted in the first round of analysis were consistently mentioned throughout the different stages of modeling. However, after review of similar papers indicated in Table I, we found that despite having filtered the corpus on healthcare-related terms, topics still appear to be consistent with analyses where corpora were filtered on limited terms like "#coronavirus".

Third, the users and conversations in Twitter are not a direct representation of the U.S. or global population. The Pew Research Foundation found that only 22% of American adults use Twitter [53] and that this group is different from the majority of U.S. adults, because they are on average younger, more likely to identify as Democrats, more highly educated and possess higher incomes [54]. The users were also not verified and should be considered as a possible mixture of human and bot accounts.

Fourth, we reduced our corpus to remove retweets for the keyword and topic modeling anlayses since retweets can obscure the message by introducing virality and altering the perception of the information [55]. As a result, this reduced the size of our corpus by nearly 77% from 23,820,322 tweets to 5,506,223 tweets. However, there appears to be variability in terms of consistent corpora sizes in the Twitter analysis literature both in Table I and other health-related studies. For example, Karami [56] used 4.5 million tweets, Zhao [57] used 1,225,851 tweets, Hong[58]used 1,992,758 tweets, Surian [59] used 285,417 tweets, Alverez[60] used 101,522 tweets, and Lim [61] used only 60,370 tweets.

Fifth, our compute limitations prohibited us from analyzing a larger corpus for the UMAP, time-series, and network modeling. For the LDA models we leveraged the gensim MulticoreLDA model that allowed us to leverage multiprocessing across 20 workers. But for UMAP and the network modeling, we were constrained to use a CPU. However, as stated above, visualizing more than 700 nodes for our graph models was unintepretable. Applying our methods across the entire 23.8 million corpora for UMAP and the network models may yield more meaningful results. Sixth, we were only able to iterate over 15 different LDA models based on changing the number of topics, whereas Syed et al. [26] iterated on 480 models to select coherent models. We believe that applying a manual

gridsearch of the LDA parameters such as iterations, alpha, gamma threshold, chunksize, and number of passes would lead to a more diverse representation of LDA models and possibly more coherent topics.

Seven, it was challenging to identify papers that analyzed Twitter networks according to their speed of retweets for public health emergencies and disease outbreaks. Zhang et al. [49] points out that there are not enough studies of temporal measurement of public response to health emergencies. We were lucky to find papers by Zhang et al. [49] and Spiro et al. [45] who published on disaster waiting times. Chew et al. [62] and Szomszor et al. [6] have published about Twitter analysis in H1N1 and the Swine Flu, respectively. Chew analyzed the volume of H1N1 tweets and categorized different types of messages such as humor and concern. Szomszor correlated tweets with UK national surveillance data and Tang et al. [63] generated a semantic network of tweets on measles during the 2015 measles outbreak to understand keywords mentioned about news updates, public health, vaccines and politics. However, it was difficult to compare our findings against other disease outbreaks due to the lack of similar modeling and published retweet cascade times and network models.

IX. Conclusion

We answered five research questions about Covid19 tweets during March 24, 2020 - April 8, 2020. First, we found high-level trends that could be inferred from keyword analysis. Second, we found that live White House Coronavirus Briefings led to spikes in Topic 18 ("potus"). Third, using UMAP, we found strong local "clustering" of topics representing PPE, healthcare workers, and government concerns. UMAP allowed for an improved understanding of distinct topics generated by LDA. Fourth, we used retweets to calculate the speed of retweeting. We found that the median retweeting time was 2.87 hours. Fifth, using directed graphs we plotted the networks of Covid19 retweeting communities from rapid to longer retweeting times. The density of each network increased over time as the number of nodes generally decreased.

Lastly, we recommend trying all techniques indicated in Table I to gain an overall understanding of Covid19 Twitter data. While applying multiple methods for an exploratory strategy, there is no technical guarantee that the same combination of five methods analyzed in this paper will yield insights on a different time period of data. As a result, researchers should attempt multiple techniques and draw on existing literature.

Acknowledgment

The authors would like to acknowledge John Larson from Booz Allen Hamilton for his support and review of this article.

References

[1] B. Wang and J. Zhuang, "Crisis information distribution on twitter: a content analysis of tweets during hurricane sandy," *Natural hazards*, vol. 89, no. 1, pp. 161–181, 2017.

[2] C. Buntain, J. Golbeck, B. Liu, and G. LaFree, "Evaluating public response to the boston marathon bombing and other acts of terrorism through twitter," in *Tenth International AAAI Conference on Web and Social Media*, 2016.

[3] A. Chatfield and U. Brajawidagda, "Twitter tsunami early warning network: a social network analysis of twitter information flows," 2012.

[4] P. S. Earle, D. C. Bowden, and M. Guy, "Twitter earthquake detection: earthquake monitoring in a social world," *Annals of Geophysics*, vol. 54, no. 6, 2012.

[5] R. Nagar, Q. Yuan, C. C. Freifeld, M. Santillana, A. Nojima, R. Chunara, and J. S. Brownstein, "A case study of the new york city 2012-2013 influenza season with daily geocoded twitter data from temporal and spatiotemporal perspectives," *Journal of medical Internet research*, vol. 16, no. 10, p. e236, 2014.

[6] M. Szomszor, P. Kostkova, and E. De Quincey, "# swineflu: Twitter predicts swine flu outbreak in 2009," in *International conference on electronic healthcare*. Springer, 2010, pp. 18–26.

[7] M. Odlum and S. Yoon, "What can we learn about the ebola outbreak from tweets?" *American journal of infection control*, vol. 43, no. 6, pp. 563–571, 2015.

[8] E. Chen, K. Lerman, and E. Ferrara, "Covid-19: The first public coronavirus twitter dataset," *arXiv preprint arXiv:2003.07372*, 2020.

[9] M. Thelwall and S. Thelwall, "Retweeting for covid-19: Consensus building, information sharing, dissent, and lockdown life," *arXiv preprint arXiv:2004.02793*, 2020.

[10] L. Singh, S. Bansal, L. Bode, C. Budak, G. Chi, K. Kawintiranon, C. Padden, R. Vanarsdall, E. Vraga, and Y. Wang, "A first look at covid-19 information and misinformation sharing on twitter," *arXiv preprint arXiv:2003.13907*, 2020.

[11] K. Sharma, S. Seo, C. Meng, S. Rambhatla, A. Dua, and Y. Liu, "Coronavirus on social media: Analyzing misinformation in twitter conversations," *arXiv preprint arXiv:2003.12309*, 2020.

[12] M. Cinelli, W. Quattrociocchi, A. Galeazzi, C. M. Valensise, E. Brugnoli, A. L. Schmidt, P. Zola, F. Zollo, and A. Scala, "The covid-19 social media infodemic," *arXiv preprint arXiv:2003.05004*, 2020.

[13] K. Jahanbin and V. Rahmanian, "Using twitter and web news mining to predict covid-19 outbreak," *Asian Pacific Journal of Tropical Medicine*, p. 13, 2020.

[14] J. M. Banda, R. Tekumalla, G. Wang, J. Yu, T. Liu, Y. Ding, and G. Chowell, "A large-scale covid-19 twitter chatter dataset for open scientific research–an international collaboration," *arXiv preprint arXiv:2004.03688*,

2020.

[15] R. J. Medford, S. N. Saleh, A. Sumarsono, T. M. Perl, and C. U. Lehmann, “An" infodemic": Leveraging high-volume twitter data to understand public sentiment for the covid-19 outbreak,” *medRxiv*, 2020.

[16] C. E. Lopez, M. Vasu, and C. Gallemore, “Understanding the perception of covid-19 policies by mining a multilanguage twitter dataset,” *arXiv preprint arXiv:2003.10359*, 2020.

[17] R. Kouzy, J. Abi Jaoude, A. Kraitem, M. B. El Alam, B. Karam, E. Adib, J. Zarka, C. Traboulsi, E. W. Akl, and K. Baddour, “Coronavirus goes viral: Quantifying the covid-19 misinformation epidemic on twitter,” *Cureus*, vol. 12, no. 3, 2020.

[18] T. Alshaabi, J. Minot, M. Arnold, J. L. Adams, D. R. Dewhurst, A. J. Reagan, R. Muhamad, C. M. Danforth, and P. S. Dodds, “How the world’s collective attention is being paid to a pandemic: Covid-19 related 1-gram time series for 24 languages on twitter,” *arXiv preprint arXiv:2003.12614*, 2020.

[19] L. Schild, C. Ling, J. Blackburn, G. Stringhini, Y. Zhang, and S. Zannettou, “" go eat a bat, chang!": An early look on the emergence of sinophobic behavior on web communities in the face of covid-19,” *arXiv preprint arXiv:2004.04046*, 2020.

[20] K.-C. Yang, C. Torres-Lugo, and F. Menczer, “Prevalence of low-credibility information on twitter during the covid-19 outbreak,” *arXiv preprint arXiv:2004.14484*, 2020.

[21] M. Yasin Kabir and S. Madria, “Coronavis: A real-time covid-19 tweets analyzer,” *arXiv*, pp. arXiv–2004, 2020.

[22] M. Röder, A. Both, and A. Hinneburg, “Exploring the space of topic coherence measures,” in *Proceedings of the eighth ACM international conference on Web search and data mining*, 2015, pp. 399–408.

[23] Z. Jin, J. Cao, H. Guo, Y. Zhang, Y. Wang, and J. Luo, “Detection and analysis of 2016 us presidential election related rumors on twitter,” in *International conference on social computing, behavioral-cultural modeling and prediction and behavior representation in modeling and simulation*. Springer, 2017, pp. 14–24.

[24] N. Genes, M. Chary, and K. Chason, “Analysis of twitter users’ sharing of official new york storm response messages,” *Medicine 2.0*, vol. 3, no. 1, 2014.

[25] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[26] S. Syed and M. Spruit, “Full-text or abstract? examining topic coherence scores using latent dirichlet allocation,” in *2017 IEEE International conference on data science and advanced analytics (DSAA)*. IEEE, 2017, pp. 165–174.

[27] C. Truong, L. Oudre, and N. Vayatis, “Selective review of offline change point detection methods,” *Signal Processing*, vol. 167, p. 107299, 2020.

[28] R. Killick, P. Fearnhead, and I. A. Eckley, “Optimal detection of changepoints with a linear computational cost,” *Journal of the American Statistical Association*, vol. 107, no. 500, pp. 1590–1598, 2012.

[29] “Cases in u.s.” Apr 2020. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html

[30] “Get your mass gatherings or large community events ready,” Mar 2020. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/community/large-events/mass-gatherings-ready-for-covid-19.html

[31] K. Liptak, “Trump says fda will fast-track treatments for novel coronavirus, but there are still months of research ahead,” *Trump says FDA will fast-track treatments for novel coronavirus, but there are still months of research ahead*, Mar 2020. [Online]. Available: https://www.cnn.com/2020/03/19/politics/trump-fda-anti-viral-treatments-coronavirus/index.html

[32] *The White House*. Presidential Memoranda, Apr 2020. [Online]. Available: https://www.whitehouse.gov/presidential-actions/memorandum-order-defense-production-act-regarding-purchase-ventilators/

[33] “Recommendation regarding the use of cloth face coverings,” Apr 2020. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/cloth-face-cover.html

[34] J. Ramos *et al.*, “Using tf-idf to determine word relevance in document queries,” in *Proceedings of the first instructional conference on machine learning*, vol. 242. Piscataway, NJ, 2003, pp. 133–142.

[35] K. Lee, D. Palsetia, R. Narayanan, M. M. A. Patwary, A. Agrawal, and A. Choudhary, “Twitter trending topic classification,” in *2011 IEEE 11th International Conference on Data Mining Workshops*. IEEE, 2011, pp. 251–258.

[36] L. Hong, O. Dan, and B. D. Davison, “Predicting popular messages in twitter,” in *Proceedings of the 20th international conference companion on World wide web*, 2011, pp. 57–58.

[37] P. Barnaghi, P. Ghaffari, and J. G. Breslin, “Opinion mining and sentiment polarity on twitter and correlation between events and sentiment,” in *2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService)*. IEEE, 2016, pp. 52–57.

[38] L. McInnes, J. Healy, and J. Melville, “Umap: Uniform manifold approximation and projection for dimension reduction,” *arXiv preprint arXiv:1802.03426*, 2018.

[39] L. McInnes, “How umap works¶.” [Online]. Available: https://umap-learn.readthedocs.io/en/latest/how_umap_works.html

[40] A. Coenen and A. Pearce, “Understanding umap.” [Online]. Available: https://pair-code.github.io/understanding-umap/

[41] K. Darwish, P. Stefanov, M. J. Aupetit, and P. Nakov, “Unsupervised user stance detection on twitter,” *arXiv preprint arXiv:1904.02000*, 2019.

[42] V. Vargas-Calderón, J. E. Camargo, H. Vinck-Posada *et al.*, "Event detection in colombian security twitter news using fine-grained latent topic analysis," *arXiv preprint arXiv:1911.08370*, 2019.

[43] P. Stefanov, K. Darwish, and P. Nakov, "Predicting the topical stance of media and popular twitter users," *arXiv preprint arXiv:1907.01260*, 2019.

[44] N. Naveed, T. Gottron, J. Kunegis, and A. C. Alhadi, "Bad news travel fast: A content-based analysis of interestingness on twitter," in *Proceedings of the 3rd international web science conference*, 2011, pp. 1–7.

[45] E. Spiro, C. Irvine, C. DuBois, and C. Butts, "Waiting for a retweet: modeling waiting times in information propagation," in *2012 NIPS workshop of social networks and social media conference. http://snap. stanford. edu/social2012/papers/spiro-dubois-butts. pdf. Accessed*, vol. 12, 2012.

[46] P. Earle, M. Guy, R. Buckmaster, C. Ostrum, S. Horvath, and A. Vaughan, "Omg earthquake! can twitter improve earthquake response?" *Seismological Research Letters*, vol. 81, no. 2, pp. 246–251, 2010.

[47] "Introduction to tweet json - twitter developers." [Online]. Available: https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/intro-to-tweet-json

[48] L. Kuang, X. Tang, and K. Guo, "Predicting the times of retweeting in microblogs," *Mathematical Problems in Engineering*, vol. 2014, 2014.

[49] L. Zhang, L. Xu, and W. Zhang, "Social media as amplification station: factors that influence the speed of online public response to health emergencies," *Asian Journal of Communication*, vol. 27, no. 3, pp. 322–338, 2017.

[50] B. Suh, L. Hong, P. Pirolli, and E. H. Chi, "Want to be retweeted? large scale analytics on factors impacting retweet in twitter network," in *2010 IEEE Second International Conference on Social Computing*. IEEE, 2010, pp. 177–184.

[51] T. Kamada, S. Kawai *et al.*, "An algorithm for drawing general undirected graphs," *Information processing letters*, vol. 31, no. 1, pp. 7–15, 1989.

[52] L. Weng, F. Menczer, and Y.-Y. Ahn, "Virality prediction and community structure in social networks," *Scientific reports*, vol. 3, p. 2522, 2013.

[53] A. Perrin and M. Anderson, "Share of u.s. adults using social media, including facebook, is mostly unchanged since 2018," Apr 2019. [Online]. Available: https://www.pewresearch.org/fact-tank/2019/04/10/share-of-u-s-adults-using-social-media-including-facebook-is-mostly-unchanged-since-2018/

[54] S. Wojcik and A. Hughes, "How twitter users compare to the general public," Jan 2020. [Online]. Available: https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/

[55] A. C. Madrigal, "Retweets are trash," Mar 2018. [Online]. Available: https://www.theatlantic.com/magazine/archive/2018/04/the-case-against-retweets/554078/

[56] A. Karami, A. A. Dahl, G. Turner-McGrievy, H. Kharrazi, and G. Shaw Jr, "Characterizing diabetes, diet, exercise, and obesity comments on twitter," *International Journal of Information Management*, vol. 38, no. 1, pp. 1–6, 2018.

[57] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing twitter and traditional media using topic models," in *European conference on information retrieval*. Springer, 2011, pp. 338–349.

[58] L. Hong and B. D. Davison, "Empirical study of topic modeling in twitter," in *Proceedings of the first workshop on social media analytics*, 2010, pp. 80–88.

[59] D. Surian, D. Q. Nguyen, G. Kennedy, M. Johnson, E. Coiera, and A. G. Dunn, "Characterizing twitter discussions about hpv vaccines using topic modeling and community detection," *Journal of medical Internet research*, vol. 18, no. 8, p. e232, 2016.

[60] D. Alvarez-Melis and M. Saveski, "Topic modeling in twitter: Aggregating tweets by conversations," in *Tenth international AAAI conference on web and social media*, 2016.

[61] K. W. Lim, C. Chen, and W. Buntine, "Twitter-network topic model: A full bayesian treatment for social network and text modeling," *arXiv preprint arXiv:1609.06791*, 2016.

[62] C. Chew and G. Eysenbach, "Pandemics in the age of twitter: content analysis of tweets during the 2009 h1n1 outbreak," *PloS one*, vol. 5, no. 11, 2010.

[63] L. Tang, B. Bie, and D. Zhi, "Tweeting about measles during stages of an outbreak: A semantic network approach to the framing of an emerging infectious disease," *American journal of infection control*, vol. 46, no. 12, pp. 1375–1380, 2018.

[64] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50, http://is.muni.cz/publication/884893/en.

[65] R. Rehurek, "Lda model parameters," Jul 2014. [Online]. Available: https://groups.google.com/forum/#!topic/gensim/aGXc0qiVBhU

APPENDIX A
TWITTER DATASET IN UTC TIME

TABLE V: Twitter Data Sets March 24, 2020 - April 8, 2020

Corpus	Time Start	Time End	Total Minutes	Size, GB	Total Tweets	No Retweets	Perc No Retweets
3/24/2020	2020-03-24 21:17:27+00:00	2020-03-24 22:00:48+00:00	44	1	132,658	27,374	20.64%
3/25/2020	2020-03-25 14:45:12+00:00	2020-03-25 16:18:47+00:00	94	2	286,405	63,649	22.22%
3/28/2020	2020-03-28 00:17:20+00:00	2020-03-28 02:01:08+00:00	105	2.3	317,498	61,933	19.51%
3/30/2020	2020-03-30 12:55:38+00:00	2020-03-30 21:44:35+00:00	530	11.5	1,618,620	365,808	22.60%
3/31/2020	2020-03-30 21:47:53+00:00	2020-03-31 13:15:36+00:00	929	20.3	2,802,069	576,741	20.58%
4/4/2020	2020-04-03 00:29:11+00:00	2020-04-04 22:05:12+00:00	2737	56.2	7,755,704	1,795,912	23.16%
4/5/2020	2020-04-05 20:41:43+00:00	2020-04-07 15:07:11+00:00	2547	49.4	6,810,216	1,599,455	23.49%
4/8/2020	2020-04-08 13:54:33+0000	2020-04-09 14:30:54+0000	1477	30.4	4,107,152	1,015,351	24.72%
Total			**8463**	**173.1**	**23,830,322**	**5,506,223**	**23.11%**

TABLE VI: Keyword Raw Counts

Corpus	bed	hospital	mask	icu	help	nurse	doctors	vent	test_pos	serious_cond	exposure	cough	fever
3/24/2020	147	1,323	1,685	139	114	215	372	1,143	28	1	11	18	1
3/25/2020	293	3,104	3,641	265	299	352	758	2,321	122	4	26	35	10
3/28/2020	191	3,218	3,607	180	248	504	891	4,073	101	2	19	19	3
3/30/2020	1,475	21,707	28,512	1,225	1,742	3,708	6,895	13,190	589	13	114	157	23
3/31/2020	1,959	28,495	67,703	1,344	3,416	5,233	9,671	16,717	948	13	141	459	137
4/2/2020	5,652	80,495	231,185	4,034	8,661	16,823	28,603	64,112	2,228	48	525	977	122
4/5/2020	5,648	81,025	159,915	6,350	7,741	14,767	27,341	45,614	2,612	36	445	786	133
Total	15,365	219,367	496,248	13,537	22,221	41,602	74,531	147,170	6,628	117	1,281	688	429

APPENDIX B
TOPIC MODELING IMPLEMENTATION DETAILS

For the LDA topic modeling, we used the gensim Python library [64, 65]. It provides four different coherence metrics. We used the "c_v" metric for coherence developed by Roder[22]. Coherence metrics are used to rate the quality and human interpretability of a topic generated. All models were run with the default parameters using a LdaMulticore model parallel computing on 20 workers, default gamma threshhold of 0.001, chunksize of 10,000, 100 iterations, 2 passes.

APPENDIX C
LIVE PRESS BRIEFINGS AND TOPIC TIME SERIES

Note - Sudden decreases in Figure 14 signal may be due to temporary internet disconnection.

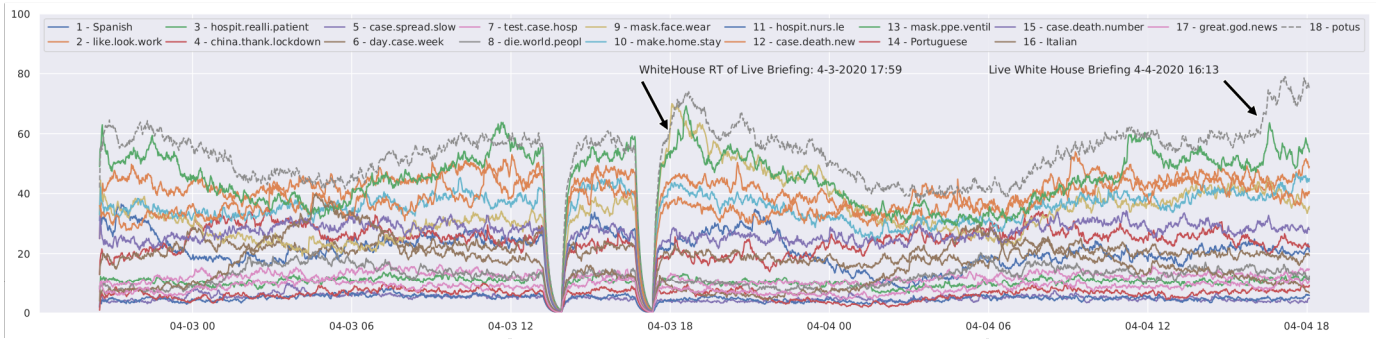


Fig. 14: April 3 8:29 PM EST to April 4 6:05 PM EST Topics Time Series

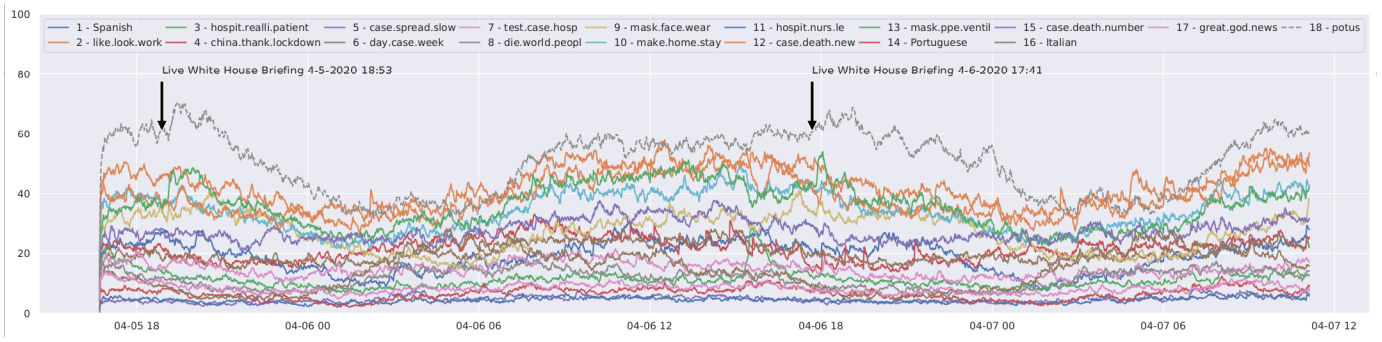


Fig. 15: April 5 4:41 PM EST to April 7 11:07 AM EST Topics Time Series

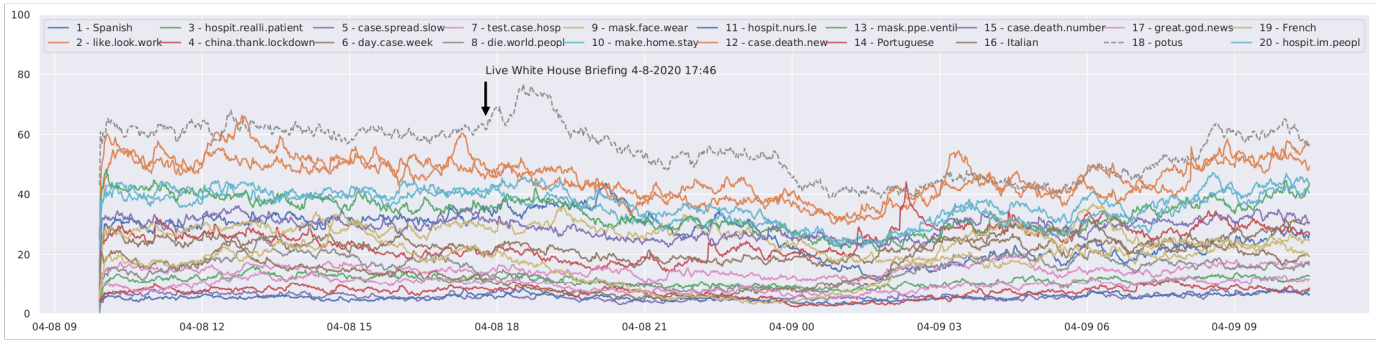


Fig. 16: April 8 9:54 AM EST to April 9 10:30 AM EST Topics Time Series

APPENDIX D
CHANGE POINT DETECTION TIME SERIES

Models were calculated using the ruptures Python package. We also applied exponential weighted moving average using the ewm pandas function. We applied a span of 5 for March 24, 2020 and a span of 20 for April 3 - 4 datasets, April 5 - 6 datasets, and April 8 - 9 datasets. Our parameters for binary segmentation included selecting the "l2" model to fit the points for Topic 18, using 10 n_bkps (breakpoints).

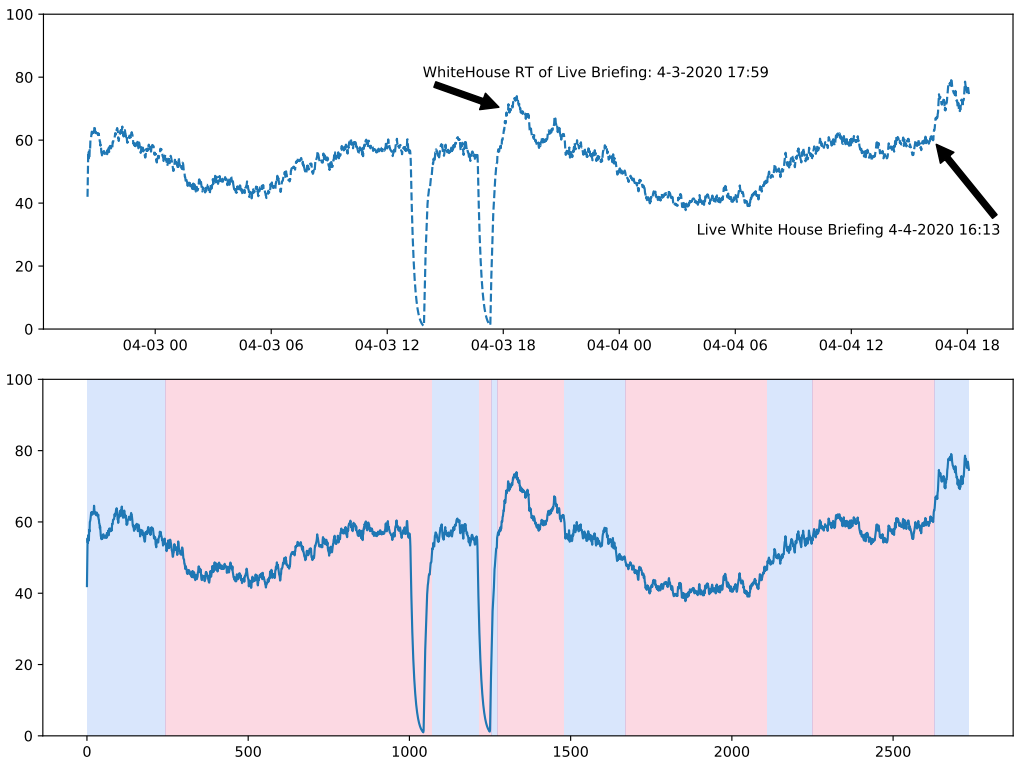


Fig. 17: Change Point Detection using Binary Segmentation for April 3 - 4, 2020

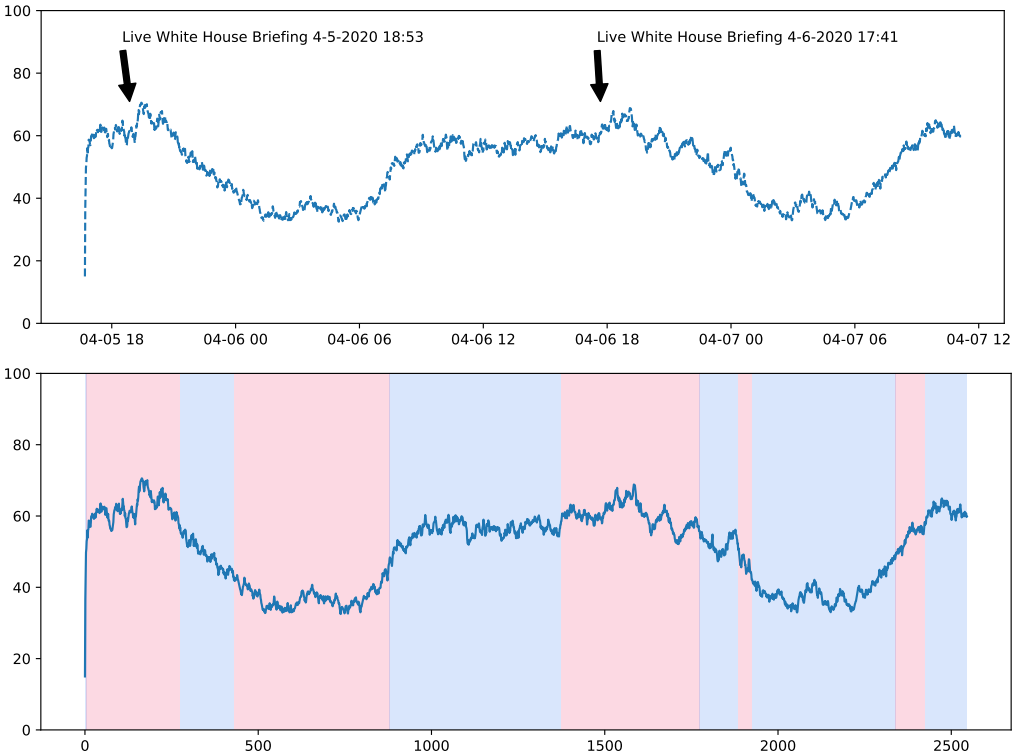


Fig. 18: Change Point Detection using Binary Segmentation for April 5, 2020

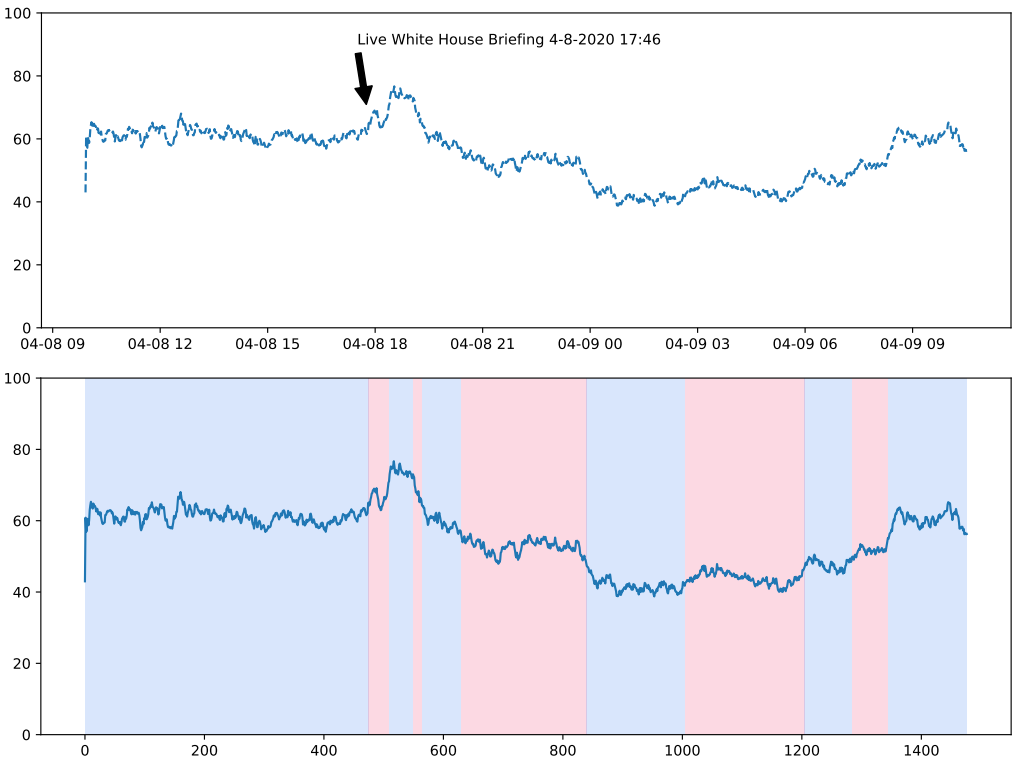


Fig. 19: Change Point Detection using Binary Segmentation for April 8, 2020

APPENDIX E
TF-IDF FREQUENCIES OF TWEETS RAPIDLY RETWEETED

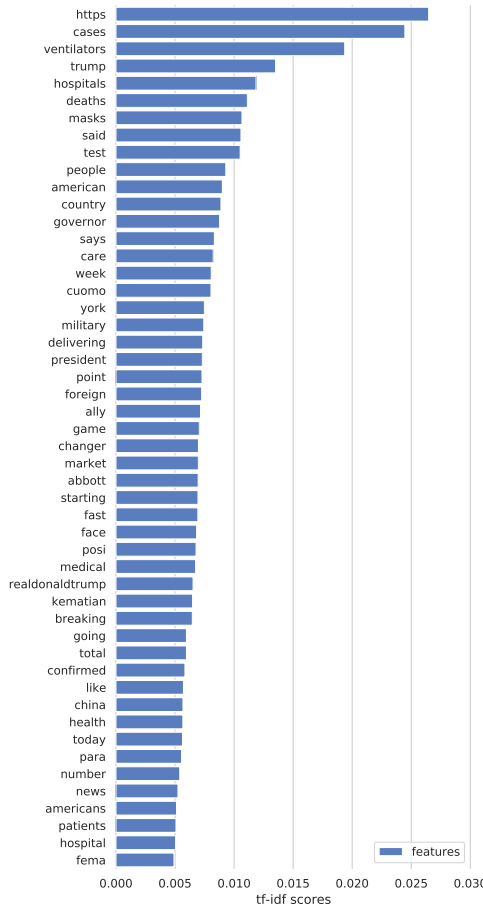


Fig. 20: TF-IDF Scores for Rapidly Retweeted Messages

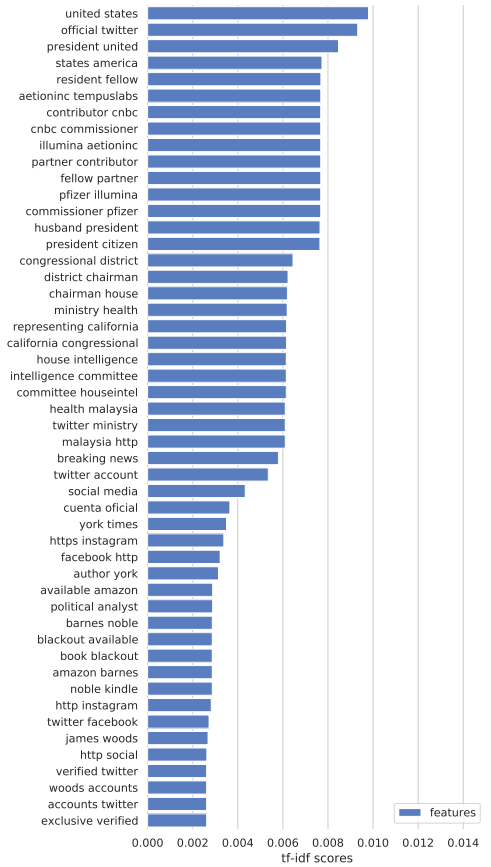


Fig. 21: TF-IDF Scores for Descriptions of Retweeted Users