TECHNICAL UNIVERSITY OF CRETE

SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

An Experimental Analysis of Twitter Suspension during the First COVID19 Period

Diploma Thesis
of

Georgios Nektarios Nikou

Committee:
Assoc. Prof. Sotirios Ioannidis (Supervisor)
Prof. Michail G. Lagoudakis
Prof. Michalis Zervakis

Chania, October 2023

ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Μια Πειραματική Ανάλυση των Αναστολών στο Twitter κατά την Πρώτη Περίοδο του COVID19

Διπλωματική Εργασία

του

Γεώργιου Νεκτάριου Νίκου

Εξεταστική Επιτροπή:
Αναπληρωτής Καθηγητής Σωτήριος Ιωαννίδης (Επιβλέπων)
Καθηγητής Μιχαήλ Γ. Λαγουδάκης
Καθηγητής Μιχάλης Ζερβάκης

Χανιά, Οκτώβριος 2023

Abstract

This study aims to capture the overall sentiment of people's tweets regarding COVID-related subjects and to examine any attempts to spread fake news and misinformation on Twitter. Our research is based on a dataset collected through the Twitter API, containing approximately 200 million tweets from two popular COVID-related hashtags. We conduct sentiment analysis using the XLM-RoBERTa-large model on several topics related to the COVID-19 pandemic. Next, we perform data analysis to identify interesting patterns and characteristics of this vast dataset. Our research also targets suspended Twitter accounts and by using the Latent Dirichlet Allocation algorithm we identify their topics of discussion. We construct the retweet social graph to analyze their social network connections, enabling us to detect any coordinated actions to retweet the same content in large quantities. The results showed a trend in sentiment towards terms like COVID-19, conspiracy, and lockdown. We observe that although suspended users made up only 0.74% of the total users in the dataset, they generated 7.52% of the total posts in the dataset.

Περίληψη

Η παρούσα μελέτη έχει ως στόχο την καταγραφή των συναισθημάτων των ανθρώπων στα tweets τους σχετικά με το COVID-19 και στην εξέταση τυχόν προσπαθειών διάδοσης ψευδών ειδήσεων και παραπληροφόρησης στο Twitter. Η έρευνά μας βασίζεται σε ένα σύνολο δεδομένων που συλλέχθηκε μέσω του Twitter API, το οποίο περιέχει περίπου 200 εκατομμύρια tweets από δύο δημοφιλή hashtags που σχετίζονται με το COVID-19. Εφαρμόζουμε ανάλυση συναισθήματος χρησιμοποιώντας το μοντέλο XLM-RoBERTa-large σε διάφορα θέματα που σχετίζονται με την πανδημία COVID-19. Στη συνέχεια, πραγματοποιούμε ανάλυση δεδομένων για να εντοπίσουμε ενδιαφέροντα μοτίβα και χαρακτηριστικά σε αυτό το μεγάλο dataset. Η έρευνά μας επίσης επικεντρώθηκε σε λογαριασμούς Twitter που έχουν ανασταλεί και χρησιμοποιούμε τον αλγόριθμο Latent Dirichlet Allocation για να εντοπίσουμε τα θεμάτα συζήτησης τους. Επιπλέον κατασκευάζουμε τον γράφο αναδημοσιεύσεων για να αναλύσουμε τις συνδέσεις τους στο κοινωνικό δίκτυο, επιτρέποντάς μας να εντοπίσουμε τυχόν συντονισμένες ενέργειες για την αναδημοσίευση του ίδιου περιεχομένου σε μεγάλες ποσότητες. Τα αποτελέσματα έδειξαν μια τάση στα συναισθήματα προς όρους όπως COVID-19, συνωμοσία και lockdown. Παρατηρούμε ότι παρόλο που οι ανεσταλμένοι χρήστες αποτελούσαν μόνο το 0.74% των συνολικών χρηστών, δημιούργησαν το 7.52% των συνολικών αναρτήσεων.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον καθηγητή κ. Ιωαννίδη για την ανάθεση αυτής της διπλωματικής εργασίας και την βοήθεια που μου προσέφερε. Θα ήθελα επίσης να ευχαριστήσω τον Αλέξανδρο Σέβτσοφ και την Δέσποινα Αντωνακάκη για την πολύτιμη βοήθεια τους κατά την διάρκεια της εργασίας μου. Επιπλέον, θα ήθελα να ευχαριστήσω τους καθηγητές κ. Ζερβάκη και κ. Λαγουδάκη που αμέσως δέχτηκαν να συμμετέχουν ως μέλη της εξεταστικής επιτροπής.
Ευχαριστώ πολύ την κοπέλα μου Ανδριάνα, για την αγάπη της και την στήριξη της όλα αυτά τα χρόνια. Ευχαριστώ πολύ τους γονείς μου Θανάση και Χριστίνα, και τις αδερφές μου Κατερίνα και Σταματία, για όλη την στήριξη τους και που είναι πάντα εκεί μαζί μου ό,τι κι αν χρειαστώ. Τέλος, θέλω να ευχαριστήσω όλους τους φίλους μου και τις παρέες μου για όλες τις ωραίες στιγμές που περάσαμε όλα αυτά τα φοιτητικά χρόνια.

Contents

List of Figures

1 Introduction

The COVID-19 outbreak, a worldwide pandemic caused by the coronavirus disease, was first reported in Wuhan, China in December 2019. On 11 March 2020, the World Health Organization (WHO) officially declared COVID-19 as a pandemic [1]. Since then, the world has witnessed the devastating impact of the virus, with approximately 6.9 million lives lost to the disease. Governments worldwide have gradually responded with strict measures, including lockdowns, social distancing, quarantines, travel restrictions, and teleworking, affecting public life and social gatherings. In addition, the extensive disruptions caused by COVID-19 had a wide-reaching impact on the global economy, leading to a significant market crash [2].

Social media platforms serve, among other purposes, as a means for users to participate in ongoing discussions and debates within the public sphere regarding emerging topics. Twitter is considered one of the most popular online social networks currently. Twitter operates in a micro-blogging way, where users create concise posts, called tweets, with a limit in characters used. Twitter's nature as a micro-blogging app, with character limitation, makes it perfect for the quick transmission of information on breaking news and real-time events. Users are able to share their tweets publicly, add hashtags to participate in ongoing discussions, attract attention, and gain more engagement for their posts.

Online discussions about COVID-19 have surged on social media platforms throughout the pandemic's progression. These discussions covered various topics, from the scientific aspects of the virus to societal responses and the broader socio-political implications. In addition, these dialogues sparked heated controversies regarding vaccination strategies and the effectiveness of government policies to handle this unprecedented situation. Also, it was noticed throughout the social media the proliferation of misinformation and fake news related to the novel disease.

At present, there is no study on Twitter posts content in an extensive timeline regarding COVID-19 disease. Our question is how the public perceived the outbreak of this novel disease and how they reacted to the applied measures by governments. Also, we would like to explore whether there were any malicious attempts by users to propagate fake news and propaganda regarding the disease and government policies.

In this study, we aim to capture the overall sentiment expressed by people in their tweets concerning various COVID-related subjects. As a next step, we perform data analysis to gain insights into the characteristics of this massive dataset and uncover interesting data patterns. Lastly, we investigate the topics discussed in the posts of Twitter users who were suspended. We also examine their connections within the social network and analyze their interactions with each other.

To achieve this, the approach taken in this study is the following:

- We perform sentiment analysis on tweets by using machine learning to investigate various aspects of the COVID-19 pandemic's discussion topics. With the help of the XLM-RoBERTa model, we analyze the content of tweets to reveal the prevailing sentiments regarding critical topics such as lockdowns, mask mandates, and vaccination efforts. The zero-shot classification used for labels is positive or negative to determine whether the public discourse on these topics tends to be positive or negative.

- We analyze the temporal distribution of tweets and the number of unique users who posted them, comparing all users and suspended users. Additionally, we have identified the most commonly used languages and the most popular hashtags in tweets over the examined months.
- Finally, we examine the content generated by suspended users with the Latent Dirichlet Allocation (LDA) algorithm to identify the prevalent topics in their tweets. Also, we construct the retweet social graph to uncover any orchestrated behaviors of these accounts in clusters of massive retweets.

2 Theoretical Background and Related Work

2.1 Sentiment Analysis

Natural Language Processing (NLP) is a subdomain of Artificial Intelligence that focuses on the development of techniques and algorithms that enable computers to understand, interpret, and analyze physical language. Transformer-based models, often also called Large Language Models (LLM), have changed the landscape of NLP. Examples of such models include BERT, RoBERTa, and GPT models. These LLMs are based on the same principle of training bi-directional transformer models on huge unlabelled text corpora in a fully unsupervised manner. This process generates a general language model that can be fine-tuned to excel in specific language processing tasks, such as classification, question-answering models, and chatbots.

Sentiment analysis or opinion mining is a sub-field of the NLP. It is defined as the field of study of people's opinions, sentiments, appraisals, attitudes, and emotions toward entities and their aspects expressed in text. These entities, also called sentiment targets, can vary from events, news, products, services, other human beings, and virtually any subject that can be a topic of opinion or discussion.

Some applications of sentiment analysis include its use in the commerce domain, where companies can utilize the customers' reviews and feedback to improve and redesign their products or services. Another application includes scanning through social media posts to get a sense of emerging events and news or the opinion of a specific product or service. Also, it can be utilised by governments or policy-makers for political opinion mining to grasp the public discourse on emerging events or new legislation that is being proposed or applied.

The development of the sentiment analysis field is intertwined with the growth of the Web, especially social media platforms. Social media platforms make a perfect place for sentiment analysis to be applied because of the vast number of texts and content posted by their users. This extremely valuable amount of recorded digital data of people's opinions on several subjects is ready to be analyzed, extract conclusions, and take actions based on that data.

The challenge of sentiment analysis is that the written text is highly subjective. Expressing opinions verbally provides additional context through the speaker's tone and non-verbal body language. However, written text represents something absolute, making it challenging to process and analyze. Written text can be interpreted differently by two humans, leading to entirely distinct conclusions. The semantics of a text can vary extremely if irony or sarcasm is involved, where its meaning can differ completely from the literal interpretation.

In previous years, the sentiment analysis field has commonly relied on lexicon-based techniques, leveraging a lexicon, which is a list of words and their associated sentiment scores, to determine the sentiment of the text. This sentiment analysis approach involves aggregating the sentiment scores of individual words within the text to determine its overall sentiment. Nowadays the leading techniques include the use of deep learning algorithms. The rise of Transformer-based models, such as BERT and RoBERTa, has dominated the NLP field. LLMs have a deep understanding of the language context. They can easily capture nuances, sarcasm, and sentiment that traditional methods might miss. Also, they can handle large datasets efficiently, making them suitable for sentiment analysis based on extensive social media datasets.

Sentiment analysis is a versatile field with various techniques for understanding and classifying sentiments in textual data. One of the fundamental approaches is binary sentiment analysis, where text is categorized as “positive” or “negative,” and sometimes, an intermediate category of “neutral” is added to capture sentiments that do not strongly lean in either direction. This method is particularly useful for tasks like sentiment classification in product reviews or customer feedback. However, sentiment analysis goes beyond the binary classification. Emotion detection, for instance, takes sentiment analysis to a more nuanced level by classifying text into specific emotions like “anger,” “surprise,” or “joy.” This approach helps in understanding the emotional states expressed in the text, providing valuable insights for applications such as social media sentiment tracking and customer sentiment in chatbots.

2.2 Topic modeling

Topic modeling is a method for unsupervised classification of documents based on their content and identifies common themes or subjects called topics. The most common method for topic modeling is Latent Dirichlet Allocation (LDA). LDA is a statistical and probabilistic machine learning algorithm that is used to discover latent topics in a collection of documents. The algorithm works by iteratively sampling the topics and words for each document, using the probability distribution over the words for each topic and the probability distribution over the topics for each document. At the end of the process, the model produces a set of topics, each represented by a distribution over the words in the vocabulary, and a set of topic assignments for each document, indicating the strength of each topic in that document. This technique is commonly used in text mining and natural language processing applications, as well as in recommendation systems and information retrieval. Therefore, it is well-suited to capture the general discussion on Twitter regarding COVID-19 and the underlying topics of the tweets.

2.3 Social graphs

Social graphs are commonly used to depict relationships between users on social media platforms. These graphs provide a structured clear way to understand the complex relationships that form the digital social landscape. Typically, nodes in the graph represent users while edges capture the connections or interactions between them. These interactions go beyond just friendships or followers and can include likes, comments, mentions, and other types of engagement.

Examining social graphs can reveal hidden patterns and insights. One powerful aspect is the identification of influential users or nodes within the network. These users have a significant impact on the network, and learning about their behavior and influence can be very valuable. Additionally, social graphs can detect communities or groups of users who share common interests or topics. This information can be useful for various purposes, like understanding how information spreads through the network, identifying potential advertising targets, or finding users who may be more likely to engage with certain content.

2.4 Related Work

The related research on the social media analysis focuses on the spread of misinformation [3, 4, 5, 6, 7] and fake news [8, 9, 10, 11, 12, 13, 14, 15], content analysis [16, 17, 18, 19, 20, 21] and sentiment analysis [22, 23, 24, 25, 26, 27, 28], while there has been an effort to apply language-agnostic analysis [29, 30].

In [31] the author retrieves from Twitter API 43.3M English tweets about COVID-19 from the day COVID-19 was initially announced in the United States on January 21, 2020, through March 12, 2020. The authors take into account the tweets on the 10^{th} and 90^{th} percentile of the bot score distribution, to compare key account characteristics, as well as their tweets content and topics of discussion. The results conclude that bot users engage in discussions about public health by promoting certain political ideologies. Also it is assumed that bot accounts take advantage of trending topics and act around the same time, based on observation of bot users' time series.

As for Twitter datasets about COVID-19, in [32] they have collected a large dataset consisting of over 800M tweets in all languages from January 2020 through November 2020 and still updating nowadays, as of 09/11/22, according to their GitHub repository. It also includes a clean version with no retweets as well as top frequent terms, bi-grams, and tri-grams. Watching the monthly number of tweets we may assume that the activity peak was in the Spring of 2020 and since then there has been a declining activity. At [33] they have focused on collecting tweets in Arabic language from COVID-19-related Arabic keywords. It includes 3.9M tweets from January 1, 2020, through April 15, 2020. They also observed an increase in activity at the end of March due to the rise of the COVID-19 outbreak.

On NLP and Twitter a BERT-based model was developed in [34] that focused on COVID-19 tweets. The authors took the original BERT-LARGE model and trained it with 22.5M tweets collected from January 20 to April 16, 2020. The evaluation with a F1 score was conducted on 5 different training datasets. Having BERT as baseline the model presented better accuracy on all datasets, with the largest being on COVID/health-related datasets and even a small increase in general Twitter or non-twitter datasets.

The authors in [21] compared the volume of URLs leading to low-credibility sources to official ones such as the New York Times and CDC. They utilized two Twitter datasets from February 1 to April 27, 2020, as well as a bot detection model to binary classify users. They found that the overall sharing of low-credibility sources is similar to official ones. Bot-like accounts are more likely to post low-credibility links and retweet other bot-like accounts than human-like users. Lastly, the content of the titles in low-credibility links was analyzed, and came up with political and economic topics.

Regarding the spread of rumors on social media in [35] a framework is proposed to detect rumors at the tweet level. Using four real-life event Twitter datasets they executed three experiments to build their framework. The experiments were conducted to find the best feature extractor (standalone and hybrid), the best textual features, and the ML model for the classifier. The results suggest that BERT is the best standalone feature extractor, and in general context-based models perform better than the content-based ones. The Random Forest model yields the best evaluation results out of the ML models. Finally, the most appropriate textual features to detect rumors are 'URLs, Trust emotions, Verbs, Adjectives, and Propositions'. The framework displayed 80-97% accuracy on the datasets, overcoming the accuracy of the three baselines.

On the subject of content and sentiment analysis of COVID-related tweets in [36] they have utilized unsupervised machine learning, thematic qualitative analysis, and sentiment analysis to discover the response of the public to this novel disease. Their dataset includes over 4 million English tweets from March 7 through April 21, 2020. With the help of the Latent Dirichlet Allocation algorithm, they have found the most common bi-grams and uni-grams. Also with the same algorithm, they have defined 13 topics that can be broken into 5 different themes. Sentiment analysis with 8 emotions was performed on the 13 topics with anticipation being the most prevalent emotion across all topics and trust not being so typical emotion as in previous studies.

In a similar manner at [24] they explore the topics and sentiment of users' tweets posted at the start of the pandemic. Collecting 107K tweets from December 13, 2019, to March 9, 2020, they performed data analysis, sentiment analysis, and topic modelling on the data. Their findings show that the initial trends and symptoms that users reported can be divided into three stages. The sentiment of the people towards the disease is mostly negative with fear being the common emotion, although an increase of positive emotions is observed as more information is shared about the new disease. Finally, the authors discovered six topics discussed by users across three main themes.

The authors in [16] analyze the content of G7 leaders' tweets about coronavirus. Specifically, 203 popular tweets were examined with 82.8% being classified 'Informative', 9.4% were 'morale-boosting' and 6.9% were 'Political'.

In [30] language-agnostic BERT sentence embedding (LaBSE) model is proposed for supporting cross-lingual sentence embeddings. The authors combine the previous best methods for learning cross-lingual sentence embeddings with pre-trained encoders on large language models. The model achieves state-of-the-art performance on bi-text retrieval/mining tasks compared to previous models while also having good performance on mono-lingual transfer learning benchmarks. Also, it performs well in over 30 languages that there are no training data.

3 Dataset

The dataset is composed of tweets fetched by Twitter API from two popular COVID-19-related hashtags, #coronavirus and #COVID19. Approximately 208M tweets were retrieved from #coronavirus and 392M tweets for #COVID19, resulting in a total of 600M tweets. For our analysis, we use a portion of the dataset including 193,459,593 tweets from around 108 million unique users posted from February 19, 2020, to July 11, 2020, spanning 144 days. We conduct a comparison of the tweets appearing in both hashtags to detect and remove the duplicate tweets. There were in total 807,796 suspended users in the dataset, which accounts for 0.74% of all users, and they created 14,557,493 tweets, making up 7.52% of all tweets.

4 Implementation

4.1 Model

The model that is used to perform sentiment analysis is XLM-RoBERTa-large [37]. RoBERTa [38] is a transformer-based model that was built on the architecture of the prior NLP model BERT and was unsupervisedly trained on a large English corpus. XLM-RoBERTa-large is an expansion of the original RoBERTa model, that was unsupervisedly trained on 2.5TB of CommonCrawl web data on 100 languages [39]. Our analysis is conducted in an instance of the XLM-RoBERTa-large model that was finetunned over xnli and anli benchmark datasets to improve its zero-shot text classification [40]. Zero-shot text classification is the process of classifying textual data without using it as training input for the model.

4.2 Preprocessing

The files generated through the Twitter API store data in JSON format. In the process of parsing a tweet object, the following fields are parsed: the tweet's unique identifier, the text of the tweet, the user's unique identifier, the tweet's creation date, any hashtags included in the tweet, information about retweeted status, and the language used in the tweet. If the tweet is a retweet, we only keep the original tweet's identifier. For these types of tweets, we also retrieve both the hashtags from the retweet and the original tweet. When parsing tweet text, we have two cases that we handle differently. Initially, for original tweets, we try to get the text from the "extended tweet", "full text", and "text" fields in that order. In the case of retweets, we attempt to get the text from the "retweeted status" object in order from the fields "extended tweet" and "full text". If these fields do not exist, we merge the original tweet's text and retweet's text from the Twitter object to get the full text for our analysis.

Additionally, we perform some cleaning in texts before storing them in our database. We remove any URLs from text as well as any special characters such as tabs or newline characters. We also remove any leading or trailing spaces from the text and replace any double spaces within the text with single spaces.

After parsing and cleaning the tweet objects, we store them in a MongoDB database. Before storing a tweet, we discard any tweets that have text not suitable for analysis, such as empty or too short texts or a piece of text automatically generated by Twitter, indicating that the account is temporarily unavailable. Lastly, we ensure that all tweets stored in the database are unique to prevent duplicates from being imported for analysis.

4.3 Sentiment Analysis

We perform sentiment analysis by inserting as input in the model data of a whole day based on the creation date of tweets. As for the preprocessing phase, we divide the handling of texts into two categories: original tweets and retweets. For original tweets we just retrieve the text and tweet identifier stored in the object. But for retweets we have several cases to consider. First of all, if we have the original tweet that is retweeted in our database, we check if its text is the same as the text of the retweet. If it is a duplicate, we increase a multiplier index for the original tweet to later multiply its sentiment scores and reduce the amount of data imported into the model. Then if we have the original tweet in the database and the text is not the same, we collect the text and the identifier of the author of the original tweet. Lastly, if we don't have we have the original tweet in the database we just get the text and the ID of the user that retweeted. Before entering the texts into the pipeline we also replace words with the same meaning with the labels that we have established. We do this to take more accurate results from the model on words and topics that have almost the same semantics. For example, we replace some variations of the term"COVID-19" (e.g., "covid19", "covid-19", "coronavirus", "pandemic") with the classification label "covid", or the word "isolation" with the classification label "lockdown".
For classification labels we decided to use 9 different labels on topics that we thought were discussed in the public discourse. The classification labels are:

- cases
- conspiracy
- covid
- deaths
- lockdown
- masks
- propaganda
- vaccine
- 5G

Focusing on the reasoning behind each label we used "covid" to capture the emotion of the public towards this novel disease. "Cases" and "deaths" were selected because of the discussions for the daily reports on new COVID-19 infections and deaths. "Masks", "lockdown", and "vaccine" are some of the precautionary measures to prevent the spread of the disease and we wanted to see how the people responded to them. "Conspiracy", "propaganda", and "5G" were chosen regarding several conspiracy theories circling the social media and Internet about this new disease being a hoax.
We also separate the classification labels to get the sentiment on two emotions: positive and negative. So we have 18 different classification labels introduced to the model.

We import all the texts of the day and the classification labels using the zero-shot classification pipeline. We use the pipeline provided by the Transformers library and we instantiate it with a batch size of 16, 4 gradient accumulation steps, and the Adafactor optimizer. To run this high-complexity model we used a NVIDIA GeForce RTX 3080Ti GPU. From the output of the model and for each tweet ID we take the scores from each label multiplied by the multiplier index if it appeared multiple times in the dataset. We store this data in a CSV file for that day, which contains all analyzed tweet IDs and their sentiment scores. Finally, we calculate the daily average score for each label and store it in a separate file.

4.4 Suspended Accounts

By extracting all the individual users in our database we can input them in Twitter API and get back the status of their accounts. The API outputs a JSON file containing information about the reason for the user's unavailability, as well as the time of the account's deactivation. The reasons for an account not being visible through Twitter API include the account being "deactivated", "deleted", "protected", or "suspended". "Deactivated", "deleted", or "protected" statuses are all actions taken by the user. In the first two cases, deactivation initiates the process to permanently delete the Twitter account, and after a 30-day deactivation window, the account is permanently deleted [41]. In the last case, being "protected" makes the tweets visible only to the account's followers and invisible to the rest of the users and search engines.

We focus on the last case where an account is suspended by Twitter for violating the Twitter Rules. We use the JSON file provided by Twitter API to identify the user IDs of suspended accounts. After searching our database, we found that 807,796 users were suspended, accounting for 0.74% of total users. We also found 14,557,493 tweets, representing 7.52% of all tweets posted by these users. We filter out the sentiment scores of suspended users' tweets using CSV files we previously generated and calculate the daily average sentiment score for their accounts. Finally, we use our database to determine the daily count of tweets posted by suspended users.

4.5 Latent Dirichlet Allocation

For the analysis of conversations related to the novel coronavirus during this period, we opt to utilize the Latent Dirichlet Allocation (LDA) algorithm. Due to the large size of our dataset, it was impossible to apply the LDA in the entirety of the data. So we decided to narrow down our area of research using LDA to discover what suspended users were discussing and try to record any spam or malicious behavior of these users. By examining the sentiment analysis conducted before on these accounts we were able to pinpoint peaks in sentiment scores on the most prevalent labels. To identify outlier days for specific labels, we filter for days with sentiment score in the 99^{th} percentile.

To begin running the LDA algorithm, we need to extract the tweet texts from the database. This process involves identifying the days that had scores in the 99^{th} percentile for each label. For those days, we extract the tweets with the highest sentiment score for the label that we are focusing on. For instance, if we want to run LDA on the "positive for covid" label on May 23, 2020, we open the CSV file with sentiment scores of all suspended users' tweets for that day. We only keep the positive scores and compare them to the tweets we are interested in. To extract those tweets, we retain only the tweet IDs and texts with the highest sentiment score on the "positive for covid" label over all the other labels.

Before we can use the LDA algorithm, we need to preprocess the tweet text to ensure optimal results. For each label and day, we input all texts from the file that we saved before. Afterward, we process each tweet text by converting it to lowercase and splitting it into individual keywords. We remove common English stop words using the NLTK corpus and perform lemmatization using the Wordnet interface. This step converts each word to its base form by removing prefixes and suffixes, with the Morphy lemmatizer returning the final lemmatized version of each word in the corpus.

After the cleaning process of texts we are mapping the words to integer IDs with the Dictionary class. To perform the LDA algorithm we use the gensim library. To determine the number of topics for the algorithm we calculate the coherence score for topics in the range [2, 30] using the UMass coherence score. It calculates how often two words, $w_i and w_j$ appear together in the corpus and it's defined as

$$C_{UMass}(w_i, w_j) = \log \frac{D(w_i, w_j) + 1}{D(w_i)}$$

,where $D(w_i, w_j)$ indicates how many times words w_i and w_j appear together in documents, and $D(w_i)$ is how many time word w_i appeared alone. The coherence score is then determined based on these calculations, with a higher score indicating better coherence. So for each number of topics from in range [2, 30], the coherence score is calculated and we pick the number of topics with the highest coherence score. Finally, we run the gensim LDA model to extract the topics and display the results by creating a plot for each topic that shows its top ten most frequent words and frequency score and we provide a pyLDAvis HTML page as an output.

Sometimes, when we examine a topic and its most commonly used words, we may notice that all of the top words have a similar high score. This could indicate that a certain tweet using those same keywords has heavily influenced the topic. Although the exact text of the tweet may not be the same, variations of those keywords are being used. To find the original tweet that generated the topic, we analyze each topic by comparing its top ten most frequent words to every tweet. If a tweet matches all ten keywords and has been repeated more than five times, there is a possibility that it generated the topic. Moreover, we inspect the content within tweet texts to investigate the discussions and extract the proportion of retweets originating from suspended users, to uncover any orchestrated bot activity involved in spreading misinformation.

4.6 Graphs

In our research to identify any coordinated malicious communities, we are using graphs as a tool. Our goal is to analyze the tweets of suspended accounts that had the greatest sentiment score as either "positive for conspiracy" or "negative for conspiracy" to see if there is a pattern. We want to recreate the relationship between the original posters and the people who retweeted them. So we create a directed graph with users as nodes and weighted edges showing the number of retweets a user made to another user in a day.

Initially, we had to create the DOT language files that would generate the directed graphs. In a similar way to the LDA implementation, we select the days that the labels "positive for conspiracy" and "negative for conspiracy" had a sentiment score on the 95^{th} percentile. Subsequently, we extract files for these days with sentiment scores for each tweet and filtered out tweets with the highest scores for the labels "positive for conspiracy" or "negative for conspiracy". We then group all users into categories of "Active", "Suspended", "Deleted", "Protected", and "Deactivated". After accessing the MongoDB we generate all the retweet relationships with a weighted edge and write them in the file in a DOT-language form like this:

$$user1 - > \ user2 \ [weight = \#Retweets]$$

Next, we utilize Gephi, an open-source visualization software for graphs and networks, to open the DOT files we created. We then assign a color to each group of users in the diagram to make them easily distinguishable. The nodes in the network are each assigned a color based on the group to which the user belongs, while the edges are assigned a color based on the group of the user who retweeted (the source node). This allows for easy identification of the different user groups and the connections between them. This can also help reveal patterns or trends in the data, such as which groups are more likely to retweet each other or which groups are more central in the network. For our study, we only kept the "active" and "suspended" user groups in the graphs.

We utilize the OpenOrd algorithm to expand and produce the final layout of the graph network. The OpenOrd algorithm is a force-directed layout algorithm that can scale to over 1 million nodes, making it ideal for large graphs. This algorithm aims to better distinguish clusters, which suits us for our goal to distinguish the clusters of retweets by suspended accounts.

To gather data on retweets from suspended users, we manually examined the areas that had the highest number of retweets. We then identify the original poster by collecting their user ID from the center of these clusters. This process was repeated for each day to accumulate statistics on the number of retweets from both suspended and non-suspended users. Additionally, we gather all tweet text associated with these clusters to examine their content.

4.7 Tools and Libraries Used

The main programming language that was used for this work is Python3. Another tool that was being used was Twitter API for accessing the data. Also, it was used to get the activity status of the users in the dataset, specifically to check for all the suspended users. For storing the data, we used the MongoDB database to store and access all the tweet datasets. For the visualization of the graph networks, it was used the open-source visualization software Gephi. The source code can be found on Github here.

4.7.1 Python libraries

Some key Python libraries/packages used are:

- Transformers used for the sentiment analysis. Transformers is a Python library by Hugging Face featuring state-of-the-art pre-trained models for Natural Language Processing (NLP) and Machine Learning. Through PyTorch, its tokenizer and pipeline were used to perform the sentiment analysis on the XLM-RoBERTa-large model.

- Pymongo was used to work with the MongoDB database from Python.

- Pandas was used for manipulating, cleaning, and transforming the data as well as performing data analysis.

- Matplotlib library was the main tool for data visualizations and plotting.

- To implement the LDA algorithm, we utilized a combination of the Natural Language Toolkit (NLTK) and gensim libraries. The NLTK library was used for the preprocessing steps of the text of the tweets, using the Wordnet lemmatizer as well as its list of stop words that were removed from the text. On the other hand, gensim provided the LDA model, the Coherence Model for determining the optimal number of topics, and the dictionary which converted the lemmatized words to integers. Additionally, the pyLDAvis visualization tool was used to display the output of the LDA topics.

5 Results

5.1 Sentiment Analysis

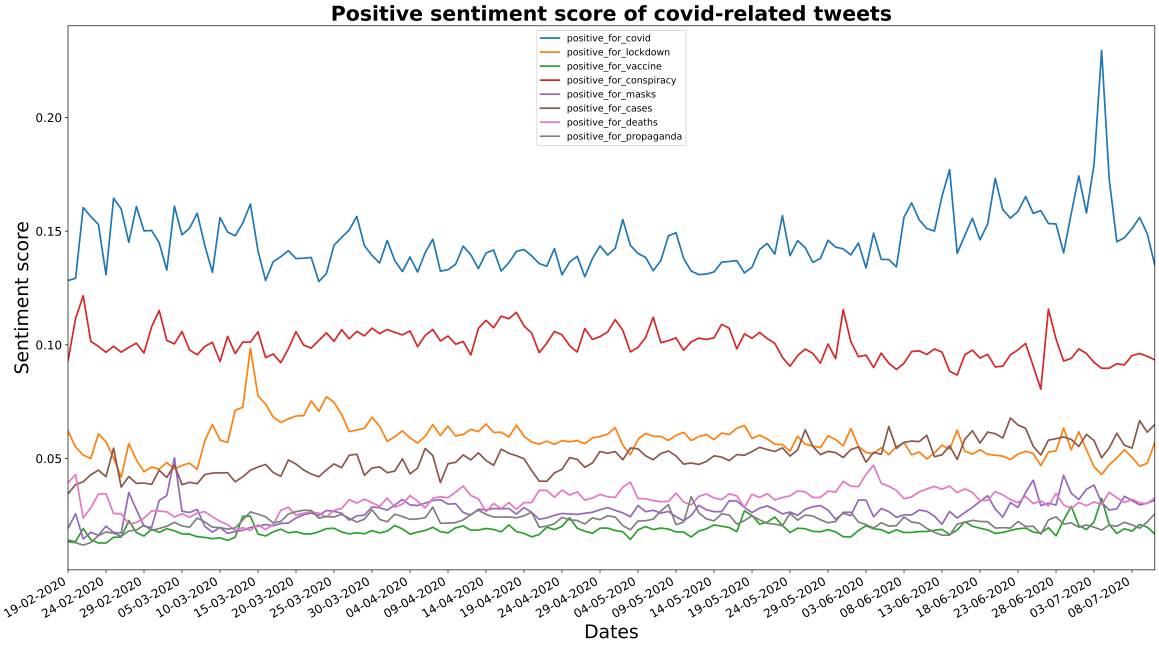

Figure 1: Daily average positive sentiment score of tweets on various labels

Figure 1 demonstrates the positive sentiment score over time for all labels. We decided to separate the plots for positive and negative sentiments for better clarity. All scores add up to one, together with negative sentiment scores. As we look at the figure we notice that the clear outstanding label that people discussed was "covid" throughout this period. A bit lower, in second place, is the "conspiracy" label which keeps its score relatively stable this whole period. The "lockdown" label rises in the middle of March, the period that the lockdowns started to begin worldwide, and falls later as time passes. The "cases" label starts lower than conspiracy but it gradually rises until it passes lockdown for several days at the start of the summer. Next are "masks" and "deaths", with masks having a peak at the start of March while the "deaths" label started rising more as months passed. As for "propaganda" and "vaccine" labels their score was relatively low with some small rises. But for "vaccine" we can note some rises at the end of the figure and it is important to note that this was a time when vaccines were not even produced so there was news and discussion about their use. Lastly, the "5G" label had negligible positive and negative sentiment scores so it was dropped from the figures.

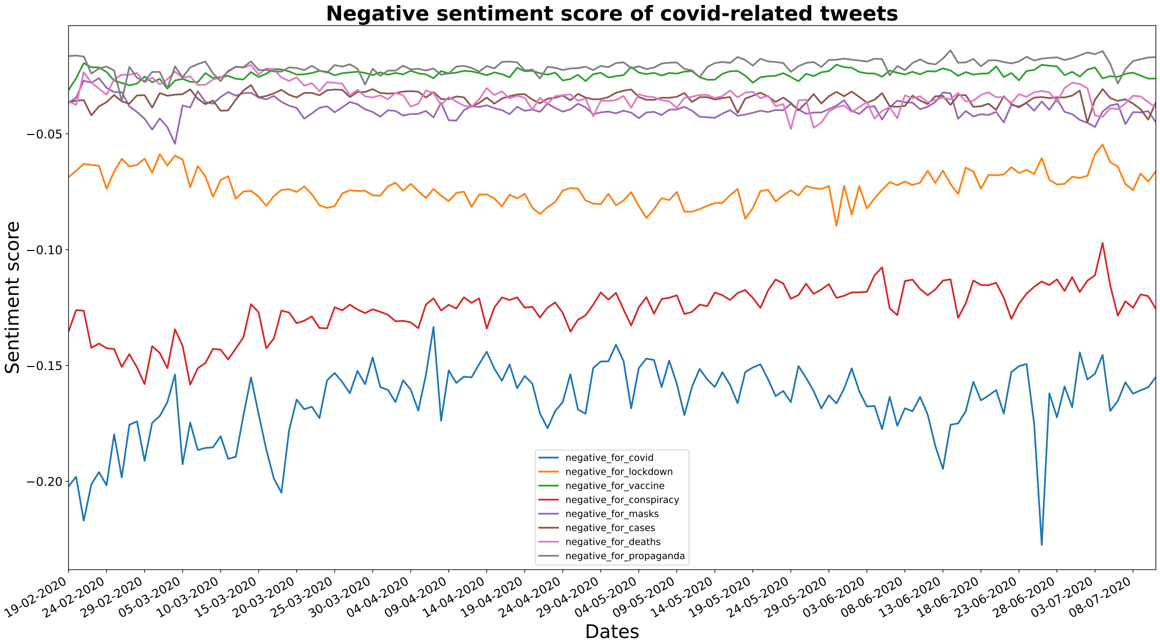

Figure 2: Daily average negative sentiment score of tweets on various labels

Figure 2 demonstrates the negative sentiment score over time for all labels. As a score gets lower, the more negative sentiment it has. We have a similar look here for the first three labels, with "covid" being the first, "conspiracy" being the second, and "lockdown" being the third. Following are "masks", "deaths" and "cases" around the same score. Last are the "vaccine" and "propaganda" labels. It’s worth noting that in both figures the first three labels consistently have more negative sentiment scores than positive ones.

5.2 Sentiment Analysis - Suspended Users

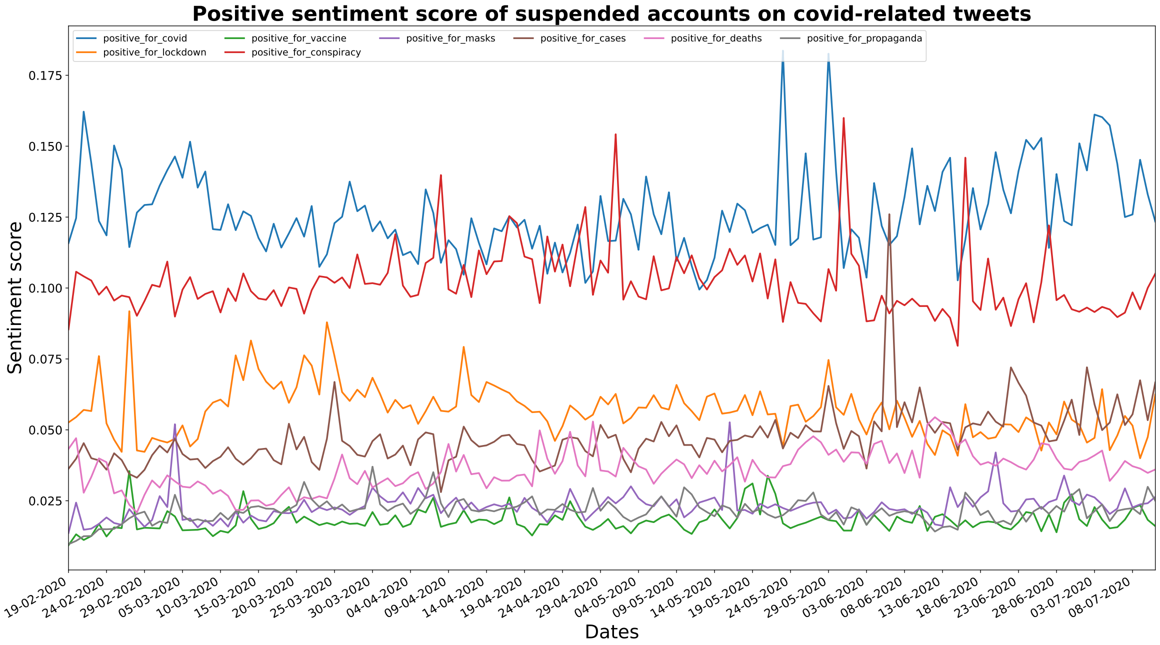

Figure 3: Daily average positive sentiment for tweets of suspended accounts

Figures 3 and 4 demonstrate the sentiment analysis of the suspended users' tweets. Comparing them to the corresponding figures 1 and 2, we observe that for positive sentiment score, the "covid" label remains in the lead. However, its mean score dropped from 0.145 to 0.126. The "conspiracy" label received higher sentiment scores and came closer to the score of the first label. During the first two months, the "lockdown" label had some high values, but its sentiment score fell later. The "cases" label, on the other hand, rose over time and passed the "lockdown" label during the last period of the data. Additionally, we noticed that the "deaths" label has some higher scores. The analysis also revealed some key peaks, such as "covid" on 23-5 and 29-5, "conspiracy" on 31-5 and 15-6, "cases" on 6-6, and "lockdown" on 27-2. We will review these peaks later in LDA.

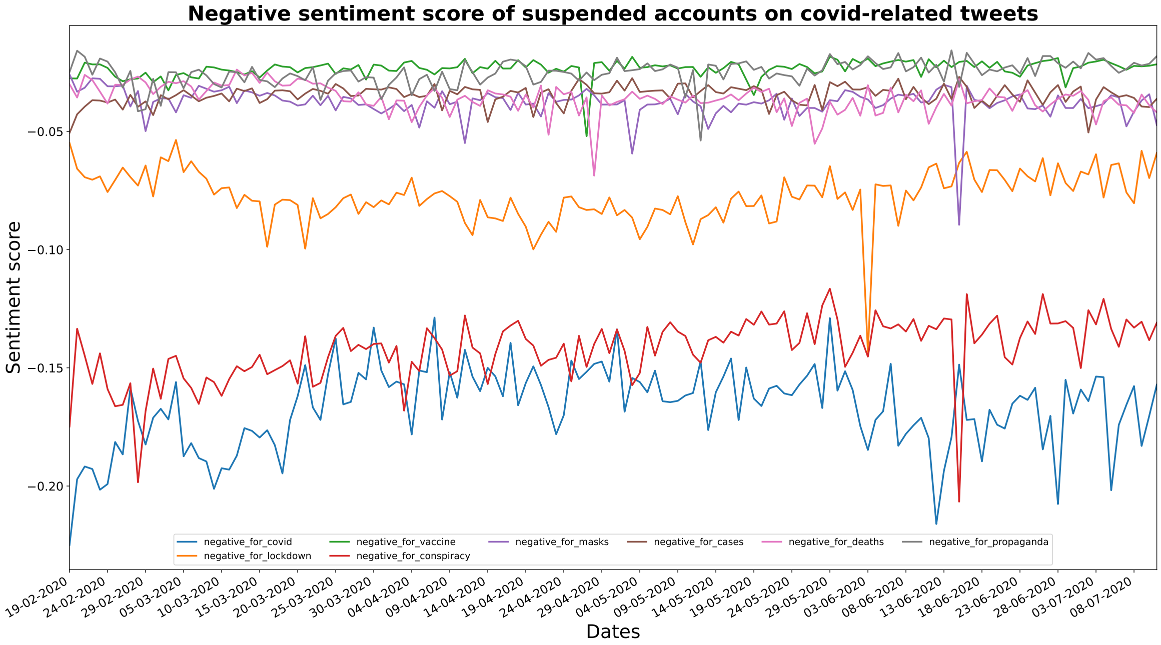

Figure 4: Daily average negative sentiment for tweets of suspended accounts

As for negative sentiment in the above figure, the labels "covid", "conspiracy", and "lockdown" hold the first three highest sentiment scores, with "covid" having a similar mean score with the general sentiment but the "conspiracy" score had an increase. All of the rest labels remained around the same score with some fluctuations. Also, some key peaks that we will examine later in LDA are for "covid" in 12-6, "conspiracy" in 15-6, "masks" in 15-6, "lockdown" in 3-6, and "deaths" in 28-4.

5.3 Data Analysis

To better understand the composition of this large dataset we examine some key parameters like the daily volume of tweets, which languages were most prominent in tweets, and which hashtags were most popular among users. For the daily volume of tweets, we depict in the same figure both the daily number of tweets found in our dataset as well as the number of users that created those tweets. Then we focus on the suspended accounts and examine the amount of tweets they post compared to the whole population of the dataset as well as the average amount of posts of every user.

5.3.1 Temporal distribution of tweets

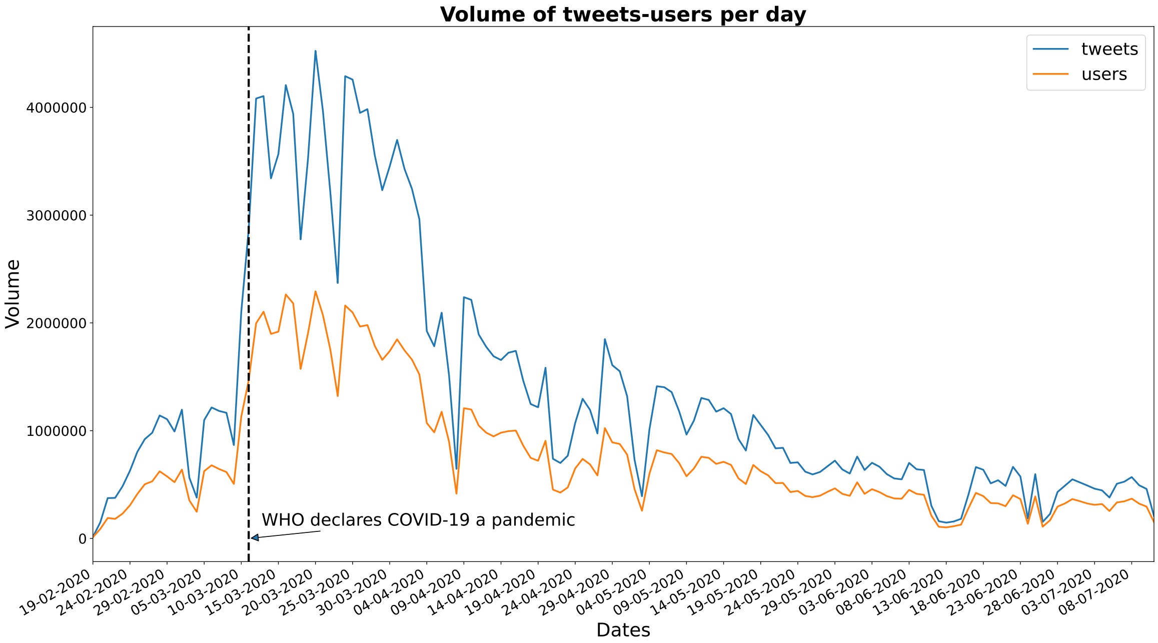

Figure 5: Temporal distribution of daily tweet activity by individual users

Figure 5 shows the number of created tweets and the amount of unique users that created them daily. We observe that there was a huge spike in mid-March after the World Health Organization declared COVID-19 a pandemic [1]. The day with the largest traffic on tweets was March 20, 2020, with 4,524,507 tweets posted by 2,292,682 unique users, which is also the highest amount of users creating a post in a day in our dataset. This high volume lasts until the start of April when we see a decline afterwards. At this high trending period users post on average almost 2 times a day but this also is reduced after a while and the number of posts and unique users converge.

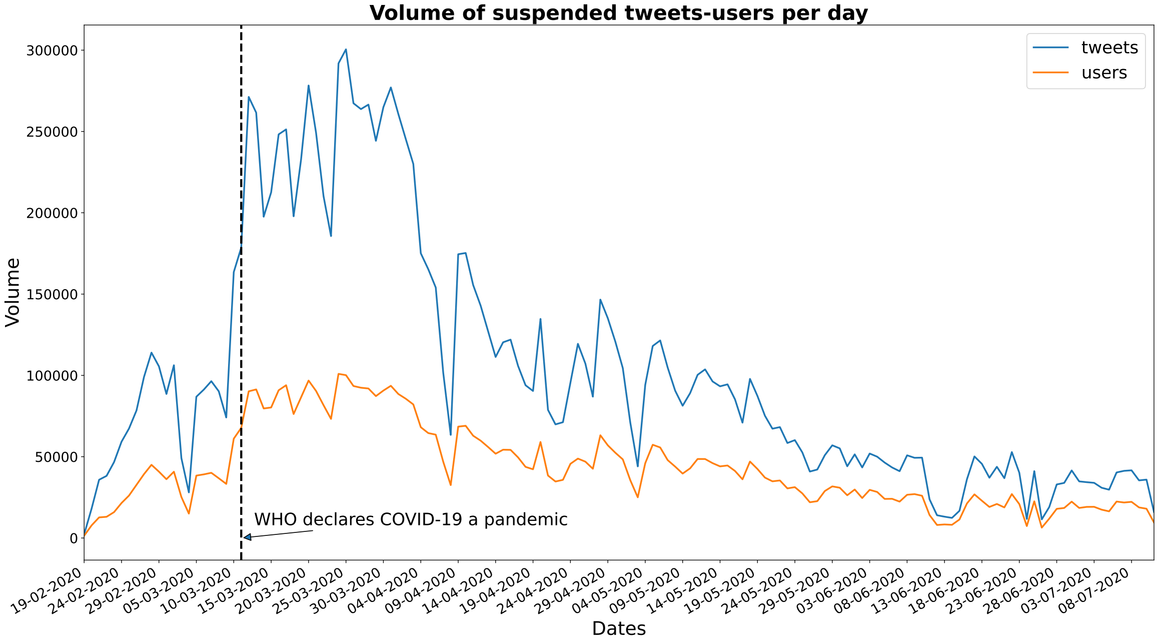

Figure 6: Temporal distribution of daily tweet activity by suspended users

Figure 6 depicts the daily volume of tweets posted by suspended accounts. We see a similar trend in the rise of tweets from March 12 and onwards and a decline since the start of April. The highest volume is recorded on March 25 with 300,513 tweets posted by 100.150 unique users. The volume of tweets created compared to normal users is one or two orders of magnitude lower, but we observe a significantly higher ratio between tweets and unique users creating these tweets. On average, a suspended user posted 2.19 tweets per day, while a non-suspended user posted 1.66 tweets per day.

5.3.2 Languages

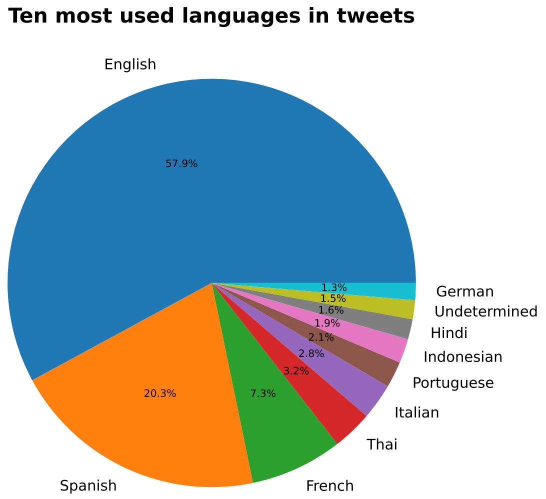

Figure 7: Pie chart for ten most popular languages in tweets

Figure 7 depicts the ten most used languages in tweets according to the Twitter metadata. English is the most prominent language used with 105,894,437 tweets. Together with Spanish accumulated almost 80% of the tweets. The third most popular language is French, followed by Thai, Italian, Portuguese, Indonesian, Hindi, and German. Something to notice is that around 1.5% of tweets are defined as undetermined language by Twitter. A tweet's language can be determined as "undetermined" for a few reasons, like the text of the tweet containing multiple languages, making it difficult for the language detection algorithm to confidently identify a single language. Also maybe the text of the tweet uses a rare language that is not supported by the language detection algorithm or it is composed of non-textual content such as emoticons.

5.3.3 Hashtags

A hashtag is a word or phrase preceded by a hash symbol (#) that is used to identify and categorize social media posts, particularly on Twitter and Instagram. Hashtags are used to help users find and follow conversations and topics that are relevant to their interests. To better understand the public discourse we grouped all hashtags that were used for each month. Below we present the most popular hashtags for the two busiest months on our dataset, March and April 2020.

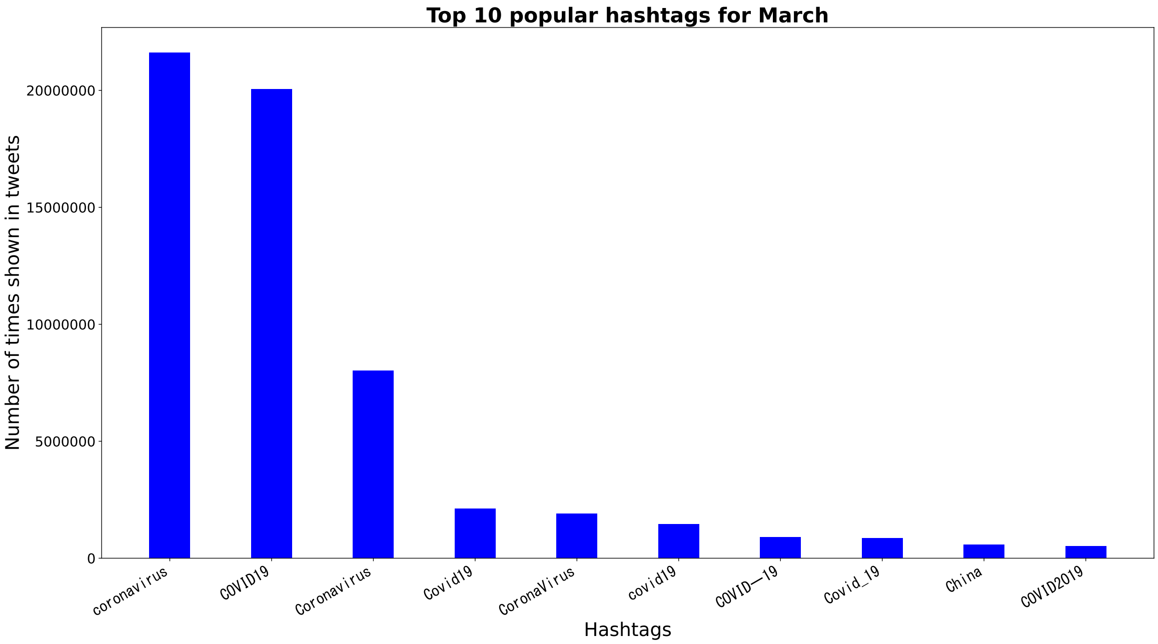

Figure 8: Most popular hashtags in March

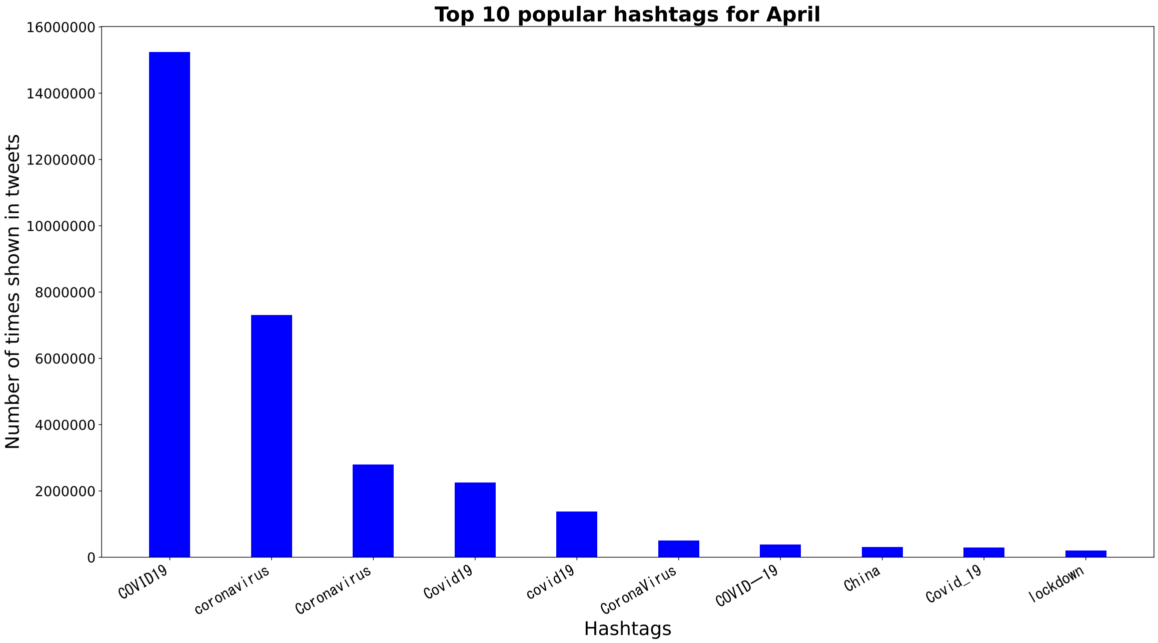

Figure 9: Most popular hashtags in April

As we can see in figure 9 the hashtags regarding the various names for the novel coronavirus dominated the first ten most tweeted hashtags. The only exception is the #China which is the origin country where the disease broke out. As for April, we notice a similar trend in hashtags with the only difference being the #lockdown, referring to the worldwide measures taken to prevent the wide of this new disease [42].

Next we take the aggregate of all the hashtags on the whole dataset and picking the 5 most popular hashtags we plot their monthly volume in figure 8.

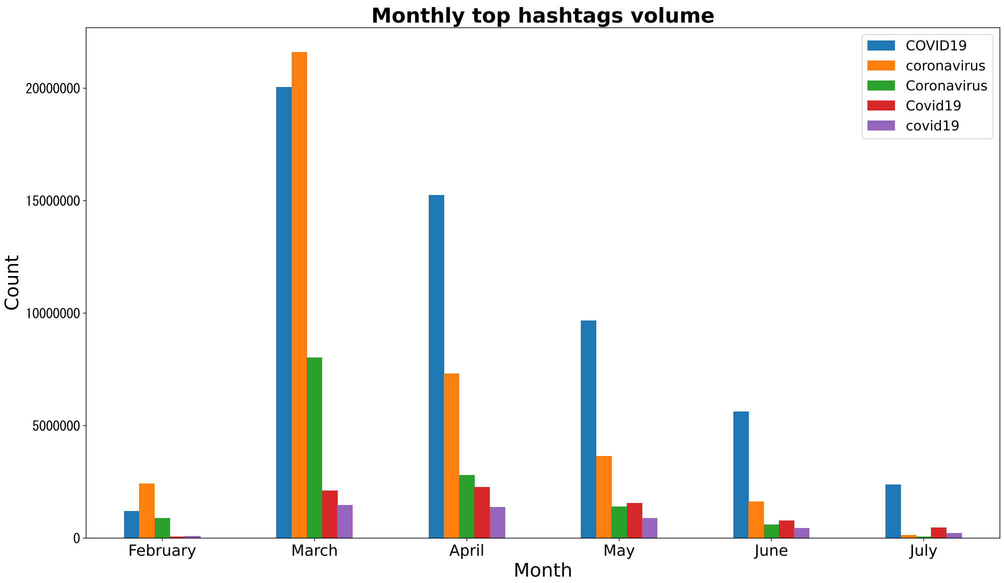

Figure 10: Monthly comparison for 5 most popular hashtags

During the first two months of the new disease, the most common hashtag used to refer to it was #coronavirus, followed by #COVID19 and #Coronavirus. However, in the following months, #COVID19 became the most prominent term with a decrease in the use of #coronavirus and #Coronavirus.

Afterwards, we remove with a regular expression any hashtags that contain the words "covid" or "coronavirus". We did this to extract information about other topics that users discussed besides the disease terms.

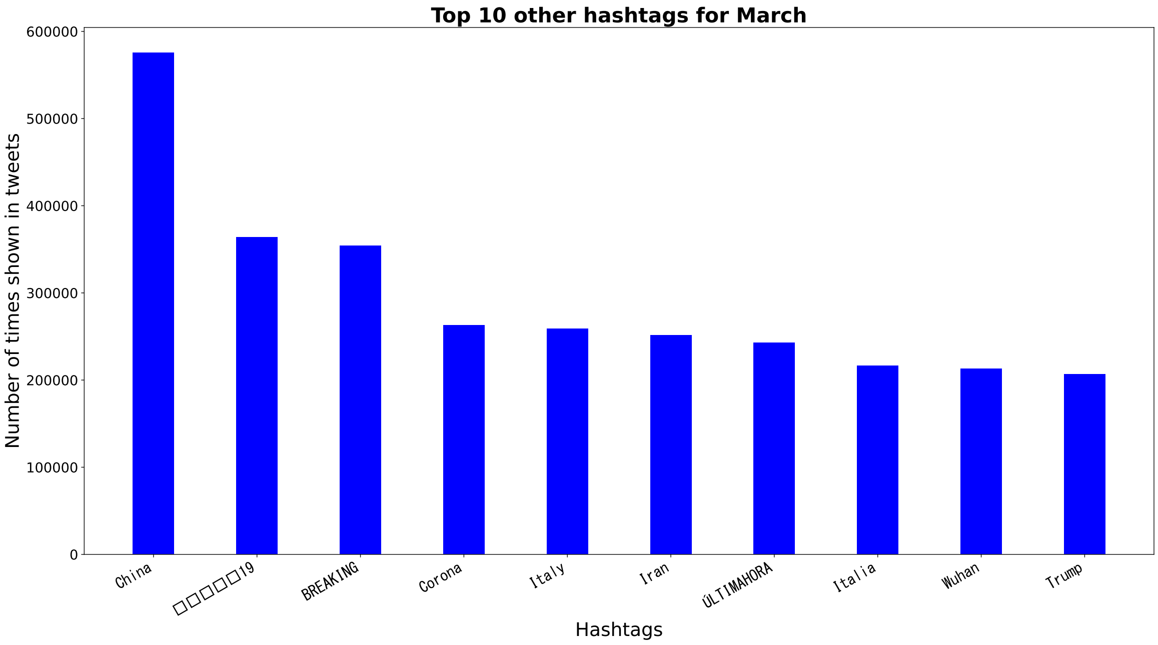

Figure 11: Most popular hashtags not including covid or coronavirus March 2020

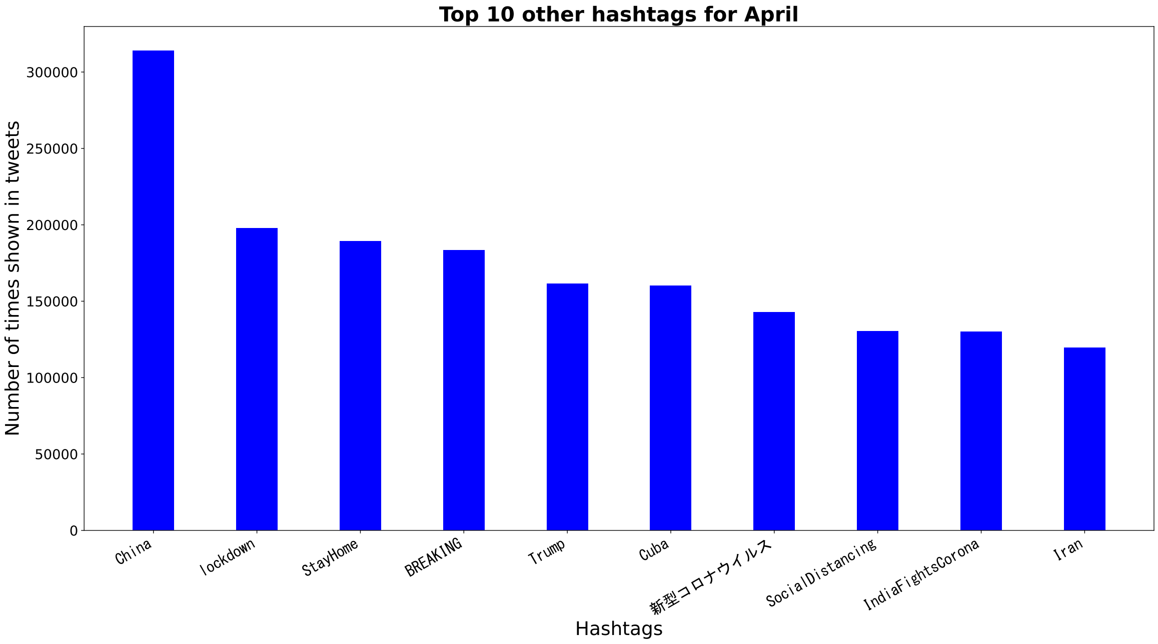

Figure 12: Most popular hashtags not including covid or coronavirus April 2020

As above in figures 9 and 10 we present the ten most used hashtags in March and April 2020. For March we observe several countries referred such as China, Italy, Iran, and the city of Wuhan where the novel coronavirus broke out. We also see #Corona and a non-Latin term that cannot be printed with Unicode with the number "19" at the suffix, so we can assume that is referring to COVID19. We observe the terms #BREAKING and #ULTIMAHORA (a Spanish-language news media) which refers to live news being posted. Lastly, we notice the #Trump, who at this time was the president of the United States of America.

In April we notice a different trend with four hashtags referring to public safety measures (#lockdown, #StayHome, #SocialDistancing, #IndiaFightsCorona). Similarly to March, we notice hashtags about countries (#China, #Cuba, #Iran). Lastly, there is a Japanese hashtag which translates as "novel coronavirus".

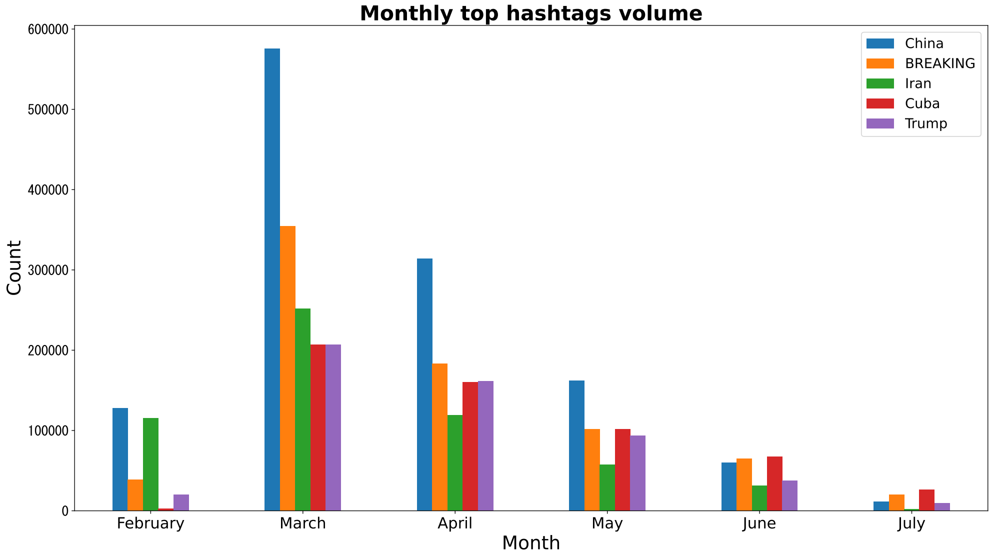

Figure 13: Monthly comparison for 5 most popular hashtags (not including covid or coronavirus)

The five most popular hashtags not including covid or coronavirus are #China, #BREAKING, #Iran, #Cuba, and #Trump. These hashtags all experienced the highest volume of tweets during March, with #China being the most tweeted hashtag for the first four months and #BREAKING, likely related to news, coming in second.

5.4 Tweet topics and their respective tweets

We are analyzing tweets from suspended users using the LDA algorithm. Specifically, we are examining selected labels on days with the highest sentiment scores. Our goal is to identify topics that are generated by frequent tweets, which may be posted or retweeted many times with the same or similar text. We examine for each topic if there are tweets that include all the top ten most frequent words. If these words have high scores, the topic is likely generated by these tweets that are probably posted or retweeted many times, with a similar text. We display the results using plots for each topic that show the top ten most frequent words and their frequency score. We have also included tables below figures to display the tweets that contain all of the top ten most frequent words for each topic, along with the tweet's ID, author ID, statistics on general retweets, and retweets by suspended users.

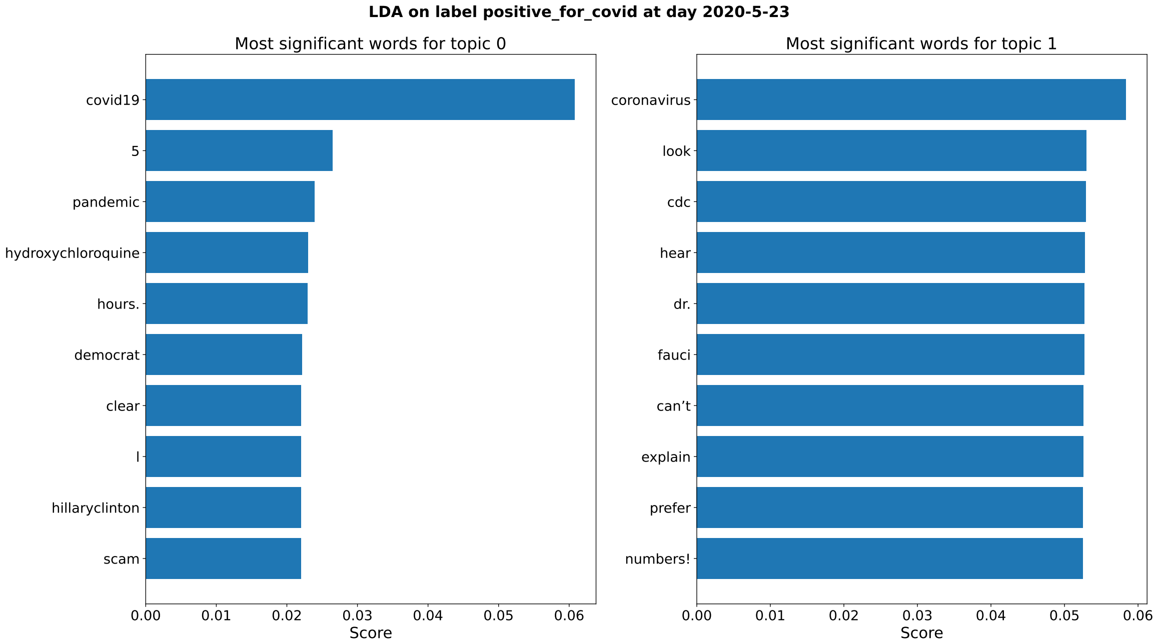

Figure 14: LDA on tweets with label "positive for covid" (May 23, 2020)

Topic	Text	Tweet ID	User ID	Retweets	Suspended RTs
0	HillaryClinton Hydroxychloroquine clears up Covid19 in 5 hours. The pandemic is a Democrat scam to hurt Trump2020. L	1264257990619566081	1250477466617090048	3173	1812
1	Wow! Look at these numbers! Can't wait to hear Dr. Fauci and the CDC explain why it is that coronavirus preferred to str	1264223201955102727	878247600096509952	14996	3996

Table 1: Topics to tweets from figure 14

Above we notice a tweet from topic 0 talking about the pandemic being a scam. Also, the tweet originating from topic 1 its text is truncated but if we access the original tweet it talks about COVID-19 preferring to strike Democrat states over Republican ones and it calls it also as "scamdemic".

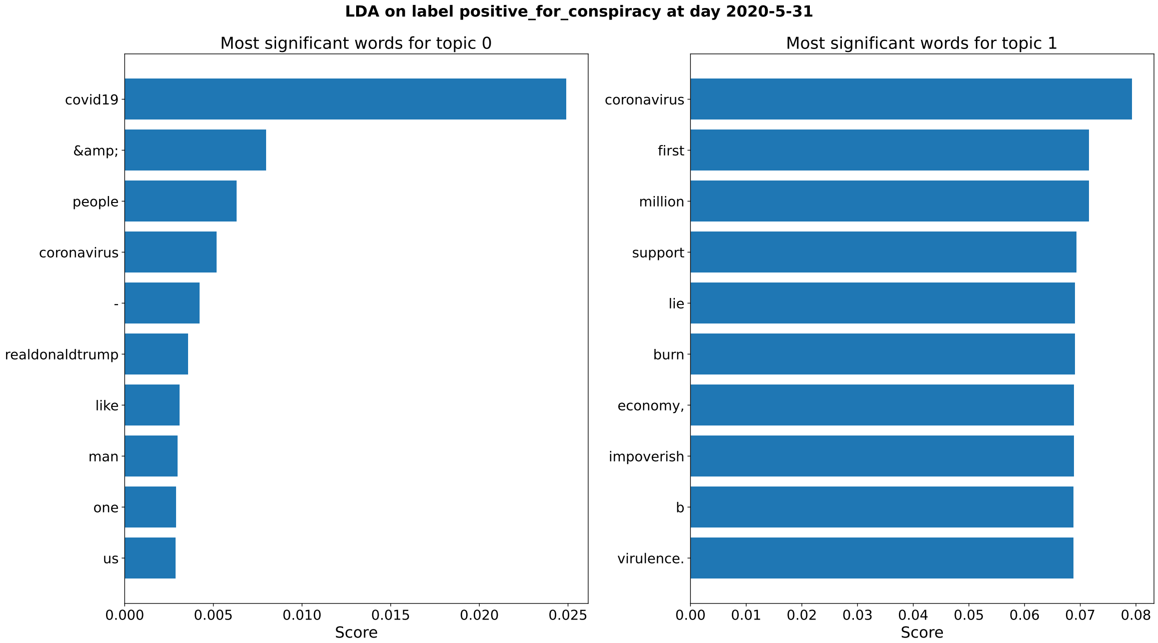

Figure 15: LDA on tweets with label "positive for conspiracy" (May 31, 2020)

Topic	Text	TweetID	UserID	Retweets	Suspended RTs
1	First they burned down our economy, impoverishing millions through the lie of coronavirus virulence. Now they support b	1267163158772408320	878247600096509952	30442	7953

Table 2: Topics to tweets from figure 15

Here in topic 1, the tweet is talking about coronavirus being a lie.

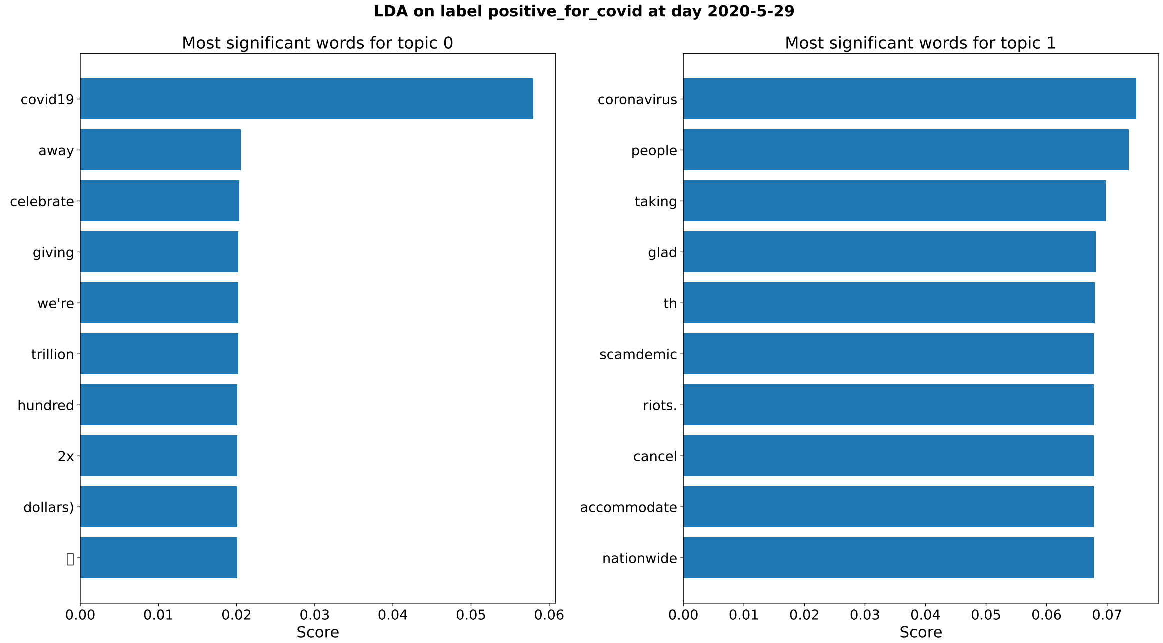

Figure 16: LDA on tweets with label "positive for covid" (May 29, 2020)

Topic	Text	TweetID	UserID	Retweets	Suspended RTs
0	GIVEAWAY We're giving away 2x 100.000.000.000.000 (two hundred trillion zimbabwe dollars) to celebrate COVID19 Mone	1266272969208131586	929318631242117120	1011	956
1	So glad the Coronavirus scamdemic was cancelled to accommodate the Left's nationwide anarchist riots. People taking th	1266396421923667968	878247600096509952	14135	4060

Table 3: Topics to tweets from figure 16

In topic 1, we also observe the term "scamdemic" which refers to the virus being a scam.

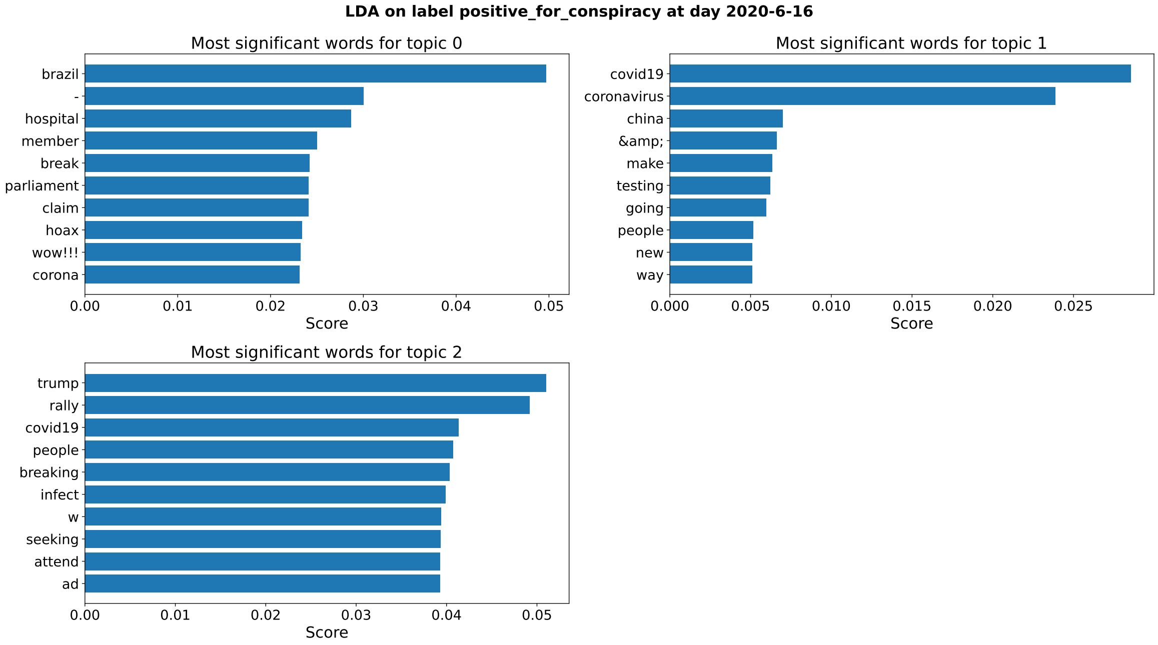

Figure 17: LDA on tweets with label "positive for conspiracy" (June 16, 2020)

Topic	Text	TweetID	UserID	Retweets	Suspended RTs
0	WOW!!! - The corona hoax was revealed in Brazil Members of Parliament of Brazil break into hospital that claimed to have	1272982565088092166	50434327	1964	913
2	BREAKING Craigslist ad seeking people infected with COVID19 to attend Trump rally in Tulsa, OK. RT—this is biological w	1272919717007765505	1249837543383859201	7888	2723

Table 4: Topics to tweets from figure 17

Here in topic 0, the disease is called as "corona hoax".

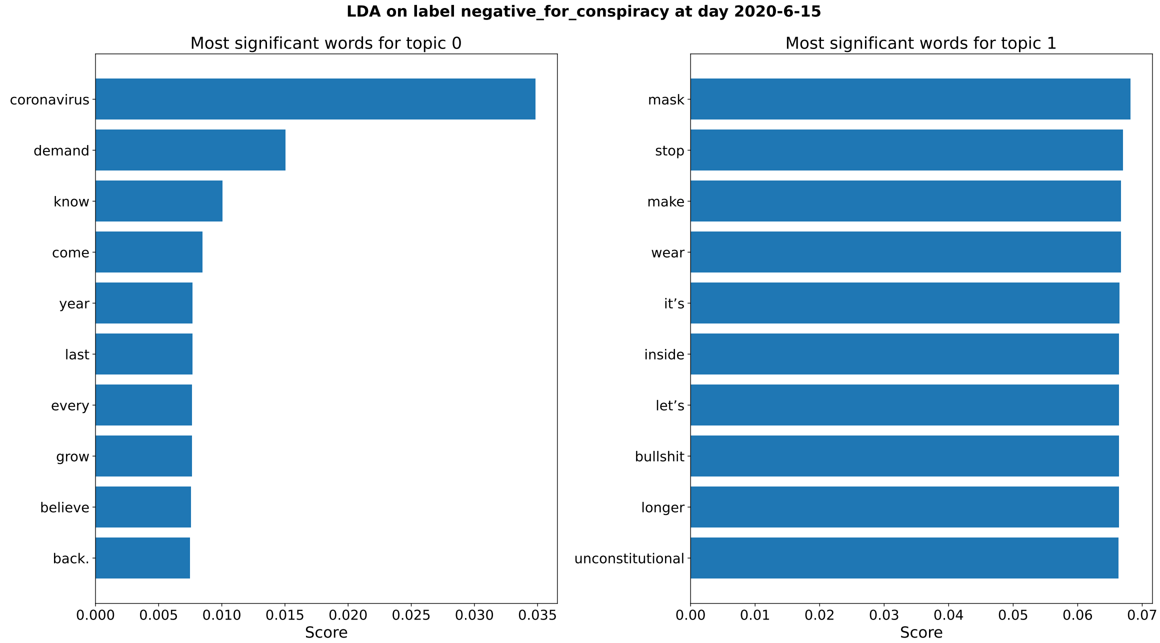

Figure 18: LDA on tweets with label "negative for conspiracy" (June 15, 2020)

Topic	Text	TweetID	UserID	Retweets	Suspended RTs
1	I will no longer wear a mask inside any business. It's unconstitutional to enforce. Let's make this bullshit stop now! Who	1272545366773096453	2410068528	8212	3070

Table 5: Topics to tweets from figure 18

The tweet in topic 1 expresses a refusal to wear a mask, claiming it's unconstitutional to enforce such a requirement.

For more topics coming from LDA and their respective tweets check the Appendix A.

5.5 Retweet Graphs

Displayed below are retweet graphs that show the relationship between users based on their retweets. The graphs highlight specific labels that received high sentiment scores on a particular day. The weighted edges represent the number of times one user retweeted another user. Red edges indicate retweeting by suspended users, while the green edges represent retweeting by active users.

Negative for conspiracy

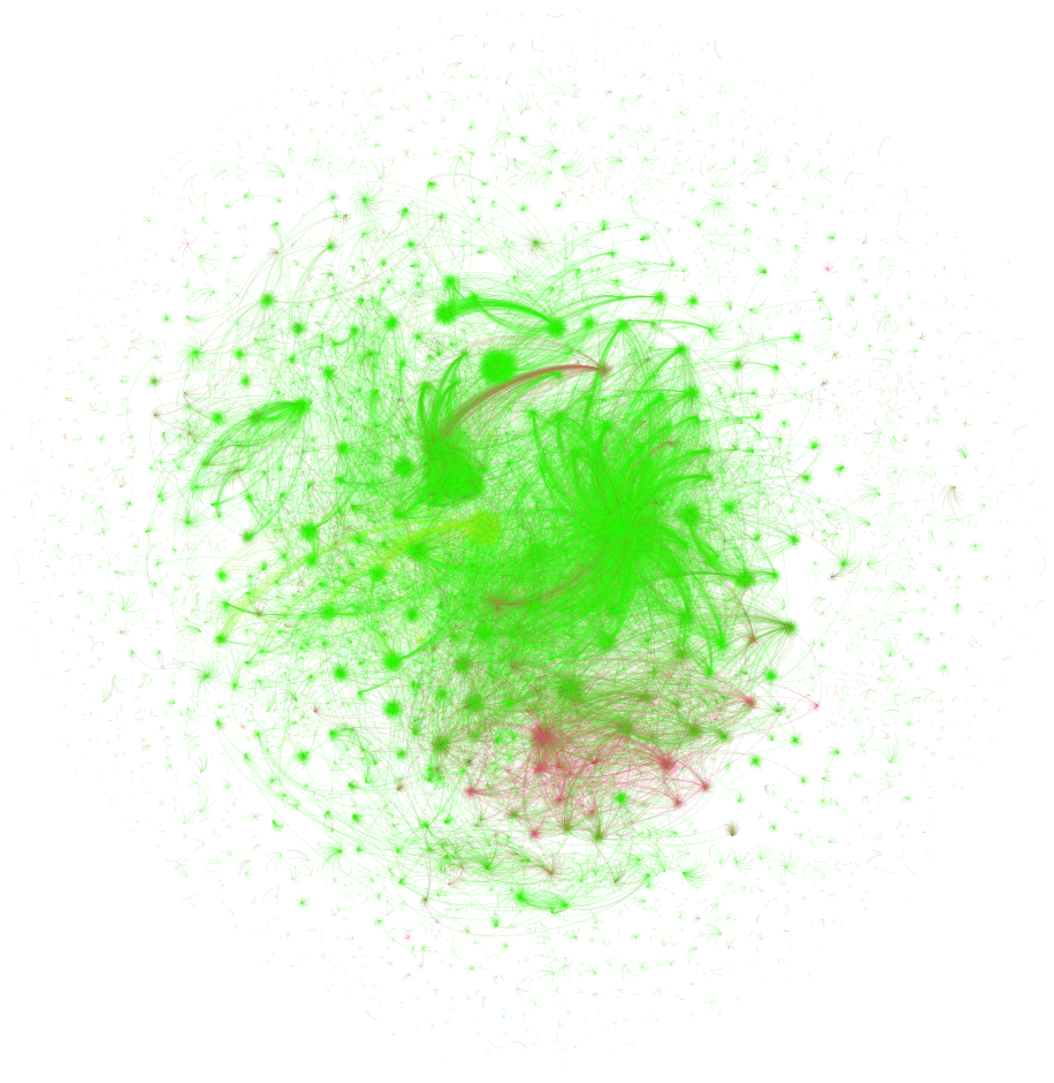

Figure 19: Retweet graph on the label "negative for conspiracy" (2020-2-29)

Positive for conspiracy

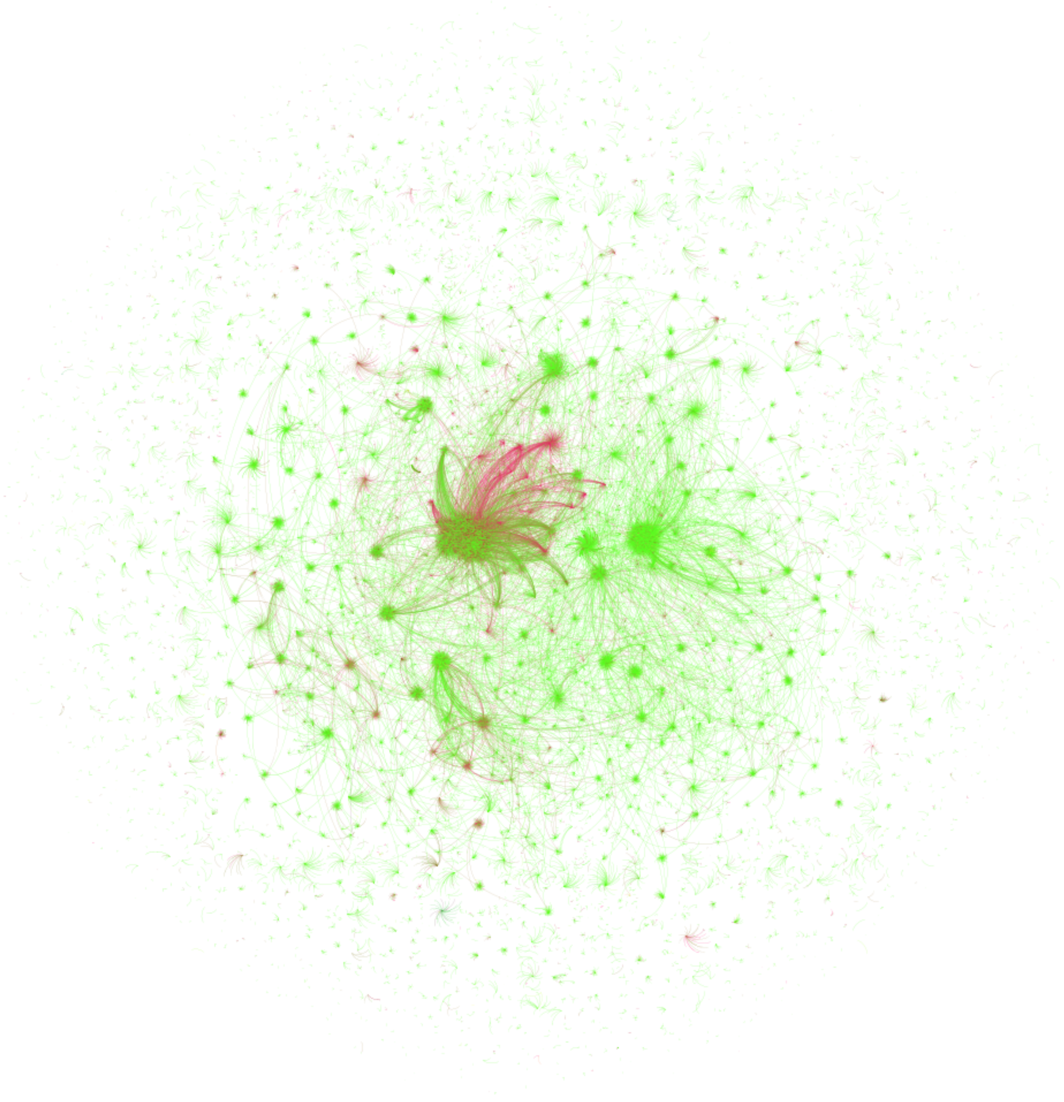

Figure 20: Retweet graph on the label "positive for conspiracy" (2020-4-18)

Here, we present the retweets per original poster on crowded clusters dominated by suspended users. We then proceed to contrast this data for retweet counts between suspended users and non-suspended users. Despite examining the texts from tweets in clusters with high retweet activity, we have not discovered any conclusive evidence.

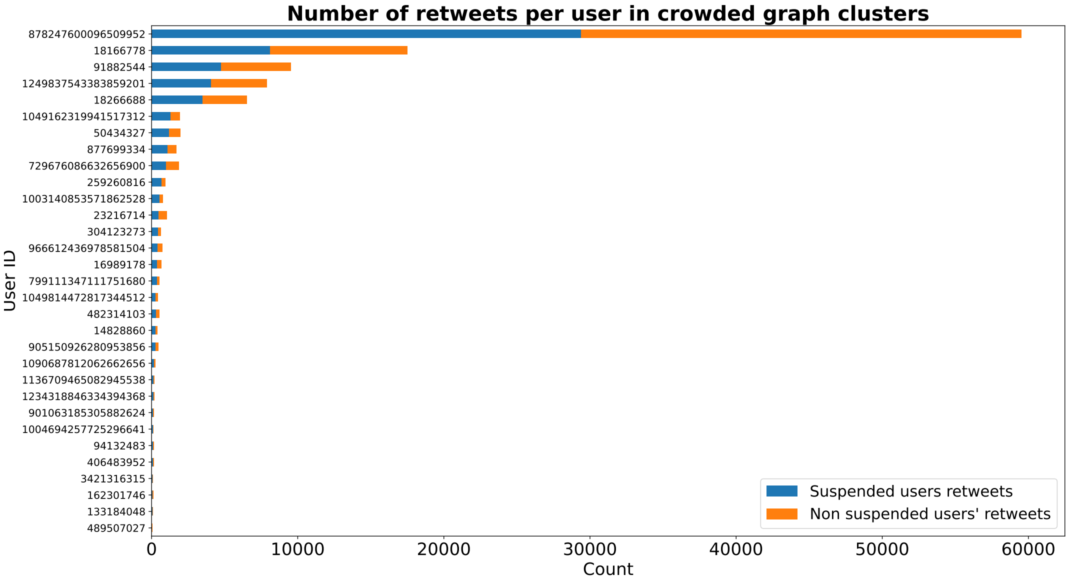

Figure 21: Comparison of retweets by suspended/non-suspended users on clusters (positive for conspiracy)

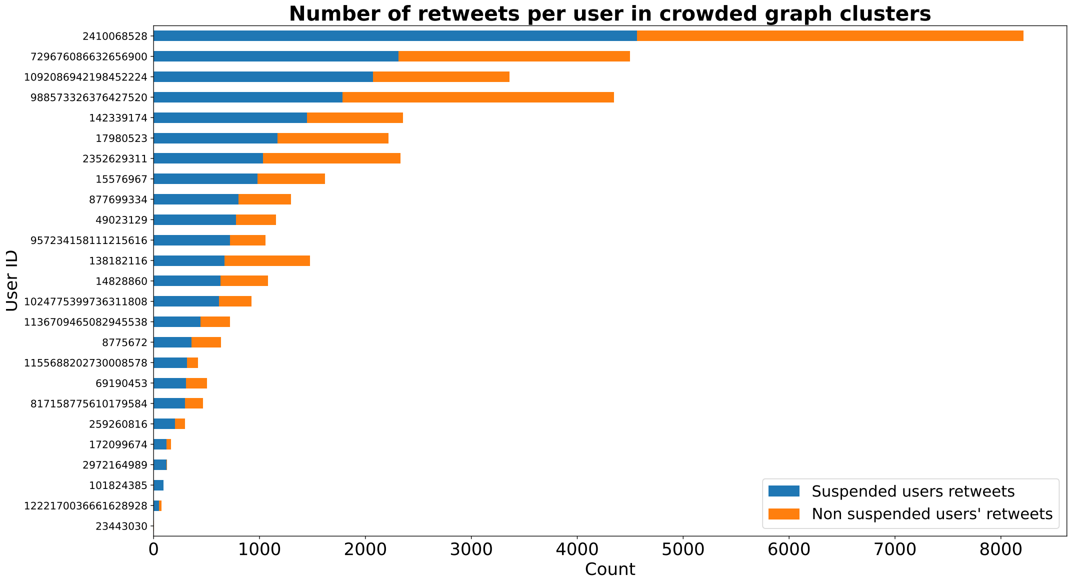

Figure 22: Comparison of retweets by suspended/non-suspended users on clusters (negative for conspiracy)

6 Conclusion

In this study, we conducted a thorough investigation of the digital discourse surrounding the COVID-19 pandemic on Twitter. Our sentiment analysis revealed a notable trend, spanning from positive to negative sentiments, around critical topics like COVID-19, lockdowns, and conspiracy theories. After examining suspended users, we noticed that these topics had higher sentiment scores compared to non-suspended users. Also, we found that the "conspiracy" label had higher sentiment scores among suspended accounts, suggesting that these users may be more likely to discuss conspiracy theories.

Exploring this vast dataset we we discovered that the highest level of traffic occurred on March 20th, 2020, with 4,524,507 tweets posted. This peak happened just a few days after the official declaration of COVID-19 as a pandemic. Regarding the suspended users there are in total of 807,796 suspended users in the dataset and 14,557,493 tweets created by them. Even though suspended users account for only 0.74% of the dataset, they contributed to 7.52% of the posts in our dataset. By comparing the volume of tweets and unique users posting over time we observe a higher ratio of suspended users tweeting than non-suspended ones. On average a suspended user posted 2.19 tweets per day while a non-suspended user posted 1.66 tweets per day. This observation may suggest that suspended users post more frequently to create more traffic and attract engagement. The most common languages detected in the tweets were English, Spanish, and French.

When looking at the most commonly used hashtags over the past few months, it's clear that many are related to the COVID-19 pandemic. Variations of the words "covid" and "coronavirus" make up the majority of the top hashtags. However, there are a few exceptions. The five most popular hashtags not including covid or coronavirus are #China, #BREAKING, #Iran, #Cuba and #Trump. During the first two months, several countries were heavily referenced in hashtags, including China, Italy, Iran, and Wuhan (where the virus was first identified). In April, we noticed a trend where four hashtags focused on public safety measures: #lockdown, #StayHome, #SocialDistancing, and #IndiaFightsCorona.

We run the Latent Dirichlet Allocation algorithm on specific days to identify frequently posted topics by suspended users. If the top 10 words in a topic have high scores, we assume that the tweet is being reposted exactly as it was before. We then provide statistics on the total number of retweets, as well as retweets by suspended users. We observe that these tweets often discuss health organizations like the CDC as well as prominent figures such as politicians and scientists. Also, they may touch upon topics like the conspiracy belief that COVID-19 is a hoax or scam and may involve calling for protests against government policies.

Finally, we created retweet graphs to detect clusters of suspended accounts that were excessively retweeting a particular person. Additionally, we have provided information on the number of retweets per original poster in congested clusters dominated by suspended users. We also compared the number of retweets between active and suspended users. Although we have identified some clusters with significantly high retweet activity by suspended accounts, we have not discovered any conclusive evidence.

The results of this study could provide valuable insights for policymakers, healthcare professionals, and communicators regarding how online conversations during crises unfold, with a particular emphasis on combating false information. One potential area for future research is to broaden the scope of the study by analyzing a larger dataset over a longer period, including the vaccine roll-out phase, to capture the initial reaction of tweets. Additionally, future investigations could focus on strategies for mitigating the spreading of misinformation and fake news on social media platforms.

References

[1] W. H. Organization *et al.*, "Who director-general's opening remarks at the media briefing on covid-19-11 march 2020," 2020.

[2] M. Mazur, M. Dang, and M. Vega, "Covid-19 and the march 2020 stock market crash. evidence from s&p1500," *Finance research letters*, vol. 38, p. 101690, 2021.

[3] G. K. Shahi, A. Dirkson, and T. A. Majchrzak, "An exploratory study of covid-19 misinformation on twitter," *Online social networks and media*, vol. 22, p. 100104, 2021.

[4] H. Rosenberg, S. Syed, and S. Rezaie, "The twitter pandemic: The critical role of twitter in the dissemination of medical information and misinformation during the covid-19 pandemic," *Canadian journal of emergency medicine*, vol. 22, no. 4, pp. 418–421, 2020.

[5] S. A. Memon and K. M. Carley, "Characterizing covid-19 misinformation communities using a novel twitter dataset," *arXiv preprint arXiv:2008.00791*, 2020.

[6] 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.

[7] 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.

[8] C. P. Rodríguez, B. V. Carballido, G. Redondo-Sama, M. Guo, M. Ramis, and R. Flecha, "False news around covid-19 circulated less on sina weibo than on twitter. how to overcome false information?," *International and Multidisciplinary Journal of Social Sciences*, vol. 9, no. 2, pp. 107–128, 2020.

[9] W. S. Paka, R. Bansal, A. Kaushik, S. Sengupta, and T. Chakraborty, "Cross-sean: A cross-stitch semi-supervised neural attention model for covid-19 fake news detection," *Applied Soft Computing*, vol. 107, p. 107393, 2021.

[10] W. Ceron, G. Gruszynski Sanseverino, M.-F. de Lima-Santos, and M. G. Quiles, "Covid-19 fake news diffusion across latin america," *Social Network Analysis and Mining*, vol. 11, no. 1, pp. 1–20, 2021.

[11] A. Glazkova, M. Glazkov, and T. Trifonov, "g2tmn at constraint@ aaai2021: exploiting ct-bert and ensembling learning for covid-19 fake news detection," in *International Workshop on Combating On line Ho st ile Posts in Regional Languages dur ing Emerge ncy Si tuation*, pp. 116–127, Springer, 2021.

[12] M. Heidari, S. Zad, P. Hajibabaee, M. Malekzadeh, S. HekmatiAthar, O. Uzuner, and J. H. Jones, "Bert model for fake news detection based on social bot activities in the covid-19 pandemic," in *2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, pp. 0103–0109, IEEE, 2021.

[13] D. S. Abdelminaam, F. H. Ismail, M. Taha, A. Taha, E. H. Houssein, and A. Nabil, "Coaid-deep: An optimized intelligent framework for automated detecting covid-19 misleading information on twitter," *Ieee Access*, vol. 9, pp. 27840–27867, 2021.

[14] S. Gundapu and R. Mamidi, "Transformer based automatic covid-19 fake news detection system," *arXiv preprint arXiv:2101.00180*, 2021.

[15] X. Zhou, A. Mulay, E. Ferrara, and R. Zafarani, "Recovery: A multimodal repository for covid-19 news credibility research," in *Proceedings of the 29th ACM international conference on information & knowledge management*, pp. 3205–3212, 2020.

[16] S. R. Rufai and C. Bunce, "World leaders' usage of twitter in response to the covid-19 pandemic: a content analysis," *Journal of public health*, vol. 42, no. 3, pp. 510–516, 2020.

[17] A. Rao, F. Morstatter, M. Hu, E. Chen, K. Burghardt, E. Ferrara, K. Lerman, *et al.*, "Political partisanship and antiscience attitudes in online discussions about covid-19: Twitter content analysis," *Journal of medical Internet research*, vol. 23, no. 6, p. e26692, 2021.

[18] Y. Li, S. Twersky, K. Ignace, M. Zhao, R. Purandare, B. Bennett-Jones, and S. R. Weaver, "Constructing and communicating covid-19 stigma on twitter: a content analysis of tweets during the early stage of the covid-19 outbreak," *International Journal of Environmental Research and Public Health*, vol. 17, no. 18, p. 6847, 2020.

[19] D. Scannell, L. Desens, M. Guadagno, Y. Tra, E. Acker, K. Sheridan, M. Rosner, J. Mathieu, and M. Fulk, "Covid-19 vaccine discourse on twitter: A content analysis of persuasion techniques, sentiment and mis/disinformation," *Journal of health communication*, vol. 26, no. 7, pp. 443–459, 2021.

[20] J. Griffith, H. Marani, H. Monkman, *et al.*, "Covid-19 vaccine hesitancy in canada: Content analysis of tweets using the theoretical domains framework," *Journal of medical Internet research*, vol. 23, no. 4, p. e26874, 2021.

[21] 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.

[22] K. H. Manguri, R. N. Ramadhan, and P. R. M. Amin, "Twitter sentiment analysis on worldwide covid-19 outbreaks," *Kurdistan Journal of Applied Research*, pp. 54–65, 2020.

[23] A. D. Dubey, "Twitter sentiment analysis during covid-19 outbreak," *Available at SSRN 3572023*, 2020.

[24] S. Boon-Itt, Y. Skunkan, *et al.*, "Public perception of the covid-19 pandemic on twitter: sentiment analysis and topic modeling study," *JMIR Public Health and Surveillance*, vol. 6, no. 4, p. e21978, 2020.

[25] B. P. Pokharel, "Twitter sentiment analysis during covid-19 outbreak in nepal," *Available at SSRN 3624719*, 2020.

[26] C. Villavicencio, J. J. Macrohon, X. A. Inbaraj, J.-H. Jeng, and J.-G. Hsieh, "Twitter sentiment analysis towards covid-19 vaccines in the philippines using naïve bayes," *Information*, vol. 12, no. 5, p. 204, 2021.

[27] R. Marcec and R. Likic, "Using twitter for sentiment analysis towards astrazeneca/oxford, pfizer/biontech and moderna covid-19 vaccines," *Postgraduate Medical Journal*, vol. 98, no. 1161, pp. 544–550, 2022.

[28] U. Naseem, I. Razzak, M. Khushi, P. W. Eklund, and J. Kim, "Covidsenti: A large-scale benchmark twitter data set for covid-19 sentiment analysis," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 4, pp. 1003–1015, 2021.

[29] O. Gencoglu, "Large-scale, language-agnostic discourse classification of tweets during covid-19," *Machine Learning and Knowledge Extraction*, vol. 2, no. 4, pp. 603–616, 2020.

[30] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic bert sentence embedding," *arXiv preprint arXiv:2007.01852*, 2020.

[31] E. Ferrara, "What types of covid-19 conspiracies are populated by twitter bots?," *arXiv preprint arXiv:2004.09531*, 2020.

[32] J. M. Banda, R. Tekumalla, G. Wang, J. Yu, T. Liu, Y. Ding, E. Artemova, E. Tutubalina, and G. Chowell, "A large-scale covid-19 twitter chatter dataset for open scientific research—an international collaboration," *Epidemiologia*, vol. 2, no. 3, pp. 315–324, 2021.

[33] S. Alqurashi, A. Alhindi, and E. Alanazi, "Large arabic twitter dataset on covid-19," *arXiv preprint arXiv:2004.04315*, 2020.

[34] M. Müller, M. Salathé, and P. E. Kummervold, "Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter," *arXiv preprint arXiv:2005.07503*, 2020.

[35] G. Ali and M. S. I. Malik, "Rumour identification on twitter as a function of novel textual and language-context features," *Multimedia Tools and Applications*, pp. 1–22, 2022.

[36] J. Xue, J. Chen, R. Hu, C. Chen, C. Zheng, Y. Su, T. Zhu, *et al.*, "Twitter discussions and emotions about the covid-19 pandemic: Machine learning approach," *Journal of medical Internet research*, vol. 22, no. 11, p. e20550, 2020.

[37] "Xlmroberta-large." https://huggingface.co/xlm-roberta-large. Accessed: 2022-09-21.

[38] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[39] “commoncrawl.” https://commoncrawl.org/. Accessed: 2022-09-21.

[40] “Vicgalle/xlm-roberta-large-xnli-anli.” https://huggingface.co/vicgalle/xlm-roberta-large-xnli-anli. Accessed: 2022-12-12.

[41] “How to deactivate your twitter account | twitter help.”

[42] J. Hamzelou, “World in lockdown,” 2020.

Appendix A

LDA topics and their respective tweets

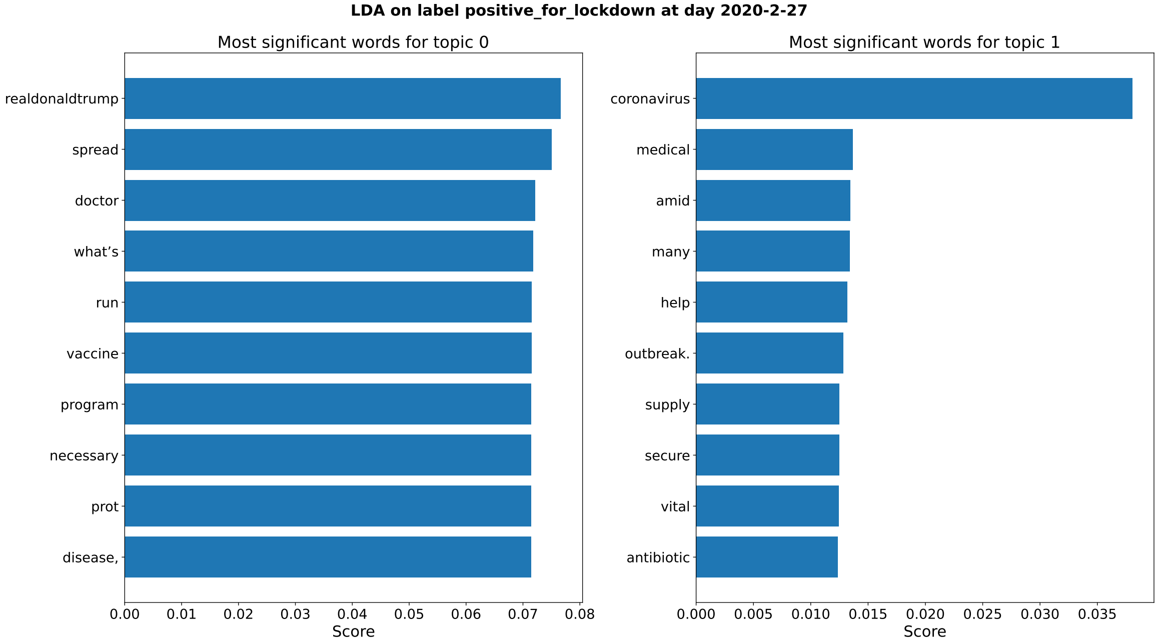

Figure 23: LDA on tweets with the label "positive for lockdown" (February 27, 2020)

Topic	Text	TweetID	UserID	Retweets	Suspended RTs
0	As a doctor who ran vaccine programs to decrease the spread of disease, realDonaldTrump is doing what's necessary to protect	1232852222414815232	55677432	24650	6682
1	My legislation will help secure our medical supply chains amid this coronavirus outbreak. Too many of our vital antibiotics	1233033351893782529	2352629311	3781	1126

Table 6: Topics to tweets from figure 23

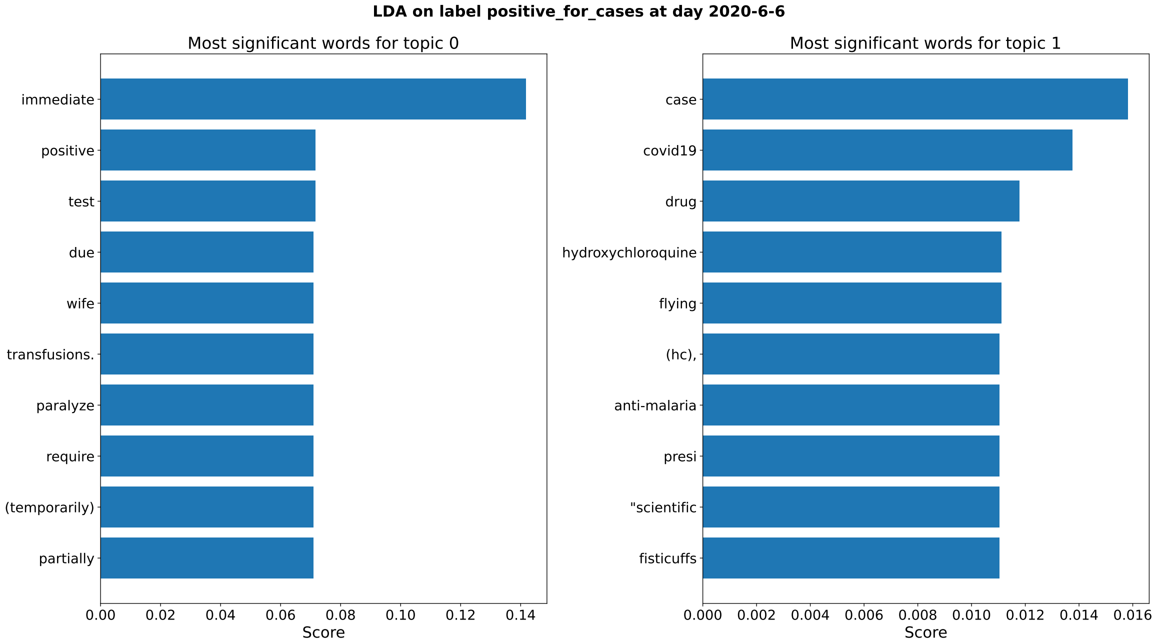

Figure 24: LDA on tweets with the label "positive for cases" (June 6, 2020)

Topic	Text	TweetID	UserID	Retweets	Suspended RTs
0	Wife tested positive for GBS, is partially paralyzed (temporarily) and requires immediate immediate transfusions. Due to BS	1269068178128322562	19091173	8725	2246

Table 7: Topics to tweets from figure 24

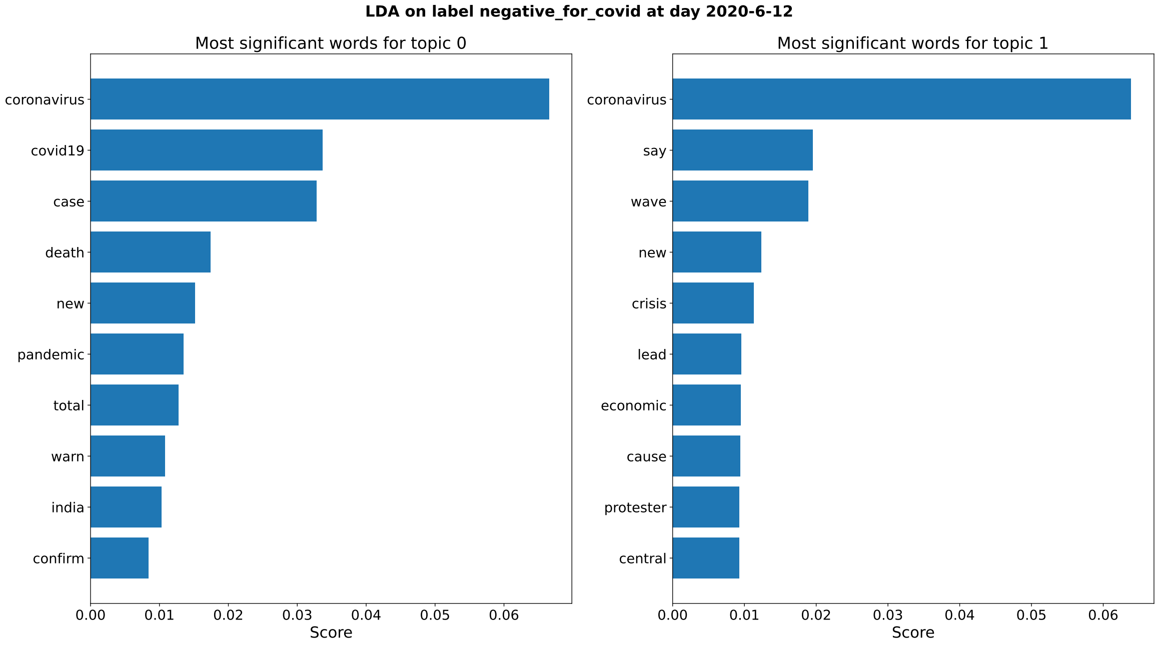

Figure 25: LDA on tweets with the label "negative for covid" (June 12, 2020)

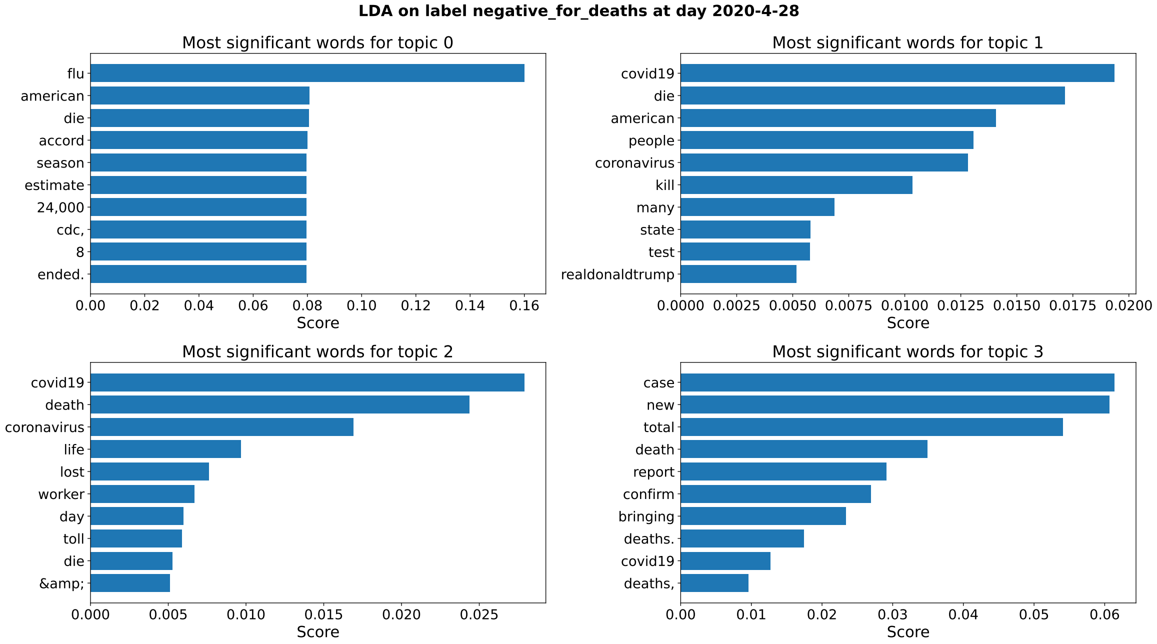

Figure 26: LDA on tweets with the label "negative for deaths" (April 28, 2020)

Topic	Text	TweetID	UserID	Retweets	Suspended RTs
0	Flu season has ended. According to the CDC, an estimated 24,000 Americans have died from the flu this season—down from 8	1255150245266067458	878247600096509952	21740	6217

Table 8: Topics to tweets from figure 26

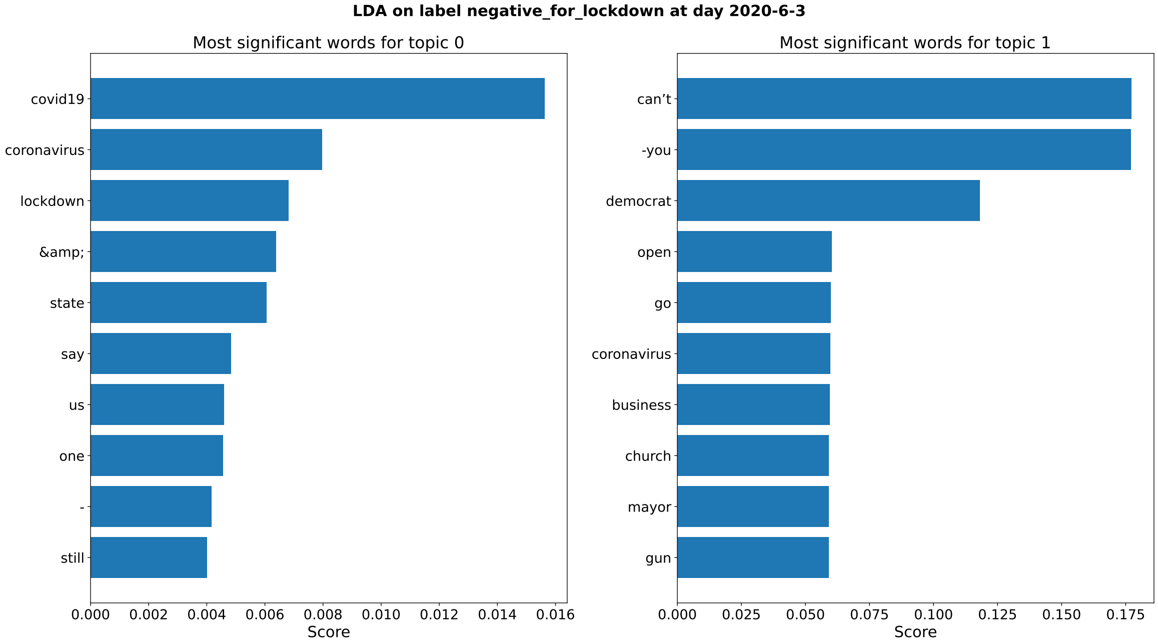

Figure 27: LDA on tweets with the label "negative for lockdown" (June 3, 2020)

Topic	Text	TweetID	UserID	Retweets	Suspended RTs
1	Democrat mayors during coronavirus -You can't open your business -You can't go to church -You can't buy a gun Democrat	1268180505725370368	18166778	29280	7957

Table 9: Topics to tweets from figure 27

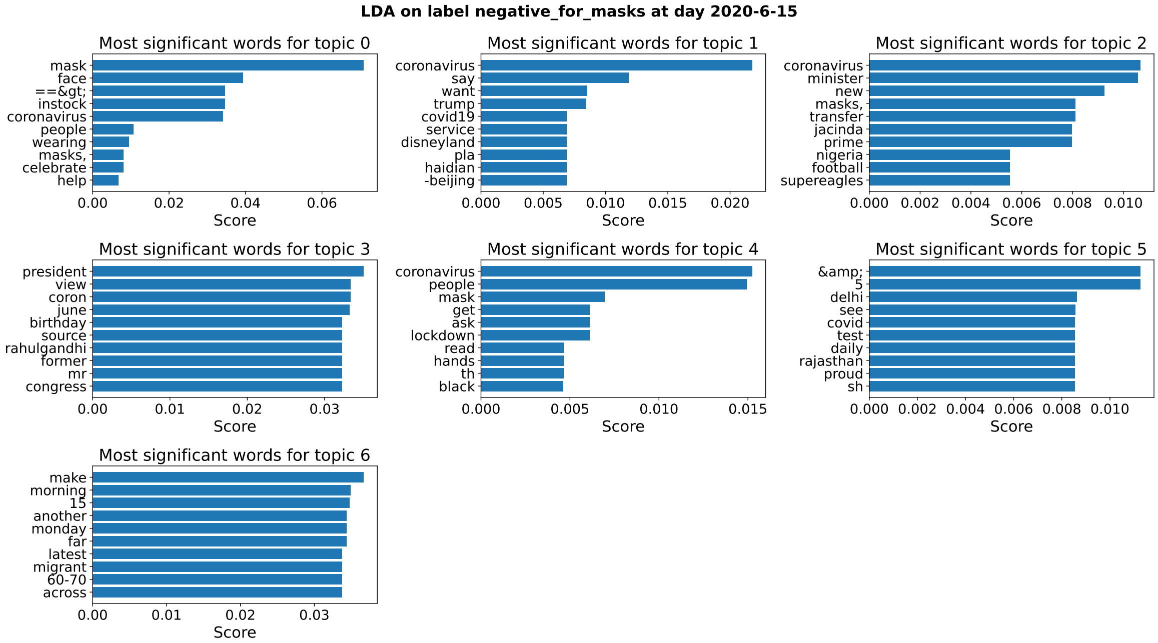

Figure 28: LDA on tweets with the label "negative for masks" (June 15, 2020)

Topic	Text	TweetID	UserID	Retweets	Suspended RTs
3	Sources Former Congress President Mr RahulGandhi decides not to celebrate his birthday on June 19th given Coron	1272502930487181312	72283791	414	34
6	DOVER LATEST MONDAY Another 60-70 Illegal Migrants have made it across the Channel so far this morning At least 15	1272484907869241345	338283123	207	58

Table 10: Topics to tweets from figure 28