Tracking Covid-19-related Trends on Social Media with Topic-Modeling

Lawrence Hessburg,
CSCI 5446,
lahe2512@colorado.edu

Abstract— **The 2020 Covid-19 pandemic, caused by the novel coronavirus SARS-CoV2 has been the most socially and economically disruptive event on a global scale since the World Wars. It also represents the first time such a large-scale global pandemic has occurred during the digital age, with much of the world's population immediately connected through social media. There are numerous epidemiological models that can be used to model the physical spread of virus itself on different scales, however an interesting area of research concerns the spread of ideas, feelings and sentiments relating to the virus and the resulting worldwide pandemic situation. In this paper, topical trends related to covid-19 pandemic were extracted and analyzed from Twitter data using Latend Dirichlet Allocation (LDA) topic modeling. The results show promise for future work to be done modeling the "spread" of these ideas between regionally-related areas of the Twitter-sphere through an epidemiological lens, and relating this model to the known spread of the physical SARS-CoV-2 virus itself.**

I. INTRODUCTION

The concept of viewing the spread of ideas through an epidemiological lens is far from a new one. The word "meme" was originally coined by evolutionary biologist Richard Dawkins in his 1976 book "The Selfish Gene" as a "unit of cultural transmission", an idea or concept spreading from person to person and group to group by "infecting" individual minds - "When you plant a fertile meme in my mind you literally parasitize my brain, turning it into a vehicle for the meme's propagation in just the way that a virus may parasitize the genetic mechanism of a host cell.". The emergence of the internet, and social media in particular, has acted to massively accelerate this process, providing an unbelievably fast and far-reaching medium for these ideas to spread.

The current covid-19 pandemic provides a unique situation for research in that it is the first global viral pandemic to occur on this scale since the advent of the internet. This raises the question of how ideas, sentiments and concepts relating to the physical pandemic spread online, and how this can be modeled and compared with the virus itself. The comparison between the physical and "digital" spread of covid-19 is a fascinating one, because the spread of ideas on the internet is not bounded by the same geographical limitations of physical space. However, this "data space" is not completely unbounded either, and although far less well-defined has its own "physical laws". For example, a concept is much more likely to spread quickly around the English-speaking Twitter-sphere before jumping to another space, such as the Italian-speaking twitter sphere. These regions can be *related to* physical regions in the "real world", however they are not strictly bounded by them except in unique cases such as China's intranet.

This project attempts to extract and analyze covid-19-related topics and trends from Twitter data, for future epidemiological region-based analysis using Latent Dirichlet Allocation (LDA) topic modeling. Given a set of text documents, LDA analysis is an unsupervised machine-learning technique for extracting a set of probability distributions of words in a given topic, and topics in a given document. These topics can then be analyzed both regionally and temporally to gain an understanding of the spread of ideas through the Twitter sphere.

A. Related Work

There has been much previous work in the area of analyzing the spread of ideas online from an epidemiological perspective across a wide range of domains including news [1], misinformation [2] and even internet memes [3]. The link to epidemiology is so deep-rooted that terms like "viral video" are even part of the colloquial internet vocabulary. However, none of this previous work has had the unique opportunity to analyze the spread of internet trends during the physical spread of a global pandemic.

II. METHODOLOGY

A. Data Collection

Individual tweets were gathered using the Twitter Premium Search API. Using the free "Sandbox Mode" feature, each Twitter Developer account is allowed 50 free requests to the Twitter Premium Full Archive search endpoint, and 250 free requests to the Twitter Premium 30-day search archive endpoint per month. After filtering out duplicates of retweets, a total of 34,246 unique tweets were collected ranging from 1 February 2020 to 3 May 2020. A distribution of the number of tweets gathered for each 7-day period within this timeframe is shown in Figure 1.

As shown in the figure, the distribution is heavily weighted towards the latter half of the time period. There are two primary reasons for this. First, because the Twitter 30-day search endpoint has more free requests available per month than the full-archive search, many more tweets were gathered from the period starting 30 days prior to the first day of data collection (13 April). Secondly, certain keywords like "covid-19", "covid" and "ncov-19" had either not been coined, or were not being frequently used on English Twitter during the month of February. There were an additional roughly 1000 tweets gathered from the latter half of January, however these

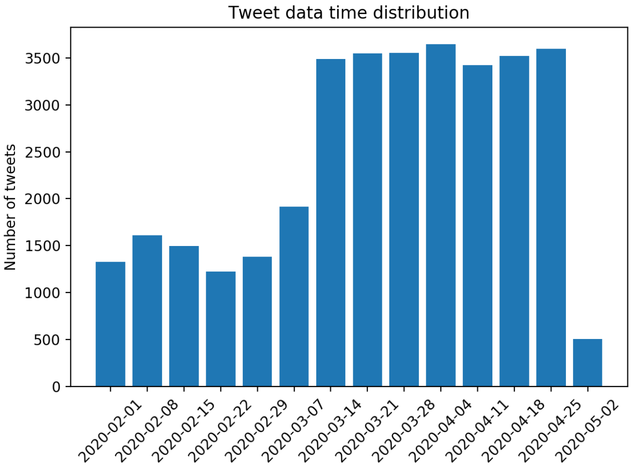


Fig. 1. Distribution of tweets in dataset by week

were not included in the dataset as they represented such a small portion of the overall data.

The search queries were constructed to get tweets containing at least one of the 8 keywords shown in Table I. These keywords were chosen to give a reasonably thorough representation of terms that may be present in tweets relating to the covid-19 pandemic.

Search Query Keywords
coronavirus
covid
covid-19
ncov-19
virus
flu
quarantine
self-isolation

TABLE I

The keywords were queried for individually, with a maximum of 100 tweets returned per query. Queries were distributed as evenly as possible throughout the search period. A distribution of keyword frequency in the final dataset is shown in Figure 2. The most common keyword was ”virus”, appearing in 11,008 unique tweets and the least common keyword was ”covid-19” appearing in 229 tweets.

The abnormally high frequency of ”virus” relative to the other keywords is likely due to the common uses of the phrase ”corona virus” instead of ”coronavirus”, as well as many tweeters referring to ”the virus” or ”this virus” when addressing Covid-19. Examples are shown in Figure 3.

B. Data Preprocessing

In order to prepare the tweet dataset for LDA topic modeling, several preprocessing steps were taken. First, the text contained in each tweet was tokenized using the SpaCy Natural Language Processing library. Each tweet was reduced to a set of individual tokens (words) in their lowercase

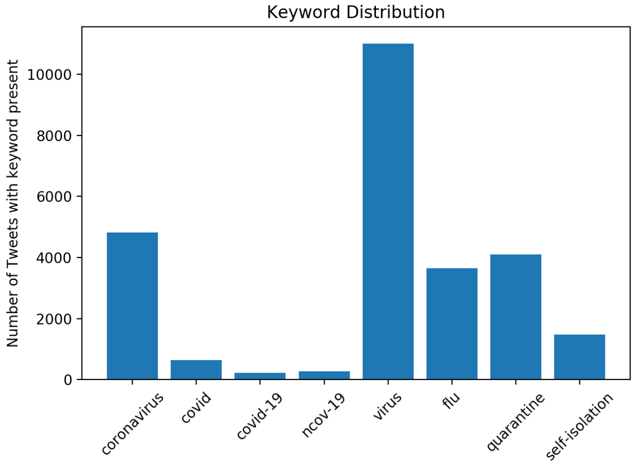
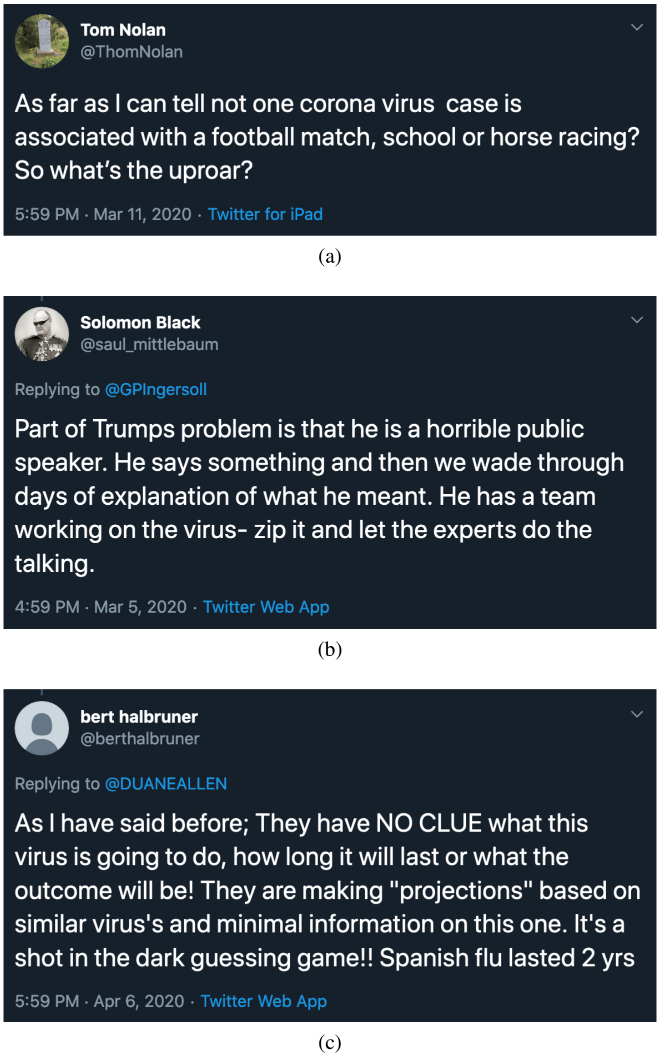


Fig. 2. Distribution of search query keywords in dataset by number of tweets



Tom Nolan
@ThomNolan

As far as I can tell not one corona virus case is associated with a football match, school or horse racing? So what’s the uproar?

5:59 PM · Mar 11, 2020 · Twitter for iPad

(a)

Solomon Black
@saul_mittlebaum

Replying to @GPIngersoll

Part of Trumps problem is that he is a horrible public speaker. He says something and then we wade through days of explanation of what he meant. He has a team working on the virus- zip it and let the experts do the talking.

4:59 PM · Mar 5, 2020 · Twitter Web App

(b)

bert halbruner
@berthalbruner

Replying to @DUANEALLEN

As I have said before; They have NO CLUE what this virus is going to do, how long it will last or what the outcome will be! They are making "projections" based on similar virus's and minimal information on this one. It's a shot in the dark guessing game!! Spanish flu lasted 2 yrs

5:59 PM · Apr 6, 2020 · Twitter Web App

(c)

Fig. 3. Examples of the many common uses of the word ”virus” in the dataset

form with all special characters removed, resulting in a large dataset of unique tokens.

Next, several filters were applied to the dataset. Non-word tokens such as urls, @ mentions and emojis were removed from the dataset. Additionally, English ”stop words”, or the most commonly used words in the language (is, and, the, etc..) were filtered out of the dataset as these words are nearly universally present but do not contribute to the meaning or subject matter of the text. The keywords used in the intial search queries were also filtered out of the dataset. This is because since at least one keyword is present in each tweet, the keywords themselves will heavily influence the topics generated by the LDA algorithm. By removing them, the goal is to extract topics with subjects *related to* the covid-19 pandemic rather than topics categorizing the virus itself.

The final step was to use the Python Natural Language Toolkit Lemmatizer to attempt to reduce each token to its root form. For example, the words ”running”, ”run” and ”runs” would all be counted towards the token ”run”.

After applying the aforementioned preprocessing steps, the final dataset consisted of 31,185 unique tokens. A histogram of the 30 most frequent tokens can be seen in Figure 4. All of the most common words appear to both have significant meaning, and are related to covid-19, suggesting the preprocessing procedure was successful.

C. LDA Topic Modeling

As mentioned previously, the topic modeling was performed using the Latent Dirichlet Allocation algorithm. [4], [5] LDA is an unsupervised algorithnm for extracting a predefined number of topics from a collection of documents by operating under the assumption each document was generated randomly with some distribution of topics, each of which is itself represented as a distribution over all individual words. LDA attempts to reverse-engineer this generative process, providing a word-distribution for each topic and a topic-distribution for each document. In this case, documents consisted of the text contained in individual tweets. For the sake of brevity, this paper will not contain a detailed description of the LDA algorithm, however the high-level description is as follows:

```
$K \leftarrow$ number of topics
$V \leftarrow$ number of words in dictionary
$M \leftarrow$ number of documents
$D \leftarrow$ a corpus (collection) of $M$ documents
$\alpha_{k=1...K} \leftarrow$ prior weight of topic $k$ in a document
$\beta_{w=1...V} \leftarrow$ prior weight of word $w$ in a topic
$\varphi_{k=1...K} \leftarrow$ distribution of words in topic $k$
$\theta_{d=1...M} \leftarrow$ distribution of topics in document $d$
$\mathbf{z}_{d=1...M} \leftarrow$ identity of topics of all words in document $d$
Approximate the joint posterior probability $P(\theta_{1:M}, \mathbf{z}_{1:M}, \beta | D; \alpha_{1:M}, \varphi_{1:K})$ iteratively through variational inference.
```

This process was performed using the python Gensim library LdaModel functionality. The algorithm was run over the corpus of tweet documents, varying the given number of topics K between 5 and 15, and the resulting esimates of the distributions θ and φ were saved for each run. The range of K values chosen is somewhat arbitrary, and was meant to reflect a reasonable guess at the range of how many distinct, high-level topics relating to covid-19 should be present within the tweet dataset.

D. Analyzing Trends

In order to analyze the trends of topic relevance over time, the tweet dataset was split into subsets containing tweets from each 1-week period within the overall dataset's timeframe. Each subset's tweets were then analyzed using the previously-calculated φ distributions to get an overall probability distribution of the topics present in the subset. By viewing the change in probability of a given topic over time, and analyzing the distribution of words within the topic qualitatively, trends in the discussion relating to covid-19 can be extracted from the tweet dataset.

III. Results

After running the LDA algorithm for each value of K, it was necessary to qualitatively analyze the extracted topics to determine the optimal value of K for the given tweet dataset. This was done using the pyLDAvis library [6], which provides a useful set of tools for LDA analysis beyond simply examining the probability distribution of words within each topic. An screenshot from an example of the pyLDAvis interface can be seen in Figure 5.

The interface provides a map of intertopic distance by plotting each topic in two axes using Principal Components Analysis on topic-space. If there are large clusters of topics close together in the intertopic distance map, this is a sign that the topics share similar word distributions, meaning the value of K is probably too high.

For each topic, pyLDAvis displays a list of the 30 most relevant terms within each topic. Relevency is calculated using a weighted sum of the posterior probability of each word given the selected topic, and the likelihood of each word appearing across all topics. The weight for this sum λ can be adjusted using a slider within the tool, and the authors suggest a value of $\lambda = 0.6$. By viewing the distribution of the most relevant words within each topic, one can qualitatively distinguish whether the given words are relating to the same subject matter, or a mix of multiple subjects (suggesting the K value is probably too low).

After performing this qualitative analysis across all values of K, it was determined that $K = 10$ was the optimal value for the tweet dataset. The pyLDAvis interface for the tweet dataset for $K = 10$ can be viewed and interacted with at https://pai-sho.github.io/CSCI-5423-Project/model10.html. A selection of some of the most frequent terms in each of the 10 topics, along with a high-level description of the topic's apparent subject matter can be found in Table II. Many important and very easily distinguishable covid-related topics seems to be present in the 10-topic model, including china, president Trump, the healthcare system and social distancing. It is interesting to

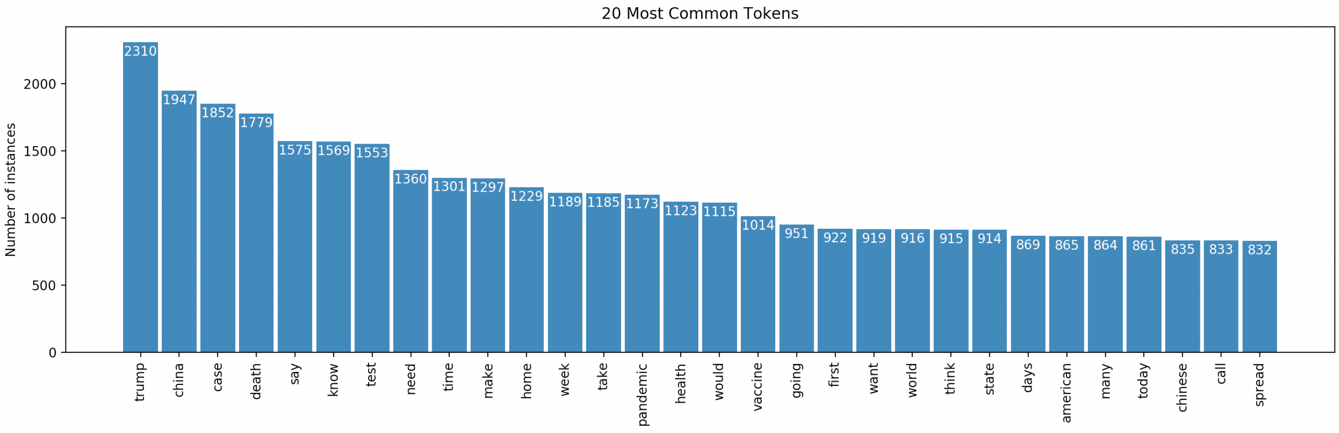
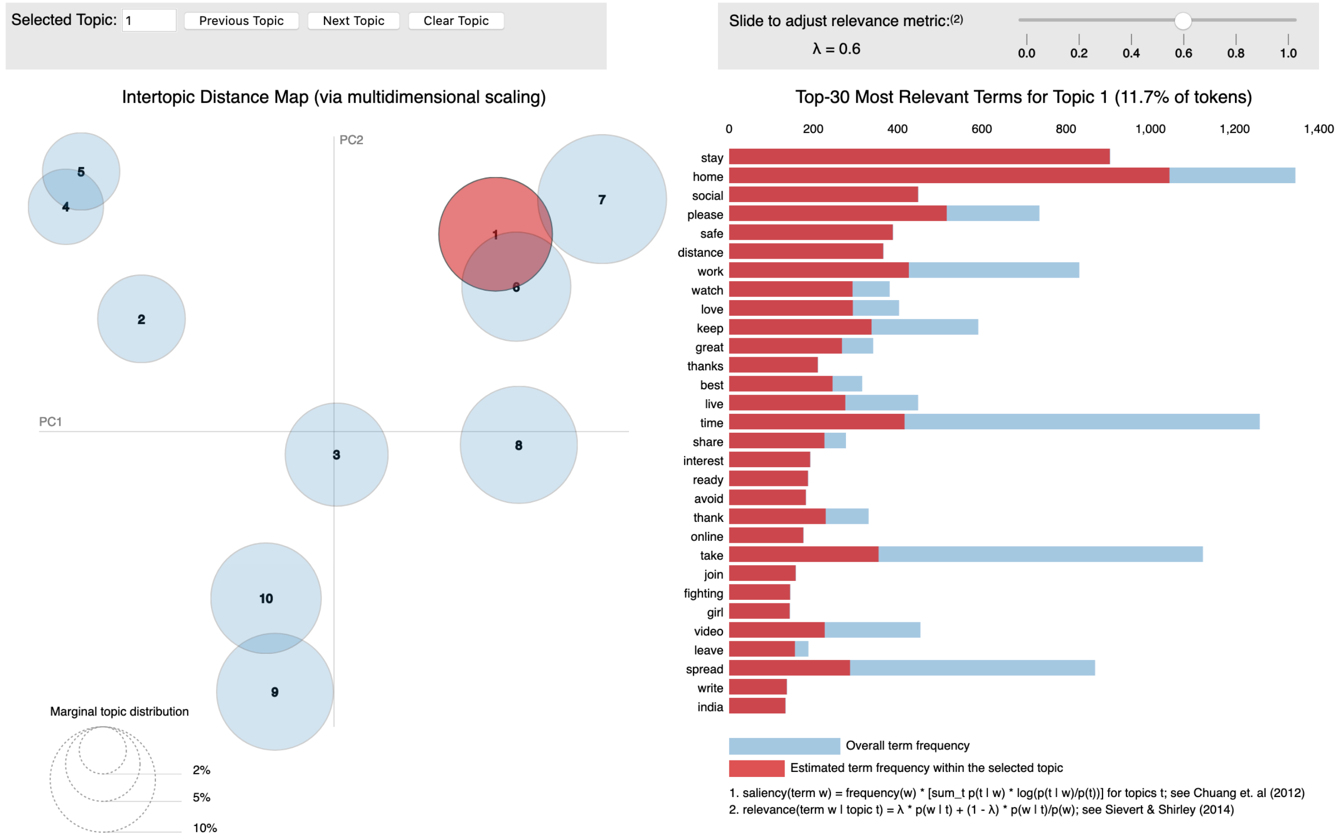


Fig. 4. The 30 most common tokens in the dataset after filtering and lemmatizing.



Selected Topic: 1

λ = 0.6

Overall term frequency

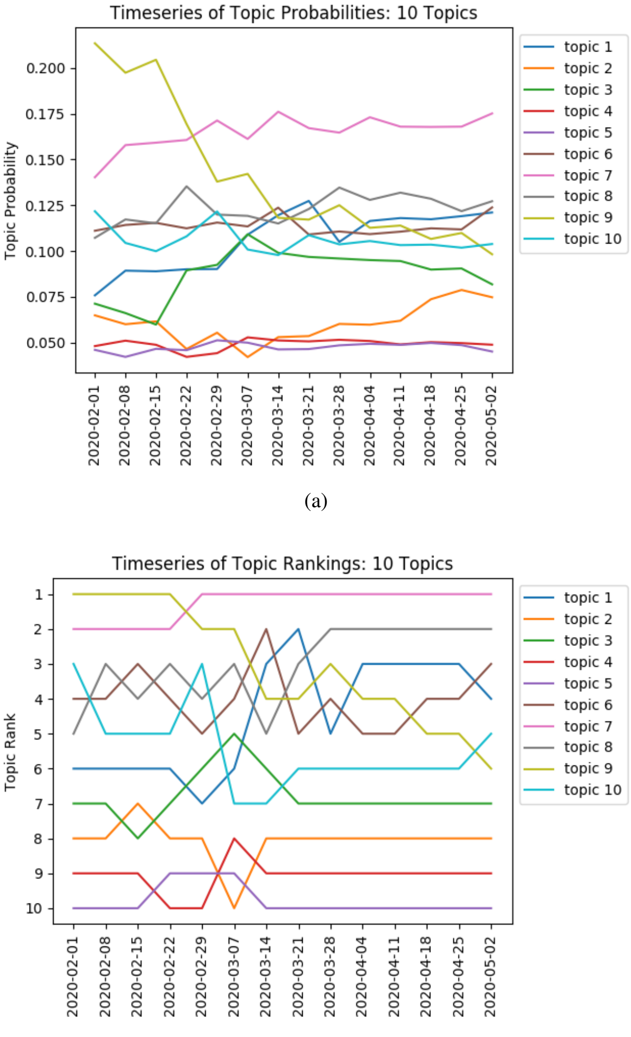
Estimated term frequency within the selected topic

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Fig. 5. An example of the pyLDAvis interface

note the presence of an apparent ”rest class” of mostly seemingly meaningless words such as ”know”, ”think”, ”really”, ”would” and ”could”. This would suggest a K value slightly too high, however reducing K to 9 resulted in undesirable combination of topics such as China and Vaccine into one larger topic. Overall, $K = 10$ qualitatively gave the best balance between distinguishability between and relevance of individual topics.

After the models were examined and $K = 10$ was chosen as the optimal value, the timeseries analysis described above was performed for each week in the dataset using the probability distributions extracted by the $K = 10$ LDA model. A plot of the topic probabilities for each week-long subset of the tweet dataset can be found in Figure IIIa. For a slightly different viewpoint, the ranking of each topic by probability across the time series is shown in FigureIIIb.

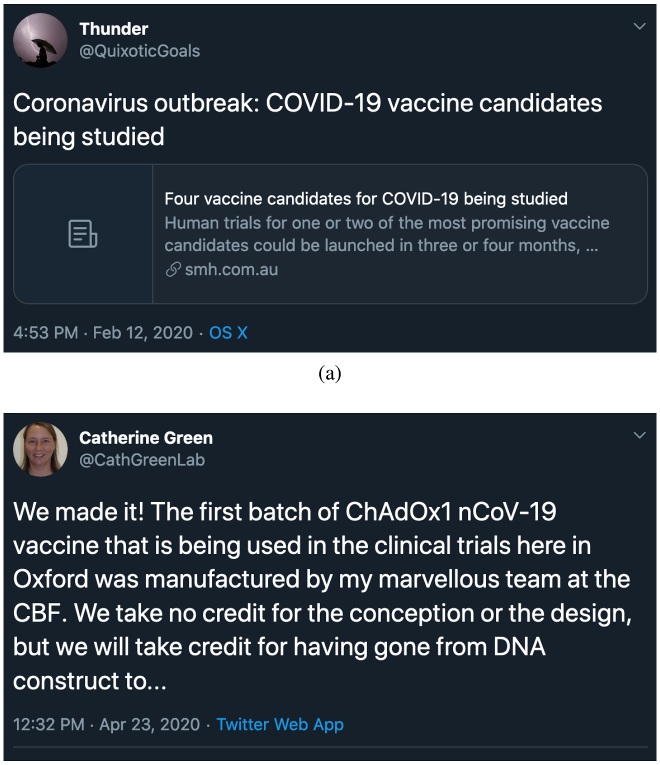


(a)

(b)

Fig. 6. (a) The probability distributions for the 10-topic LDA model, plotted over each week-long subset of the tweet dataset. (b) The ranking of each topic by relevance over each week-long subset of the tweet dataset.

There are several trends readily apparent in the timesearies visualization. First of all, In the first few weeks of the dataset, the discussion seems to be mainly focused towards topic 9, which sharply falls off in Late February. This makes sense, as topic 9 seems to mainly concern the initial outbreak in China, and as covid-19 evolved into a worldwide pandemic, the discussion became much less focused on China in general. Around the same time, topic 3 (Donald Trump and his pandemic response) experienced sharp growth, before slowly declining for the latter half of the time period. This indicates a period of highly increased discussion around Trump and his refusal to acknowledge covid-19 as a serious threat as the virus started to become a global problem. Topic 1 (related to social distancing) experienced sharp growth during the month of March as talks of stay-at-home orders and social distancing measures rose in the english-speaking Twitter space. Topic 2 (related to vaccines) fell in relevance during the first several weeks before gradually rising with a sharp increase in mid-April. This is likely due to the fact that in the early days of the pandemic, there was much discussion that a vaccine would soon be created. This fell as the problem grew worse, but was reinvigorated by Oxford University’s announcement of successful manufacturing of the candidate vaccine ChAdOx1. Examples of vaccine-related tweets from both time periods can be found in Figure 7


Thunder
@QuixoticGoals

Coronavirus outbreak: COVID-19 vaccine candidates being studied

Four vaccine candidates for COVID-19 being studied
Human trials for one or two of the most promising vaccine candidates could be launched in three or four months, ...
smh.com.au

4:53 PM · Feb 12, 2020 · OS X

(a)

Catherine Green
@CathGreenLab

We made it! The first batch of ChAdOx1 nCoV-19 vaccine that is being used in the clinical trials here in Oxford was manufactured by my marvellous team at the CBF. We take no credit for the conception or the design, but we will take credit for having gone from DNA construct to...

12:32 PM · Apr 23, 2020 · Twitter Web App

(b)

Fig. 7. Examples of tweets relating to vaccines from Febrary 12 and April 23

The aforementioned ”rest class” remains fairly consistent and very relevant throughout the timeseries, suggesting that better results may be obtained by adding additional stop words to the filtering process.

Topic Number	Topic Keywords	Topic Subject
1	stay, home, safe, please, social, distance	Social Distancing / Quarantine
2	vaccine, chadox1, human, oxford, scientist	A Covid-19 Vaccine (Specifically ChAdOx1 being developed by Oxford University)
3	trump, president, panic, response, blame, fear, market	President Donald Trump's handling of the pandemic and its effects
4	total, problem, press, yesterday, tomorrow	The media coverage of the pandemic
5	body, attack, survive, catch, feeling, f**king	The effects of covid-19 on the body once caught
6	week, day, month, family, friend, toll, miss, kid	The social effects of the increasingly extended mitigation efforts
7	know, think, need, make, would, really, could	A "rest class" of words that aren't necessarily stop words, but have little meaning
8	kill, die, death, million, rate, country, mortality	Projections, predictions and fears related to overall death toll
9	china, wuhan, outbreak, case, update, travel, test	The initial outbreak in Wuhan, China and its potential spread
10	hospital, doctor, patient, worker, mask, supply, help	The strain on the medical system, medical workers and other essential workers

TABLE II

IV. CONCLUSIONS

This work provided a successful proof-of-concept for analyzing topical trends related to covid-19 from twitter data, however there are several major areas for improvement and potential future work. First of all, the roughly 35,000 tweets gathered for this project represent an extremely small fraction of the overall covid-related discussion in the given time period, which is likely on the order of tens of millions of tweets. Using Twitter's paid premium search API to gather at least a few orders of magnitude more data would be extremely helpful in both extracting a higher number of more fine-grained topics, and analyzing the trends of said topics on much smaller timescales. This would also allow for more keywords to be included in the search queries, and expansion of the dataset to include non-english tweets. Secondly there are several other popular topic modeling techniques such as Latent Semantic Indexing and Non-Negative Matrix Factorization which can be applied and compared to better understand the dataset. Lastly and perhaps most importantly, the Twitter paid premium search API allows for additional filtering during search - namely the ability to only include tweets that contain location metadata. By constructing a dataset of exclusively geotagged tweets, similar analysis could be performed with the LDA-extracted probability distributions *regionally* in addition to temporally. In this way, the "spread" of different topics in relevance from regionally related areas of the overall twitter-sphere. It would be extremely interesting to analyze these trends from an epidemiological perspective, and relate it to the spread of the physical virus itself.

REFERENCES

[1] Fang Jin, Edward Dougherty, Parang Saraf, Yang Cao, and Naren Ramakrishnan. Epidemiological modeling of news and rumors on twitter. In *Proceedings of the 7th workshop on social network mining and analysis*, pages 1–9, 2013.

[2] Marcella Tambuscio, Giancarlo Ruffo, Alessandro Flammini, and Filippo Menczer. Fact-checking effect on viral hoaxes: A model of misinformation spread in social networks. In *Proceedings of the 24th international conference on World Wide Web*, pages 977–982, 2015.

[3] Lin Wang and Brendan C Wood. An epidemiological approach to model the viral propagation of memes. *Applied Mathematical Modelling*, 35(11):5442–5447, 2011.

[4] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[5] Jonathan K. Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

[6] Carson Sievert and Kenneth Shirley. Ldavis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pages 63–70, 2014.