Understanding the pandemic: An exploration of the public perception of COVID-19 in the United States using topic modeling and event-sentiment analysis on Twitter data

Asher R. Schrijvers[a], Joey M.A. Spronck[a] and Ioannis Konstantakoupoulos[a]

[a] *Utrecht University, Science Park de Uithof 3584 CE, Utrecht, The Netherlands*

Abstract

COVID-19 is a novel contagious disease that has rapidly spread across the globe and has caused a worldwide pandemic. The impact of COVID-19 encompasses many aspects of society, such as healthcare, the economy and the socio-physiological well-being. The unprecedented impact and scale of the pandemic complicates governmental decision making, which requires insights in all aspects related to the crisis, including the public perception. In this study the public perception on the COVID-19 crisis was explored in a corpus of 127,128 COVID-19 related tweets from the United States. A combination of topic modeling and sentiment analysis was used to identify sudden shifts in discussed topics and their sentiment. This paper shows that using our methodology, we are able to identify event-related fluctuations in the public perception on the ongoing pandemic. With additional proposed improvements and follow up research, governmental institutes may use similar research strategies to gain additional insights for future decision making.

Keywords

COVID-19, twitter data, natural language processing, topic modeling, sentiment analysis, event analysis

1. Introduction

As of January 21th, 2021 the total number of people that have contracted COVID-19 is over 90 million, of which over 2 million people have died, and 70 million have recovered. In other words, around 10% of the world's population has contracted COVID-19, with a mortality rate of 3,5% . The increasing total amount of cases has seen no steep declines globally [1]. Furthermore, the effects of COVID-19 are not limited to its symptoms. According to the WorldBank [2], the impact of the pandemic is forecast to contract 5.2% in global GDP, and the deep recessions triggered by the pandemic are expected to leave lasting scars. Next to the effects on the public heath and the economy, the physiological and social impact of COVID-19 is also becoming more prevalent [3].

✉ Asherasherasher@hotmail.com (A.R. Schrijvers); joeyspronck@gmail.com (J.M.A. Spronck); gianniskwnstantakopoulos@hotmail.gr (I. Konstantakoupoulos)
🌐 https://github.com/Asherrav (A.R. Schrijvers); https://github.com/JoeySpronck (J.M.A. Spronck); https://github.com/Konpoulos (I. Konstantakoupoulos)

CEUR Workshop Proceedings (CEUR-WS.org)

With the unprecedented impact and scale of COVID-19, governmental organizations now cope with the task to tackle consequences of the pandemic effectively. Therefore, there is a large requirement for insights in the impact of COVID-19 on healthcare, the economy and the socio-psychological well-being. Many insights for governmental decision-making are based on numerical data, such as the infection rates, the number of occupied intensive care beds, and the amount of companies requesting financial aid. However, the impact on socio-psychological well-being and the public perception of the ongoing pandemic are harder to capture with these kinds of statistics. As such, there is a requirement for more elaborate research methods in order to quantify these kind of impacts.

Over the last decade, it has become evident that with the dawn of digital public spaces such as Facebook, Twitter, Youtube, etc. the way people express their views and opinions has changed [4]. Every day, a large amount of tweets are posted in the form of text, and this rich source of data is soaked with sentiment [4]. In order to help health professionals and governmental institutions gain insight in the emotional reactions and the feelings expressed by their civilians, twitter data can be analyzed to reveal frequently discussed topics and sentiment over time, so that they can benefit from that data. One example of an insightful finding comes from Xue et al., [3] who aimed to understand twitter users' discourse and psychological reaction to COVID-19 using topic modeling and sentiment analysis. The findings of this research pointed towards a negative trend in the sentiment of discussed topics, and the authors reported that fear of the unknown nature of COVID-19 was dominant across all topics. In order to make recommendations for battling the impact of COVID-19 on the societal well-being, and tailoring these insights to a specific situation or event, the current study will explore further applications of topic modeling and sentiment analysis of twitter data.

This study is one among many studies modeling twitter topics [3, 5, 6] and analyzing twitter sentiment [3, 5, 7], and the use of such techniques have been evaluated and validated [4]. However, among these studies there is a general emphasis on topical shifts and gradual changes of sentiment, and more thorough inspections of large fluctuations during small periods have yet to bet conducted. With our research, we are therefore interested in detecting rapid changes in the amount of tweets that are being posted related to meaningful topics in relatively small time periods. We will then explore the sentiment of such a topic during the relevant time period by visualising the most distinctive words for negative and positive tweets. We propose that by comparing these changes of meaningful topics to plausibly related events for the respective time period, we can assess how citizens respond to certain events and evaluate the effect of an event on the public perception of the pandemic situation and government's response strategies (i.e the trustworthiness of a government after they announced a response strategy; how citizens estimate the impact of the situation etc.).

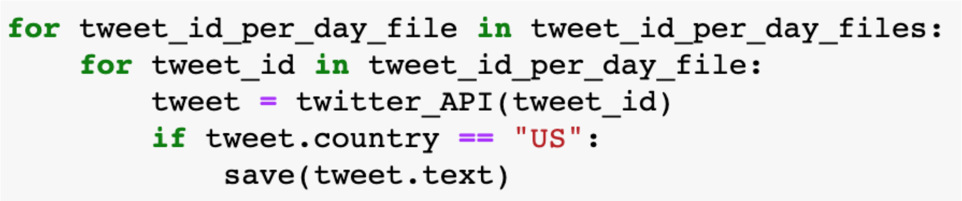
The current framework of Twitter analysis on public perception of the ongoing COVID-19 pandemic is relatively novel. However, with this framework, it has already been established that twitter discourses can be used to explore frequently discussed topics, their respective sentiment and shifts thereof. To expand on this framework, we aim to further explore (1) how the public perception in terms of discussed topics and their respective sentiment in COVID-19 related tweets changes over time, and (2) explore the possibilities of using these methods to identify events that can be related to sudden fluctuations of topic frequency and their respective sentiment.

2. Background

In recent times, there have been many studies modeling twitter topics and analyzing twitter sentiment around COVID-19-keywords. Boon Skunkan (2020) [5] sought to increase understanding of public awareness of COVID-19 trends and uncover ”meaningful themes of concern” posted by twitter users by using topic modeling and sentiment analysis on tweets. They showed that this approach produces useful information about the trends in the discussions concerning COVID-19. Another such research, conducted by (Xue, J., et al., 2020) [3] aimed to understand twitter users' discourse and psychological reaction to COVID-19 using LDA for topic modeling. The findings of this research also point towards a negative trend in the sentiment of discussed topics, and the authors report that fear of the unknown nature of COVID-19 is dominant across all topics. Kaur, S., Kaul, P., Zadeh, P.M. (2020) [7] took on another perspective, and explored the dynamics and flow of behavioural changes among twitter users during the COVID-19 pandemic by looking at emotional changes in tweets across three time intervals characterised by infection rates and mortality rates. Most of these studies focus primarily on shifts in topical trends and either sentiment or emotions, but no mentions are made of sudden fluctuations of discussion frequencies of certain themes and their sentiment. Chen, Lerman and Ferrara (2020)[6] , however, describe a large multilingual COVID-19 Twitter dataset and showcase some visualisations of descriptive statistics. Interestingly, they found that Twitter discourse statistics reflect major events at those times. They identified these events of interest by leveraging Business Insider, NBC and CNN released timelines. With this research, Chen, Lerman and Ferrara showed that events from news timelines could be reflected in the fluctuations of tweets containing specific substrings such as ”wuhan” or ”covid”. While the timelines by Chen et al.[6], allow for capturing of relatively large changes in a small time period, the substrings they showcase are not informative beyond the fact that they show the level of public awareness of the pandemic situation.

3. Data collection

The dataset used for this study was collected by sampling from an open-access dataset from IEEE-dataport [8, 9]. Because the most sophisticated methods for the purpose of text analysis are widely available for the English language, this dataset was selected. Furthermore, this was the largest English subset with available Geo-tags. The IEEE-dataport dataset consisted of N English COVID-19 related tweet id's, accompanied with Geo-tags. To scrape the tweet text of these tweet id's, the python library ”tweepy”[10] was used. To retrieve these tweet texts, Twitter developer accounts were created so that an application programming interface (API) could be accessed. Only tweets with the Geo-tag ”US” were selected. The scraping pseudo code is shown in Figure 1. After scraping, a sample of 127,128 tweets was obtained, and each tweet id was accompanied by date, tweet text and a sentiment score computed by the sentiment model used by Lamsal (2020) [9].



```
for tweet_id_per_day_file in tweet_id_per_day_files:
    for tweet_id in tweet_id_per_day_file:
        tweet = twitter_API(tweet_id)
        if tweet.country == "US":
            save(tweet.text)
```

Figure 1: Pipeline of tweets collection using the tweepy API in python

4. Methods

4.1. Overview

In our data analysis, we aim firstly to apply topic modeling to our corpus, and to analyze the sentiment of tweets that are related to meaningful COVID-19 related topics found. The next step in our analysis will then be to visually inspect time-series plots of these topics and sentiment scores. With this inspection, we are looking for a relatively small amount of topics that can be deemed relevant and we try to identify fluctuations in the tropic frequencies and ins sentiment that can be linked back to topic related events in the corresponding time frame. Once we have identified possible topic-related events, we will analyze the sentiment of the tweets related to that topic in that time period by plotting the amount of positive and negative topic tweets over time, and, creating wordclouds for the most distinctive words between the most positive and most negative topic tweets. Lastly, we will inspect whether this discourse of tweets can be explained by looking at the events in the COVID-19 time line.

4.2. Topic modeling

In order to apply topic modeling to the corpus, Latent Dirichlet Allocation (LDA) was used. LDA uses sampling out of the Dirichlet distribution to create a text with the specific polynomial distribution.[11] Then, the tweets may be viewed as a sample of many topics where each tweet is considered to have a set of topics allocated to it by the LDA model.

4.2.1. Preprocessing

Before applying LDA, the corpus was preprocessed by means of lemmatizing, removing stop words and punctuation signs. It was examined whether different part of speech (POS) selections yielded more coherent LDA output. Here we compared the selection of nouns, verbs and adjectives with the selection of nouns only. This procedure was done in a Jupyter Notebook, and the package ”SpaCy”[12] was used for the preprocessing steps described.

4.2.2. LDA settings

After preprocessing the corpus, 6 different LDA models were created. Here, the LDA was compared to the output of the two POS selections and the amount of topics. For each POS selection an LDA model was trained for 1000 iterations to find 20, 35 and 50 topics. The LDA output of these models were evaluated (LDA output evaluation section), and the optimal model was trained for a total of 10000 iterations. Eventually, the LDA model was provided with the POS selection containing nouns, verbs and adjectives[13].

4.2.3. LDA output evaluation

For the selection and evaluation of the trained LDA models, criteria were set for identification of themes that we expected to find. Here, it was expected to find themes related to: lockdown periods, vaccines, effects of COVID-19 on economical and educational structures, and COVID-19 measures introduced by the government. Furthermore, the additional determined topics were manually inspected to identify topics that were over looked in our evaluation criteria. The interpretability of the top 10 words that are associated with the topics were also evaluated. Since we are interested in the effect of certain events on the proportion of tweets associated with the respective topic, the found topics were further inspected by plotting the topic proportions over time (next section). Here we visually inspected whether we were able to identify sudden shifts in the topic proportions that could by explained by certain events. The model that yielded the most interpretable topics and showed the most sudden fluctuations in topic proportions was deemed the most optimal and fitting model for our research goals.

4.3. Time series analysis

After creating the LDA model, a time series analysis was conducted to get an overview of the fluctuations of topics and their sentiment over time.

4.3.1. Preprocessing

In order to visualise the frequency of the discussed topics over time, the topic distribution output required additional preprocessing. In Figure 2, an example of the topic proportion preprocessing of a single tweet is shown. The lower bound of the topic proportion output (Fig. 2, top) was subtracted from the individual topic proportion values, and the remaining values were rescaled to sum up to 1 (Fig. 2, mid). Eventually, the individual topic proportions larger than 0, were set to one. Here, we assume that a tweet is 100% associated with a topic, if the individual topic proportion is bigger than the lower bound. This results in a binary topic association for each tweet, where each tweet can be associated with multiple topics (Fig. 2, bottom).

4.3.2. Time series figures

In the time series figures (Fig. 4, 5 and 6), the topic frequency, the topic sentiment mean and the topic sentiment standard deviation per day are shown. The topic frequency was calculated by taking the mean of the binary topic association per day, for each topic. In order to gain insights into the general sentiment fluctuations and the spread of the sentiment values for each day, the mean sentiment and its standard deviation were calculated. To calculate the mean and standard deviation of the sentiment of a topic, only the tweets that were positively associated to the topic were taken, and the mean and standard deviation were calculated per day. To reduce the noise in these figures and allow easier event identification, these values were plotted after

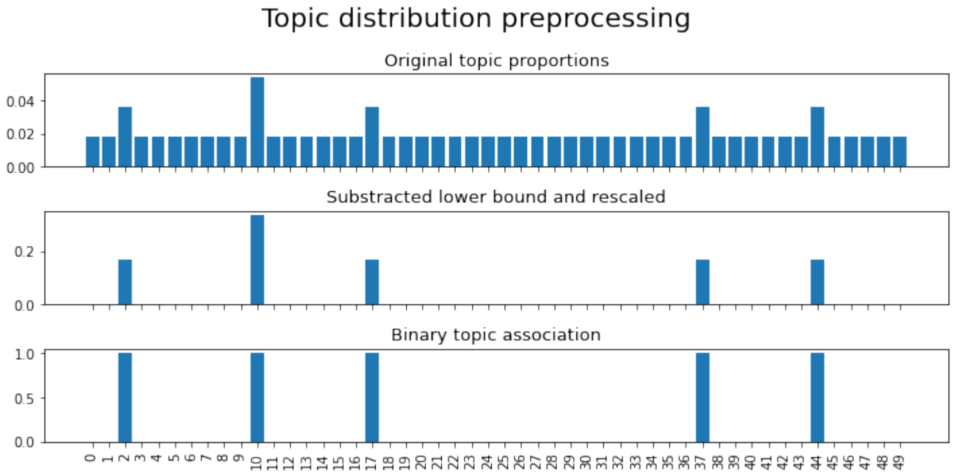


Figure 2: Topic distribution re-coding

calculating the rolling mean of 7 days. Additionally, the time series figures are accompanied with a wordcloud of the 10 most important words for the corresponding topic, where the size of the word is related to the importance of the word.

4.4. Event analysis

4.4.1. Event sentiment time series figures

After a possible identification of sudden topic frequency and/or sentiment shifts the time series plots, a manual search on the web was done to relate this sudden shift back to a possible causal events. Here, events are defined as any planned public occasion related to COVID-19 that applied to the United States. Furthermore, a more in depth analysis is done on a subset of the data (from here onward named ”event subset”). This subset contains the relevant time period for the defined event, and only contains the tweets that are positively associated to the topic in which the sudden shift was detected. Both the number of tweets with a positive as a negative sentiment are plotted within this time frame. Here, the raw counts are plotted, in contrast with with the rolling mean values in the time series data. Furthermore, the possible causal event are depicted by vertical lines on the corresponding date.

4.4.2. Event sentiment-related words analysis

To gain additional insights into the sentiment of the discussed themes, the words in the tweets of the selected subset were investigated further. The tweets were preprocessed in a similar manner as the LDA preprocessing, however, only nouns were selected. The first and third quartiles were calculated for the sentiment scores of the tweets in the event subset. Subsequently, the most distinctive words were calculated between the most positive and the most negative sentiment quarters, using the log likelihood ratio (llr) metric[14]. The most distinctive words for both the most positive and most negative sentiment were again plotted in a wordcloud, where the size of the word is related to the llr value.

5. Results

5.1. Topic modeling

The outputs of the 6 trained LDA models were evaluated according to our evaluation criteria. It was found that an LDA model with 50 topics and the POS selection with nouns, verbs and adjectives resulted in the highest number of expected themes found, with the best interpretability of the most important keywords. Furthermore, an inspection of the topic frequency over time showed that these settings showed more sudden shifts, in contrast with mainly slow shifts, that are more probable to relate back to events, which again supports the chosen LDA settings.

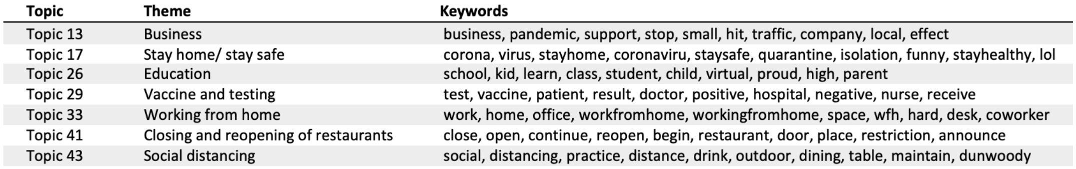
The output of the most important keywords for each of the 50 topics of the final LDA model is shown in the appendix A. For this study, a selection of 7 clear COVID-19 related topics were identified. In Fig. 3, a summary of these named topics extracted from the selected topics can be found. As can be seen in this figure, words such as ”positive” and ”negative” in topic 29, and ”close”, ”open” and ”dining” in topic 41, add up to the interpretability of these topics, supporting the decision to include verbs and adjectives in the POS selection.

From the LDA model with 50 topics and 10,000 iterations all topics were individually inspected and evaluated. Among those 50 topics, the following 7 themes emerged as can be seen in Figure 3. Theme ”Business” was found in topic 13, with keywords such as ”stop” and ”small”. ”Business” was the most important keyword followed by ”pandemic”. Theme ”Stay home” was found in topic 17, with keywords such as ”staysafe” and ”quarantine”. ”corona” was the most important keyword followed by ”virus”. Theme ”Education” was found in topic 26, with keywords such as ”learn” and ”class”. ”School” was the most important keyword followed by ”kid”. Theme ”Vaccine” was found in topic 29, with keywords such as ”patient” and ”result”. ”Test” was the most important keyword followed by ”vaccine”. Theme ”Working from home” was found in topic 33, with keywords such as ”office” and ”space”. ”work” was the most important keyword followed by ”home”. Theme ”Closing and reopening of restaurants” was found in topic 41, with keywords such as ”restriction” and ”reopen”. ”Close” was the most important keyword followed by ”open”. Theme ”Social distancing” was found in topic 43, with keywords such as ”practice” and ”distance”. ”Social” was the most important keyword followed by ”distancing”.

A remarkable finding was that expected ”lockdown” topic of the model evaluation criteria, was not captured in a clear and interpretable single topic. Figure 9 in the appendix shows the 3 topics including ”lockdown”, where the ”lockdown” keyword shows to have a relatively inferior importance rank. For the topic with the highest ”lockdown” keyword rank, the time series data did not show sudden fluctuations around specific dates (e.g. a lockdown announcement date), while it was expected to identify these (Appendix Fig. 10).

5.2. Time series

The figures used to explore the topics and sentiment over time will follow the same structure for all topics: The Y-axis of the upper subplot (1) contains the proportion of tweets associated



Topic	Theme	Keywords
Topic 13	Business	business, pandemic, support, stop, small, hit, traffic, company, local, effect
Topic 17	Stay home/ stay safe	corona, virus, stayhome, coronaviru, staysafe, quarantine, isolation, funny, stayhealthy, lol
Topic 26	Education	school, kid, learn, class, student, child, virtual, proud, high, parent
Topic 29	Vaccine and testing	test, vaccine, patient, result, doctor, positive, hospital, negative, nurse, receive
Topic 33	Working from home	work, home, office, workfromhome, workingfromhome, space, wfh, hard, desk, coworker
Topic 41	Closing and reopening of restaurants	close, open, continue, reopen, begin, restaurant, door, place, restriction, announce
Topic 43	Social distancing	social, distancing, practice, distance, drink, outdoor, dining, table, maintain, dunwoody

Figure 3: Summary of 7 selected topics with their top 10 Keywords

with the topic, whereas the Y-axis of the middle subplot (2) contains the sentiment of that topic. The Y-axis of the lower subplot (3) contains the standard deviation of the sentiment scores. On the X-axis of these three subplots, time is plotted using a rolling mean of 7.

In Figure 4, the ”Stay Home / Stay safe” (upper figure) theme and the ”Vaccine and testing” (lower figure) theme are shown. Regarding the ”Stay home / Stay safe” theme, on the starting date (20th of March) of our dataset, the proportion of tweets associated with this topic is already above .5, and during April, the proportion of tweets associated with this theme steeply declines to under .25 until July. This pattern raised the suspicion that the event that caused the high proportion of tweets to be associated with this theme preceded the start date of the current dataset.

In the ”Vaccine and testing” theme time series, an increase from under .1 to above .2 in the proportion of tweets associated with this topic was detected in December 2020, which sustained until the end date of the dataset. This pattern indicated a possibility of an event causing this sudden increase. Given the recent developments surrounding COVID-19 vaccines, it was expected that an event in December could have been the cause of this increase in proportion of tweets associated with the theme. Therefore an event-sentiment analysis will be conducted for this theme.

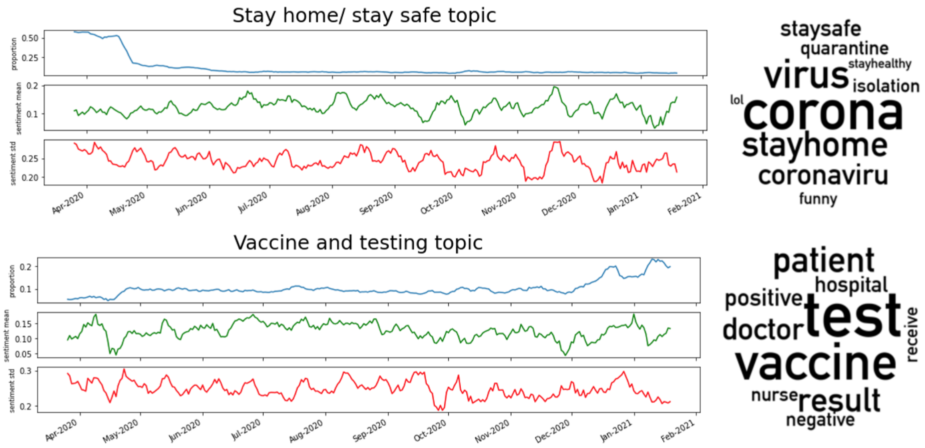


Figure 4: Time series for ”Stay home / Stay safe” and ”Vaccine and testing” topics containing topic frequency and sentiment accompanied with their respective Wordclouds

In Figure 5, the ”Restaurants closing and reopening”, ”Social distancing” and the ”Business” themes are shown. For these themes, relatively similar patterns of fluctuations were found. From around the start date of our dataset, for these topics an increase in the proportion of tweets associated with them was detected. These proportions shift downwards again until the end date of our dataset. Upon inspecting the COVID-19 event timeline by AJMC [15] and NBC[16], no clear suspicion arose for any particular event to be linked to these fluctuations. However, the Social distancing theme shows the largest increase in proportion of tweets associated with the theme in the smallest time-frame. Therefore, further insights into possibly related events to this theme shall be further inspected with an event-sentiment analysis.

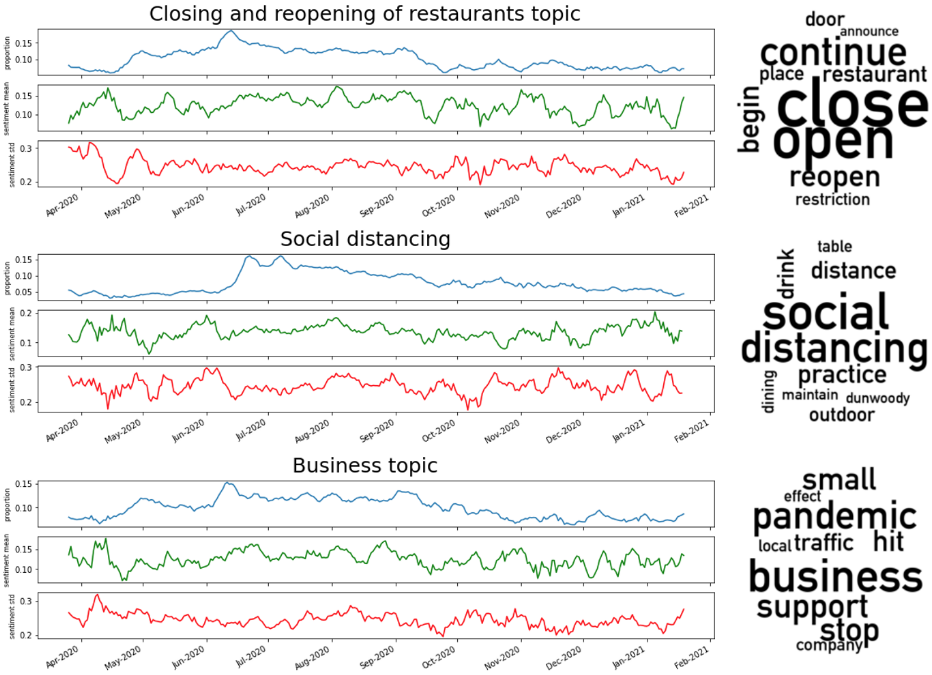


Figure 5: Time series for ”Restaurants closing and reopening”, ”Social distancing” and ”Business” topics containing topic frequency and sentiment accompanied with their respective Wordclouds

Finally, in Figure 6, the ”Education” theme (upper figure) and the ”Working from home” theme (lower figure) are shown. Both of these themes show more gradual changes in the proportion of tweets associated with them, and inspecting the timeline by AJMC [15] did not result in possible fluctuation-related candidate events.

5.3. Event analysis

Once a possible event-related fluctuation has been identified in the time series data, a closer and more in depth analysis was carried out on the time frame surrounding the fluctuation.

In Figure 7, two examples of event analysis results are shown. Here, the sentiment of tweets surrounding the event-related fluctuation was plotted, the red line indicating negative tweets and the green line indicating positive tweets. The number of tweets are shown on the Y-axis, while time is shown on the X-axis. Vertical lines represent events indicated by the legend. Below

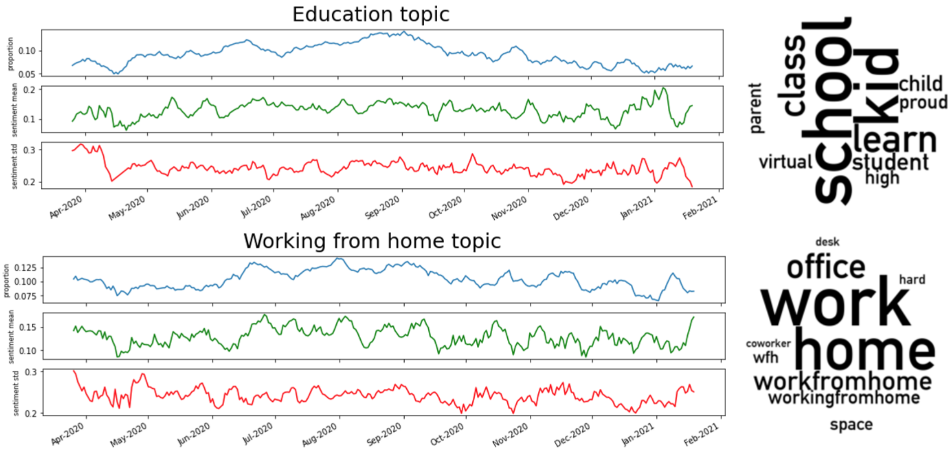


Figure 6: Time series for "Education" and "Working from home" topics containing topic frequency and sentiment accompanied with their respective Wordclouds

these time series, wordclouds of quartile most positive tweets are shown on the left (P), and wordclouds of the quartile most negative tweets are shown on the right hand side (N).

In Figure 7 A, the event analysis result surrounding the vaccine fluctuations from December 2020 onward are shown. Over the whole time interval, more tweets that are positively associated with the topic have been posted than tweets that are negatively associated with the topic. From December the 15th until December the 22nd, a spike in the amount of positive tweets associated with the topic can be seen. The events of "FDA Approving Pfizer-BioNTech vaccines" and "Pfizer vaccines administrated" precede this spike in positive tweets. We therefore assume these events are related to the observed patterns. The negative sentiment tweets fluctuate during this time period, but indicate no clear pattern. The P wordcloud shows "Ready" "AMP", "thing" and "healthcareworker" are important words for positive tweets. The N wordcloud shows "yesterday", "conspiracy", "antibody" and "anxiety" are import words for negative tweets.

Figure 7 B shows the "Social distancing" theme event analysis results. Again, over the whole time interval, more tweets that are positively associated with the theme have been posted. From June 10th until June 30th, a large increase in positively associated tweets can be seen for the topic. The event of "Washington DC reopening and social distancing rules" precedes the peak of the first spike, and is therefore assumed to be related to the observed pattern. Another such spike can be observed from the first of July until the 6th of July. Two smaller spikes occur between the 8th of July until the 12th, and between the 15th and the 21st. The Independence day event, celebrated on July the 4th, appears to be the center of the peak of positively associated tweets. However, in the P wordcloud, no mentions of the independence day are made. The P wordcloud shows that "Birthday", "kid", "sure" and "public" are important words for positive tweets. The N wordcloud shows that "place", "safety", "downtown" and "sun" are important words for negative tweets.

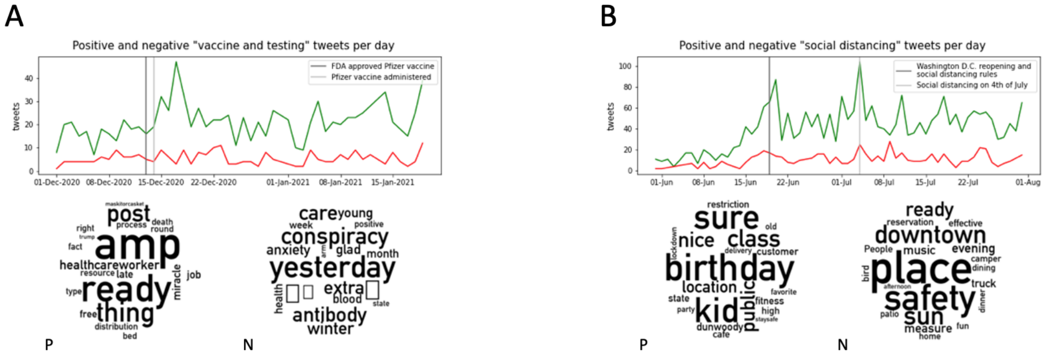


Figure 7: Sentiment analysis for the ”Vaccine” and ”Social distancing” topics accompanied with the most positive and most negative wordclouds

6. Discussion

After evaluation of the best LDA model output, the topic modeling results showed that most of the expected themes were identified in clearly interpretable topics (Fig. 3). We found that using more topics (50) resulted in the best interpretable topics (Appendix Fig. 8). Using a smaller number of topics, showed more gradual fluctuations in the time series data (data not shown). On the other hand, using more (and therefore more fine grained) topics showed larger and more sudden fluctuations that are more likely to have a causal relation ship with certain events. This finding emphasizes the importance of proper topic modeling evaluation itself, combined with the forthcoming results (in this case the time series analysis). The evaluation of the topic model itself could be further improved by looking at the convergence scores of the LDA models and the corresponding topic coherence scores[3, 4, 5].

Regarding the found topics, about all selection criteria topics were identified, as is shown in Figure 3. A remarkable finding was that the lockdown topic, was not clearly defined in a single topic (Appendix Fig. 9) and the lockdown time series data exhibited slowly changing fluctuations (Appendix Fig. 10), which makes defining causal relations with lockdown-related events hard. We presume that the inferior quality of the lockdown topic, is partially due to the nation-level structure of our data, while individual states have their own lockdown announcements on different time points. Our results showed we are able to identify clear nation-level events, such as Christmas, the black lives matter protests in June and the elections in November (Appendix Fig. 11). Therefore, follow up research could add the state-level layer in order to identify more state-level events.

However, while this state-level layer would be a enticing addition for improved event-detection, in Figure 7 it was shown we identified a possible state-level related event (Washington D.C. reopening and social distancing rules announcement). Furthermore, the lockdown-related topics were in general hard to interpret (Appendix Fig. 9). Therefore, the absent state-level layer in our data cannot be the only reason of the inferior quality of the lockdown topic. We presume that the expected lockdown topic is too big to be clearly identified in a set of 50 topics, and the LDA model may actually identify sub-topics of the overarching ”lockdown” topic.

In order to investigate these possible overarching topics, it would be interesting to inspect the concurrence of the 50 topics. This may identify certain groups of topics that are often mentioned together, which may yield additional interesting insights. This may for example identify the over arching lockdown topic, or identify closely related topics, such as the expected co-occurrence of the ”closing and reopening” topic and the ”business” topic (Fig. 5).

The time series analysis revealed that some extracted themes show clear patterns of discussion frequency and sentiment. It was found that patterns of discussion frequency and sentiment concur with the discourse of the pandemic. For example, staying at home during the COVID-19 pandemic is a new reality many find difficulty to cope with [3]. However as time goes on, such realities become the norm and this is reflected in the decreasing amount of tweets associated with this topic, while the sentiment keeps fluctuating. With the optimal LDA model, a vaccine theme was identified, and from Figure 4 it became evident that over the course of 2020 the vaccine theme was not frequently discussed. However, news of recent developments on vaccinations has spread around the globe, and the fluctuations in the theme discussions concur with these developments. The discourse of the impact of COVID-19 on different institutions of society such as the business sector and the educational sector is reflected in the themes in Figures 5, 6. From March 2020 onward, millions of Americans are hit by COVID-19 and its consequences, as people lose their jobs and many businesses, restaurants and shops close. On top of that, educational institutions have to introduce major changes to the current educational system, and for many, home is now the new work-space. Concurrently, the discussion frequency of themes ”Working from home”, ”Closing and reopening of restaurants”, ”Education”, ”Business” and ”Social distancing” increases from March 2020 until October 2020.

The sentiment means and standard deviations in these figures were however less useful than expected, and should be interpreted carefully. The number of tweets per day are unequally distributed, which leads to unreliable values for days with few topic related tweets, which makes identification of true fluctuations less intuitive in comparison with the topic proportion fluctuations. While some fluctuations in the sentiment time series data may have interesting causal relations with certain events, our current approach, with manual related event lookup, does not provide an easy fluctuation-event relation identification pipeline. Future research could improve on this by combining the current time series data, with an additional timeline dataset with COVID-19 related events/ news articles. This may enhance the fluctuation-event relation identification pipeline for the topic proportions as well as its sentiment data.

In order to link events to fluctuations in topic frequencies and their sentiment, information about event dates and tweet dates were combined with topic frequencies and sentiment scores. In Fig. 7 A, we showed that two important events related to the vaccine preceded an increase in tweets associated with this theme. For the positive tweets ”AMP” is the most important keyword, which indicates that positive tweets concerning the vaccine theme mention Americans For Medical Progress, a charity organization that aims to protect and advocate for society’s investment in medical research. A general finding is that most tweets in our dataset have a positive association with the vaccine theme during the selected time period, however some tweets have a negative disposition towards the vaccine theme, where mentions of anxiety and conspiracy are made. The finding of the words ”conspiracy” and ”anxiety” in the negative tweets associated with the vaccine theme could indicate a requirement for further elaboration on the mechanisms of vaccines and elucidate certain aspects citizens could be anxious about.

This finding exemplifies a practical implication of this study, and future research could obtain similar insights by improving on the methods used in this study. In Fig. 7 B, two events that were possibly related were compared to positive and negative tweets concerning the ”social distancing” theme. The wordclouds for these tweets however revealed no words indicative of the event relating to the theme and sentiment fluctuations. One reason for this could be that the events actually did not relate to these fluctuations. This finding illustrates the shortcomings of the used methods, and that improvements in future research can be made by improving on event analysis. For example, by setting thresholds for selecting coherent themes, or for example improving on methods for selecting events.

This type of exploratory research could be considered the first step in two possible directions. First, a qualitative approach could be taken in order to inspect individual tweets more in depth. With this approach practical implications are limited in their generalizability, however more detailed insights could be obtained. Secondly, a quantitative approach could be further elaborated by scraping more tweets, using more keywords to scrape these tweets, for example keywords occurring in the wordclouds in Fig. 7, and refining selection criteria for the LDA models.

There are several limitations in the current study. First, the findings of this study are limited to the English speaking Twitter population in the United States, and the sample of tweets for every day was limited in size and unequally distributed. Second, the the time interval used in the current study cannot capture the full discourse of the pandemic, as it is still ongoing, and results could be biased towards time periods with more tweets. Furthermore, the validity of the extracted themes was not assessed, and no coherence indicators were consulted. Lastly, the assumption that the public perception of the pandemic relates to the tweets included in the current dataset has not been tested. Future research could possibly expand upon the current findings by improving on the aforementioned limitations.

7. Conclusion

In this study, topic modeling was performed on 127,12 COVID-19 related tweets from the United States. Elaborate evaluation of the identified topics yielded clear and interpretable topics of themes related to the ongoing COVID-19 pandemic. Topic modeling and sentiment analysis were combined with event analysis in order to identify event-related fluctuations in the discussed topics and their respective sentiment data. This study shows that the proposed methodology can be used to identify these event-related fluctuations. Furthermore, it was shown that fluctuations in the frequency of topic discussion and their sentiment may provide insights in the public perception of previously made governmental decisions and announcements. These insights in turn may support future decision-making of similar situations. While our approach shows promising results, shortcomings of the current methodology and dataset are also showcased. However, with the proposed improvements, these research methods may provide a first step in the identification of interesting event-related fluctuations in topics and their sentiment.

8. Acknowledgments

Acknowledgments

Thanks to joris for scraping tips Joris_Github_url and to twitter giving us free dev accountsLATEX templates https://twitter.com/home.

References

[1] WHO, Coronavirus disease (covid-19) dashboard (2021). URL: https://covid19.who.int/.

[2] I. M. Fund, Covid-19 has triggered a steep increase in debt, particularly in emerging market and developing economies (2020). URL: https://www.worldbank.org/en/who-we-are/news/coronavirus-covid19.

[3] J. Xue, J. Chen, C. Chen, C. Zheng, S. Li, T. Zhu, Public discourse and sentiment during the covid 19 pandemic: Using latent dirichlet allocation for topic modeling on twitter, PLOS ONE 15 (2020) 1–12. URL: https://doi.org/10.1371/journal.pone.0239441. doi:10.1371/journal.pone.0239441.

[4] V. A. Kharde, S. Sonawane, Sentiment analysis of twitter data : A survey of techniques, CoRR abs/1601.06971 (2016). URL: http://arxiv.org/abs/1601.06971. arXiv:1601.06971.

[5] S. Boon-Itt, Y. Skunkan, Public perception of the covid-19 pandemic on twitter: Sentiment analysis and topic modeling study, JMIR Public Health Surveill 6 (2020) e21978. URL: http://publichealth.jmir.org/2020/4/e21978/. doi:10.2196/21978.

[6] E. Chen, K. Lerman, E. Ferrara, Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set, JMIR Public Health Surveill 6 (2020) e19273. URL: http://publichealth.jmir.org/2020/2/e19273/. doi:10.2196/19273.

[7] S. Kaur, P. Kaul, P. M. Zadeh, Monitoring the dynamics of emotions during covid-19 using twitter data, Procedia Computer Science 177 (2020) 423–430. URL: https://www.sciencedirect.com/science/article/pii/S1877050920323243. doi:https://doi.org/10.1016/j.procs.2020.10.056, the 11th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2020) / The 10th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH 2020) / Affiliated Workshops.

[8] R. Lamsal, Coronavirus (covid-19) geo-tagged tweets dataset (2020). URL: https://dx.doi.org/10.21227/fpsb-jz61. doi:10.21227/fpsb-jz61.

[9] R. Lamsal, Design and analysis of a large-scale covid-19 tweets dataset (2020). URL: https://doi.org/10.1007/s10489-020-02029-z.

[10] J. Roesslein, Tweepy: Twitter for python! (2020). URL: https://github.com/tweepy/tweepy.

[11] Z. K. Stine, J. E. Deitrick, N. Agarwal, Comparative religion, topic models, and conceptualization: Towards the characterization of structural relationships between online religious discourses, in: CHR, 2020.

[12] M. Honnibal, I. Montani, S. Van Landeghem, A. Boyd, spaCy: Industrial-strength Natural

Language Processing in Python (2020). URL: https://doi.org/10.5281/zenodo.1212303. doi:10.5281/zenodo.1212303.

[13] R. Rehurek, P. Sojka, Gensim–python framework for vector space modelling, NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic 3 (2011).

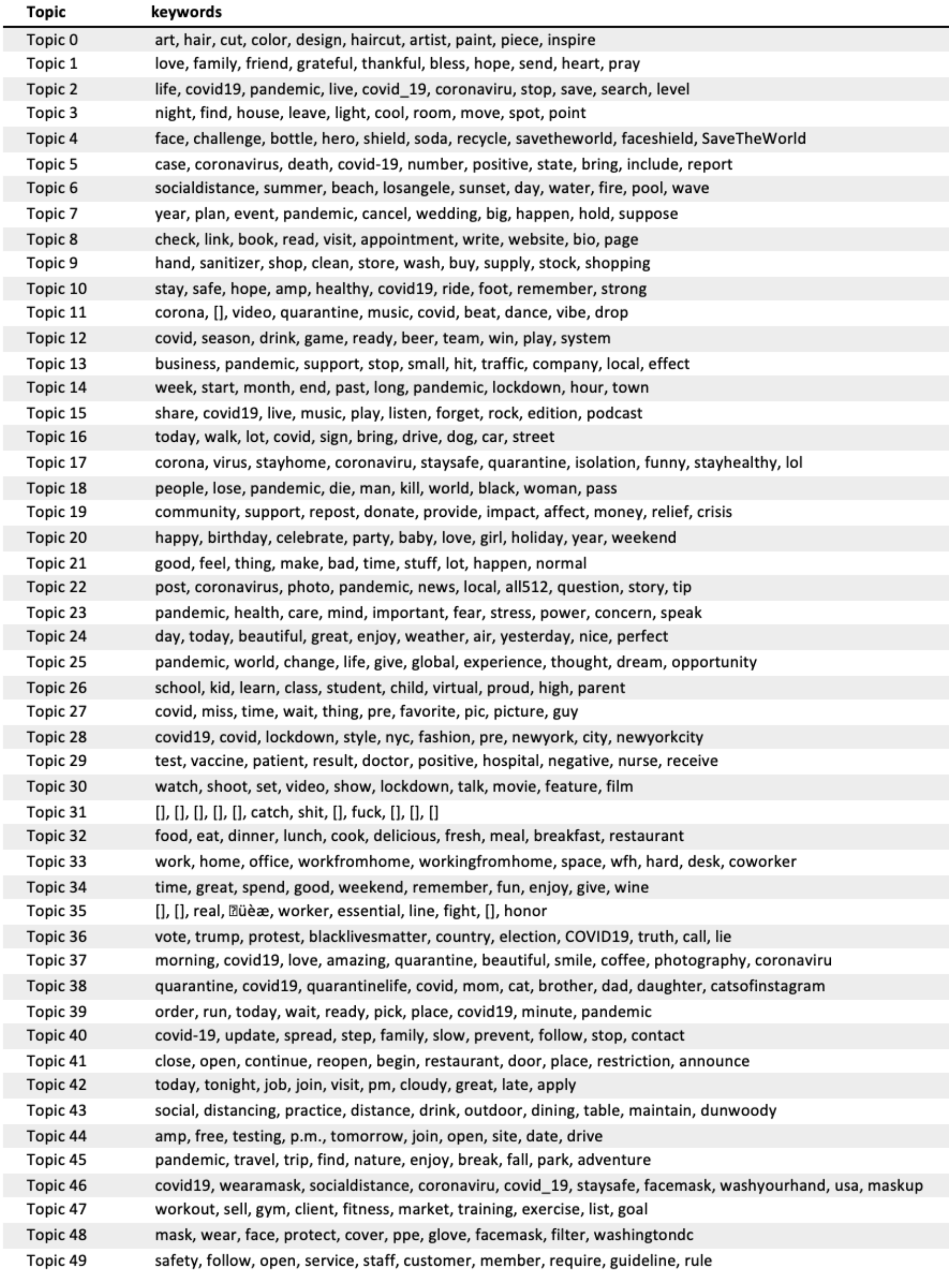
[14] R. Luh, S. Schrittwieser, S. Marschalek, Llr-based sentiment analysis for kernel event sequences (2017). doi:10.1109/AINA.2017.47.

[15] AJMC, Covid-19 timeline (2020). URL: https://www.ajmc.com/view/a-timeline-of-covid19-developments-in-2020.

[16] NBCnews, Covid-19 timeline (2020). URL: https://www.nbcnews.com/health/health-news/coronavirus-timeline-tracking-critical-moments-covid-19-n1154341.

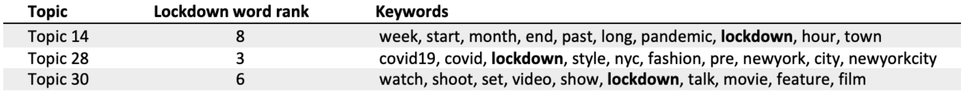
9. Appendices

A. Additional figures



Topic	keywords
Topic 0	art, hair, cut, color, design, haircut, artist, paint, piece, inspire
Topic 1	love, family, friend, grateful, thankful, bless, hope, send, heart, pray
Topic 2	life, covid19, pandemic, live, covid_19, coronaviru, stop, save, search, level
Topic 3	night, find, house, leave, light, cool, room, move, spot, point
Topic 4	face, challenge, bottle, hero, shield, soda, recycle, savetheworld, faceshield, SaveTheWorld
Topic 5	case, coronavirus, death, covid-19, number, positive, state, bring, include, report
Topic 6	socialdistance, summer, beach, losangele, sunset, day, water, fire, pool, wave
Topic 7	year, plan, event, pandemic, cancel, wedding, big, happen, hold, suppose
Topic 8	check, link, book, read, visit, appointment, write, website, bio, page
Topic 9	hand, sanitizer, shop, clean, store, wash, buy, supply, stock, shopping
Topic 10	stay, safe, hope, amp, healthy, covid19, ride, foot, remember, strong
Topic 11	corona, [], video, quarantine, music, covid, beat, dance, vibe, drop
Topic 12	covid, season, drink, game, ready, beer, team, win, play, system
Topic 13	business, pandemic, support, stop, small, hit, traffic, company, local, effect
Topic 14	week, start, month, end, past, long, pandemic, lockdown, hour, town
Topic 15	share, covid19, live, music, play, listen, forget, rock, edition, podcast
Topic 16	today, walk, lot, covid, sign, bring, drive, dog, car, street
Topic 17	corona, virus, stayhome, coronaviru, staysafe, quarantine, isolation, funny, stayhealthy, lol
Topic 18	people, lose, pandemic, die, man, kill, world, black, woman, pass
Topic 19	community, support, repost, donate, provide, impact, affect, money, relief, crisis
Topic 20	happy, birthday, celebrate, party, baby, love, girl, holiday, year, weekend
Topic 21	good, feel, thing, make, bad, time, stuff, lot, happen, normal
Topic 22	post, coronavirus, photo, pandemic, news, local, all512, question, story, tip
Topic 23	pandemic, health, care, mind, important, fear, stress, power, concern, speak
Topic 24	day, today, beautiful, great, enjoy, weather, air, yesterday, nice, perfect
Topic 25	pandemic, world, change, life, give, global, experience, thought, dream, opportunity
Topic 26	school, kid, learn, class, student, child, virtual, proud, high, parent
Topic 27	covid, miss, time, wait, thing, pre, favorite, pic, picture, guy
Topic 28	covid19, covid, lockdown, style, nyc, fashion, pre, newyork, city, newyorkcity
Topic 29	test, vaccine, patient, result, doctor, positive, hospital, negative, nurse, receive
Topic 30	watch, shoot, set, video, show, lockdown, talk, movie, feature, film
Topic 31	[], [], [], [], [], catch, shit, [], fuck, [], [], []
Topic 32	food, eat, dinner, lunch, cook, delicious, fresh, meal, breakfast, restaurant
Topic 33	work, home, office, workfromhome, workingfromhome, space, wfh, hard, desk, coworker
Topic 34	time, great, spend, good, weekend, remember, fun, enjoy, give, wine
Topic 35	[], [], real, ⍰üèæ, worker, essential, line, fight, [], honor
Topic 36	vote, trump, protest, blacklivesmatter, country, election, COVID19, truth, call, lie
Topic 37	morning, covid19, love, amazing, quarantine, beautiful, smile, coffee, photography, coronaviru
Topic 38	quarantine, covid19, quarantinelife, covid, mom, cat, brother, dad, daughter, catsofinstagram
Topic 39	order, run, today, wait, ready, pick, place, covid19, minute, pandemic
Topic 40	covid-19, update, spread, step, family, slow, prevent, follow, stop, contact
Topic 41	close, open, continue, reopen, begin, restaurant, door, place, restriction, announce
Topic 42	today, tonight, job, join, visit, pm, cloudy, great, late, apply
Topic 43	social, distancing, practice, distance, drink, outdoor, dining, table, maintain, dunwoody
Topic 44	amp, free, testing, p.m., tomorrow, join, open, site, date, drive
Topic 45	pandemic, travel, trip, find, nature, enjoy, break, fall, park, adventure
Topic 46	covid19, wearamask, socialdistance, coronaviru, covid_19, staysafe, facemask, washyourhand, usa, maskup
Topic 47	workout, sell, gym, client, fitness, market, training, exercise, list, goal
Topic 48	mask, wear, face, protect, cover, ppe, glove, facemask, filter, washingtondc
Topic 49	safety, follow, open, service, staff, customer, member, require, guideline, rule

Figure 8: All topics created by the LDA model accompanied with their top 10 keywords



Topic	Lockdown word rank	Keywords
Topic 14	8	week, start, month, end, past, long, pandemic, **lockdown**, hour, town
Topic 28	3	covid19, covid, **lockdown**, style, nyc, fashion, pre, newyork, city, newyorkcity
Topic 30	6	watch, shoot, set, video, show, **lockdown**, talk, movie, feature, film

Figure 9: Topics containing ”lockdown” keyword and their rankings

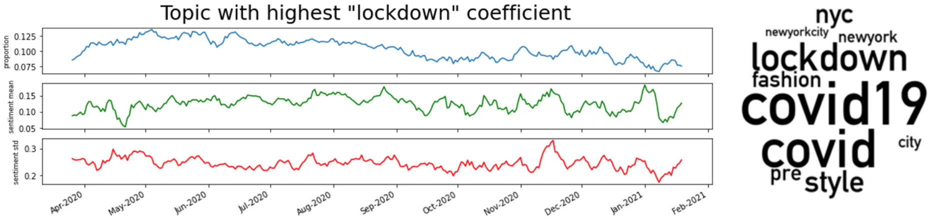


Figure 10: Time series for ”Lockdown” topic containing topic frequency and sentiment accompanied with its Wordcloud

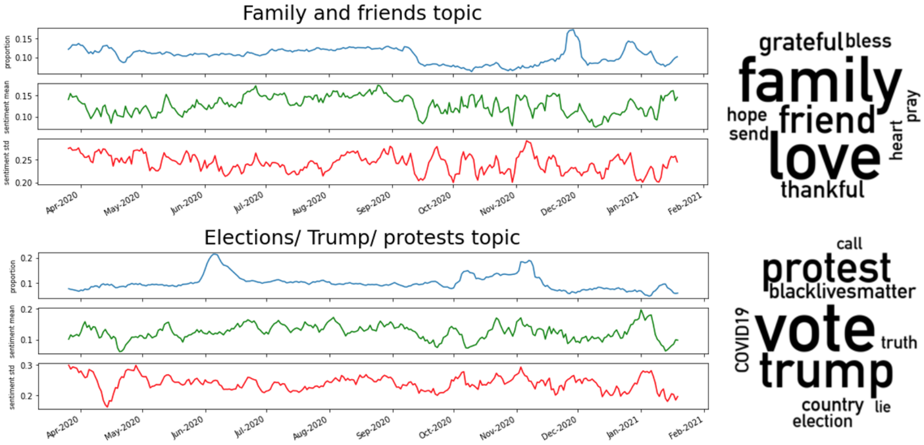


Figure 11: Time series for ”Family and friends” and ”Elections/Trump/protests” topics containing topic frequency and sentiment accompanied with their respective Wordclouds