Capstone Project - CIND820

Final Project Report

Ryerson University

Identifying Trends and Sentiments in Tweets

Name: Pallavi Thirunavukarasu

Student Number: 501114919

Supervisor: Dr Ceni Babaoglu

Date: April 4, 2022

Table of Contents

Introduction

Social media has become a rich corpus of text and a communication channel that reflects the events happening in the community and opinions/reviews of products. It has become important to identify the trends in social media. Trends may be broadly described as the top noteworthy topics that are discussed.

Sentiment detection and classification is also of great importance in the analytics of social media data. There are several practical applications for identifying trends and sentiments in social media data such as in recognising consumer sentiments towards a product, understanding public opinion on government policies, financial predicting, etc.

Leveraging social media content presents multiple challenges in interpreting the data due to the presence non- standard words(slang), use of symbols, abbreviations and emoticons.

Substantial amount of research work has been done in this area and there is ongoing research to address the unique challenges of mining social media data.

Abstract

The purpose of this project is to apply Text Mining and Sentiment Analysis using Natural Language Processing techniques on Twitter data:

- Identify the top topics/context within the tweets
- Group the tweets based on the topics
- Classify the tweets by the sentiment they express within each topic
- Visualize the results of the sentiments expressed by topic, sentiment

Literature Review

Sentiment Analysis:

The basic idea of sentiment analysis is to take a text and classify that text as positive, negative or neutral. Sentiment mining can be done at different levels [1].

- Document level sentiment classification: This level classifies the entire document as positive, negative or neutral
- Sentence level sentiment classification: This level classifies each sentence as positive, negative or neutral
- Aspect or Entity level sentiment classification: This level identifies the aspect in each sentence and then classifies the aspect as positive, negative or neutral.

Liu and Zhang **[2]** defined an opinion as a quintuple, consisting of an entity, aspect of the entity, orientation(sentiment) of the opinion of the user about that aspect, the user and the time. For example, if the following review was posted by user user1, on 14.02.2022,

> The camera quality of my new iPhone13 is great.

The entity for which the opinion is expressed is the iPhone 13, the aspect is the camera, the sentiment is positive, the user is user1 and the time is 14.02.2022. The quintuple would be (iPhone13, camera, positive, user1, 14.02.2022). All the five components are considered as essential to sentiment analysis.

Given a collection of opinionated documents, there are 5 steps to find all the quintuples(opinions) in the document.

- Extracting all the entities in the document
- Extracting all the aspects of the entities
- Extracting the opinion holder and the time
- Determining if the opinion of the aspect is positive, negative or neutral
- Producing all the possible quintuples for this document

Immense quantities of user created web based and social networking content has several applications. Some examples include:

- Stock prediction – in this study **[3]** conducted in 2007, stock discussions on message boards were collected using web scraping programs, various classifier algorithms were used to classify the sentiments for the messages, a voting mechanism was used on the classifier outcomes to determine the sentiment index. Relationship of the sentiment index to the stock values was analysed.
- Government Policy making – in this study **[4],** conducted in 2015, sentiment analysis was used to determine how opinion of local government social media posts influences citizen involvement on Twitter. Relationship of positive sentiment expressed on social media to citizen's digital participation was studied
- Political results – in this study **[5],** conducted in 2010, text analysis software was used to conduct a sentiment analysis of 100,000 messages which mentioned a politician or political party. An analysis of the tweets political sentiments showed that public sentiment is close to the political positions of the parties and politicians.

Twitter Sentiment Analysis (TSA)

Twitter is considered as one of the most popular microblogging sites. As of Q4 of 2020, twitter has 192M monetizable daily active usage*. Each tweet is a single message posted on twitter that can be 280 characters long. A user can register with the platform to post the tweets. Each user can follow other users. Retweets are tweets that are re-distributed by other users.

Twitter is considered as one of the largest datasets of user generated content due to the large number of users, easy access to data/downloading of published posts.

Twitter is a dynamic forum with continuous updates and also has unique challenges, hence a novel approach is required for Twitter sentiment analysis.

Challenges:

Some of the challenges associated with Twitter data sentiment analysis are discussed in detail in this survey **[6]**.

1. Text Length – The short length of the twitter messages makes the sentiment analysis very different from the sentiment analysis of blogs, movie reviews or product reviews. Bermingham and Smeaton conducted a study **[7]**, to check if the brevity of text was an advantage by comparing Support Vector Machine (SVM) and Multinomial Naïve Bayes. They concluded that SVM performed better in Twitter Sentiment Analysis.
2. Incorrect English – Tweets have peculiarities of use of slang, emphatic lengthening, emphatic uppercasing, abbreviations. Preprocessing and cleanup of these inaccuracies is required
3. Topic Relevance – To classify tweets and to take into account the topic relevance of tweets, the keywords and hash tags in the tweets are considered
4. Data Sparsity - Tweets contain a lot of noise due to misspellings, slang and this results in data sparsity.
5. Stop Words – Stop words that are normally filtered out in text processing such as 'like' may usually have discriminatory powers when it comes to tweets. Saif et al. [2014] **[8]** work focussed on comparing the different methods of stop word removal on 4 different datasets and analysing the impact on the classifiers and the data sparsity.
6. Tokenization – Tokenization by white spaces may not work well for Twitter. Some studies have developed twitter specific tokenization.
7. Multilingual content – Tweets can contain mixed languages even in a single tweet, tweets can be done in several languages. Multi lingual classifiers have been developed for this.
8. Multimodal – Tweets can also videos, images. Extracting features from multimodal content is ongoing research.

Methods for Twitter Sentiment Analysis:

There are several approaches to Twitter Sentiment Analysis – Supervised Machine Learning methods, Ensemble methods, Lexicon Based, Hybrid and Deep Learning based.

Supervised Learning – Twitter data is easily available and also in large quantities, yet there is a challenge in getting labelled data for sentiment analysis which makes the use of Supervised Learning methods a challenge. In one of the early Twitter Sentiment Analysis done by Go et al., 2019**, [9]** the approach was to collect the tweets with emoticons, then use the emoticons as a kind of noisy labels. The labelled tweets where then stripped of the emoticons and different features like unigrams, unigrams and bigrams, unigrams with part of speech tagging were built. The labelled data was now used to built classifier models using Support Vector Machine, Naïve Bayes and Max Entropy algorithms.

Ensemble Methods – In this study [**10**] , the mined tweets were subjected to an ensemble classification method. Multiple classifiers were trained to obtain better predictive power. The results of the different classifiers were averaged to obtain the sentiment predictions

Lexicon based – In this study **[11],** tweets were collected using keywords, lexicon or dictionary-based approach was used to classify the sentiments. The words from the tweet were matched with the words from the dictionary. If the dictionary word was positive the tweet word was tagged positive. If the dictionary word was negative the tweet word was tagged negative. If the dictionary word was not positive or negative, the tweet word was tagged as neutral. A sentiment score was calculated to classify the tweet. Sentiment mining was also done at the aspect level.

Hybrid methods – Balage et al [**12]**, built a hybrid system for tweet classification. The system was built with 4 main components. A normalization component that was used for correcting and normalizing the collected tweets. The three other components were created as a pipeline - a rule-based classifier, a lexicon-based classifier and a machine learning based classifier. The tweet was passed through each component in the pipeline sequentially until the tweet was classified with a certain confidence. This hybrid pipeline of classifiers was found to give more accuracy than the Support Vector Machine classifier.

Deep learning based – Several deep learning-based approaches to sentiment analysis are underway now.

Feature Selection for Twitter Sentiment Analysis:

Features such as automatic part-of-speech tags and resources such as sentiment lexicons are useful for sentiment analysis in other domains. However, feature selection for microblogging sites like twitter is not a trivial task.

Kouloumpis et al, 2011 **[13],** conducted a study to check the utility of linguistic features for determining sentiments in microblogging sites such as twitter. They used three datasets in this study. The Hash Tagged dataset HASH, is a subset of the Edinburgh twitter corpus. The top 15 hashtags were first identified from the Edinburg twitter corpus, then the hashtags with at least 1000 tweets were taken. Emoticon dataset EMOT, was created by Go, Bhayani, and Huang for a project at Stanford University by collecting tweets with positive and negative emoticons. The iSieve dataset contains a set of 4000 tweets which were collected and annotated by the iSieve corporation.

The data was preprocessed – tokenization, normalization, part of speech tagging was done. Abbreviations and emoticons were converted to individual tokens. Intensifiers such as upper case were identified and converted to lowercase. Emphatic lengthening was identified and repeated characters were replaced with single characters.

Variety of features like unigrams, bigrams, part of speech features, lexicon-based features were used in the classification. Microblogging domain specific features were also included. Accuracy of the classifiers for different feature combinations was analyzed. Part of Speech features were found to have a negative impact on the twitter sentiment analysis.

Agarwal et al [14**]** [2011], conducted a study in which they compared different combinations of feature sets. The dataset consisted of manually annotated twitter data that was collected. Two models were built. One was a binary class classifier that classified the tweets as positive or negative. The second was a multi class classifier that classified the tweets as positive, negative or neutral. Three types of models were built, a unigram model, a feature-based model and a tree-based model. Using the unigram model as baseline, the study compared combination of models with different combinations of features. The combinations were found to outperform the baseline unigram model. Contrary to the previous study's finding, this study found that part of speech tagging increased accuracy. Twitter specific features did not impact the outcome.

Topic modelling and Sentiment Analysis on Twitter Data

In the past tweets have been analyzed at the time of disasters and spread of diseases. For example, this study on Zika outbreak**[17]** used twitter data, analyzed how timely topics like the Zika virus are addressed on social media. The study examined the emerging themes during a Center for Disease Control (CDC) hosted live Twitter chat and used text mining to evaluate the public's concerns about the Zika virus and the CDC's response to the public's questions.

Similar studies have been conducted at the time of Ebola outbreak **[18],** Hurricane Irma **[19]** and Japanese earthquake of 2011**[20]**.

In the present pandemic situation, many studies have been conducted using twitter data to gauge public opinions and identify trends.

In this study **[15]** , Using the Twitter API and tools, tweets which contained a predefined set of keywords related to covid19 were downloaded from twitter. The top topics in these tweets was identified using the Latent Dirichlet Allocation (LDA). Topics were grouped together as themes manually. The sentiment analysis was done. The mean sentiment was found to be positive for ten topics and negative for two topics.

In this study **[16] ,** Data from Twitter was collected using a streaming application over a period of time. Tweets were collected with keywords related to covid. The data was cleaned and preprocessed (removal of stop words, removal of URLs and hyperlinks, etc.). The word cloud of frequently used words was created. Sentiment analysis was done using the National Research Council (NRC) lexicon, Total of 8 emotions were evaluated based on the lexicon. Latent Dirichlet Allocation (LDA) was applied to fit a topic model. Coherence was used as an evaluative measure to choose the best number of topics for the topic model. The tweet trends and results of the sentiment analysis were visualized.

Project Methodology

The purpose of this project is to apply Text Mining and Sentiment Analysis using Natural Language Processing techniques on tweets collected from Twitter with the covid19 hashtag. The purpose of this project is to propose a framework for intelligent analysis of the covid related Twitter data by extracting the semantic topics and the sentiments along those topics and presenting the results in an easy-to-follow visual format.

The following is the methodology for the project.

1. The Twitter API will be used to download the English language tweets with the covid19 and related hashtags.
2. The dataset will be cleaned and preprocessing techniques will be applied.
3. Using the topic modelling approach of Latent Dirichlet Allocation (LDA), the top topics from the tweet dataset will be identified
4. The tweets will be grouped into clusters based on the topics identified in step 3
5. The feature selection for sentiment analysis will be done
6. The sentiments expressed in the tweets within each topic cluster will be identified
7. The opinion impact of the sentiments expressed by topic, sentiment will be visualized effectively

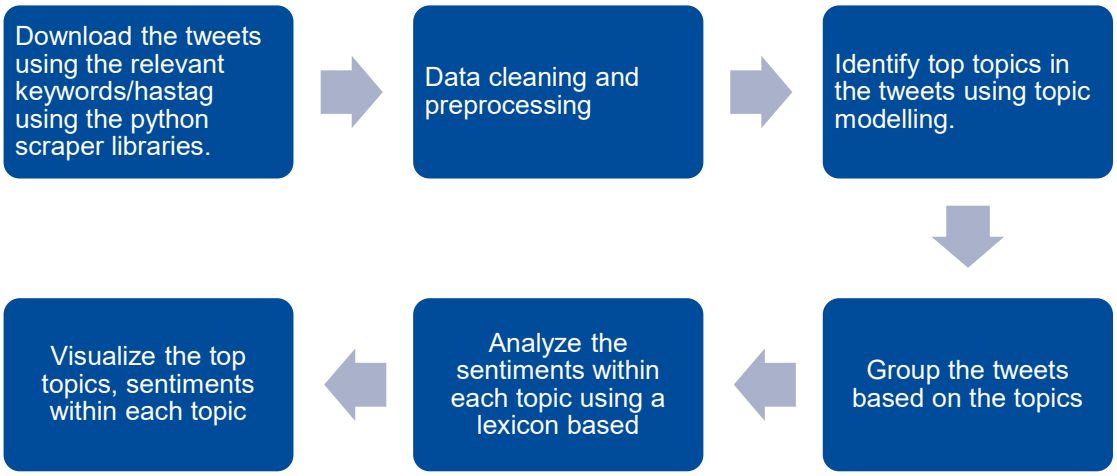


Figure 1: Project Methodology

Data description

The dataset used for the project consists of Tweets scraped using the python twitter scraper libraries.

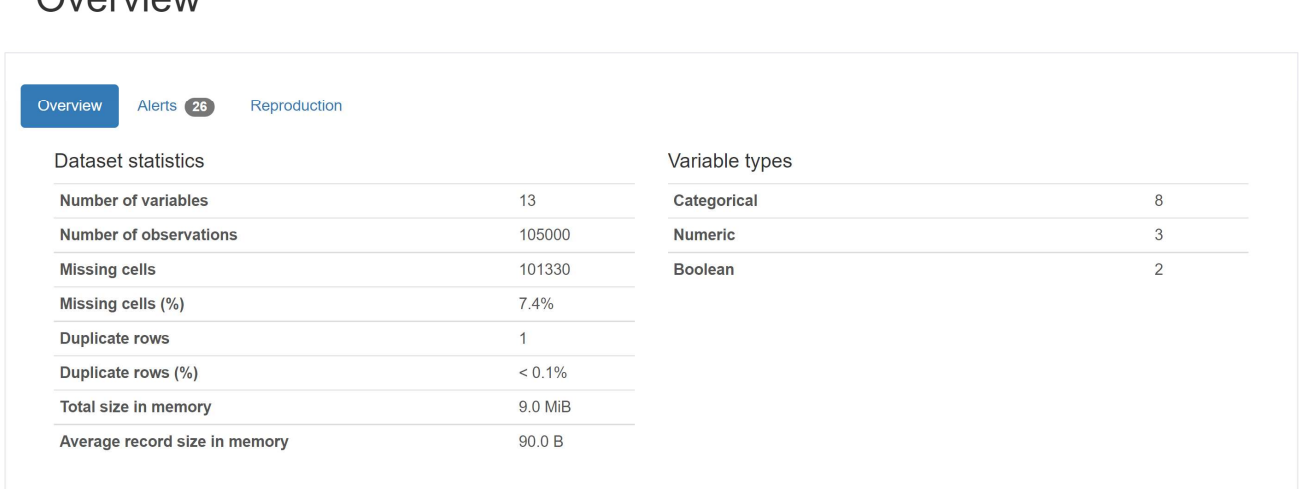
This dataset consists of tweets in English covering dates from 2022-03-08 to 2022-03-14. 15,000 tweets from each day were scraped.

The tweets were downloaded from no specific region and are global. Retweets are not included in this dataset.

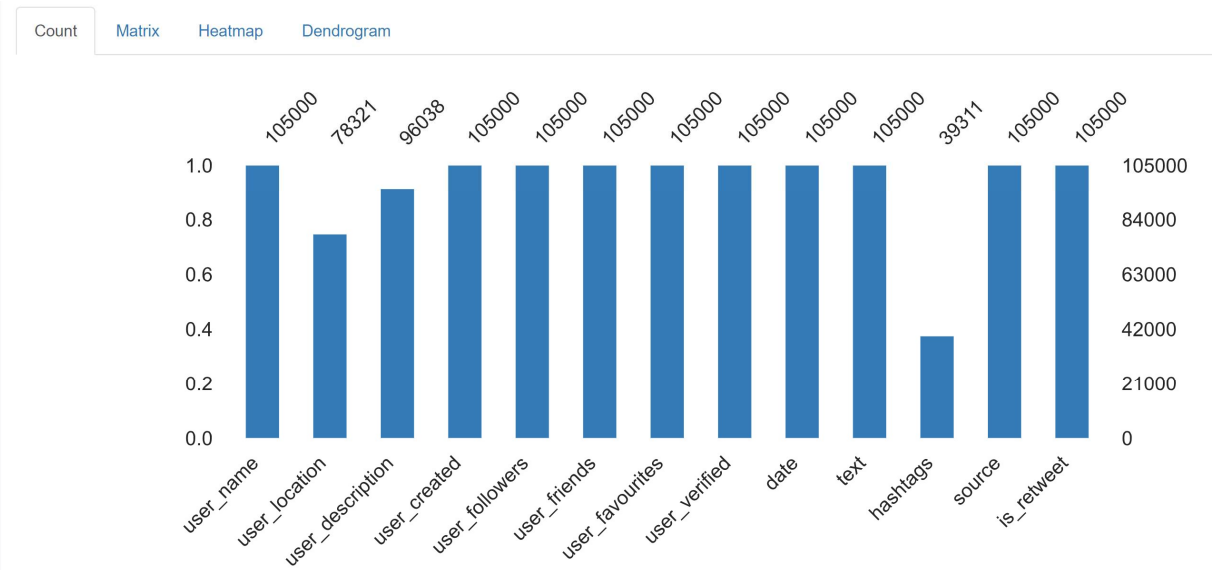
There are 13 columns in this dataset:

1. user_name – Name of the user on twitter
2. user_location - Location of the user
3. user_description – Description of the user on twitter
4. user_created – When the user was created
5. user_followers – Number of followers of this user
6. user_friends – Number of friends of this user
7. user_favourites – Number of favourites of this user
8. user_verified – Is the user verified
9. date – Tweet date
10. text -Text of the tweet
11. hashtags – List of hash tags
12. source – Source of the tweet
13. is_retweet – Tweet is a retweet

Overview of the Dataset



Missing Values



Variables

user_name
Categorical

HIGH CARDINALITY

Distinct	59974
Distinct (%)	57.1%
Missing	0
Missing (%)	0.0%
Memory size	820.4 KiB

corona19_stats	727
kcvaccinewatch	401
CanNews24	182
VippusaO	166
Victori44685362	149
Other values (59969)	103375

Toggle details

user_location
Categorical

HIGH CARDINALITY
MISSING

Distinct	18558
Distinct (%)	23.7%
Missing	26679
Missing (%)	25.4%
Memory size	820.4 KiB

United States	1510
Washington, DC	1227
Canada	1171
USA	965
New York, NY	890
Other values (18553)	72558

user_description

Categorical

HIGH CARDINALITY
MISSING

Distinct	53932
Distinct (%)	56.2%
Missing	8962
Missing (%)	8.5%
Memory size	820.4 KiB

Author: @comster	727
Notifying you when COVID-19 test appo…	401
Canada's #1 news aggregator. LIVE 24/…	182
NON-VOTERS ARE UNDEFEATED. T…	166
A bot designed to put the #COVID19 pa…	141
Other values (53927)	94421

Toggle details

user_created

Categorical

HIGH CARDINALITY

Distinct	59962
Distinct (%)	57.1%
Missing	0
Missing (%)	0.0%
Memory size	820.4 KiB

2020-09-22 10:11:34+00:00	727
2021-03-02 16:59:23+00:00	401
2020-06-20 22:29:52+00:00	182
2020-02-07 14:49:27+00:00	166
2021-01-31 22:03:05+00:00	149
Other values (59957)	103375

Toggle details

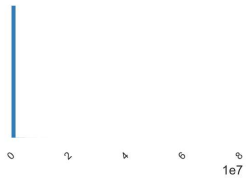
user_followers

Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
SKEWED
ZEROS

Distinct	20313
Distinct (%)	19.3%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	95439.04061

Minimum	0
Maximum	77306922
Zeros	1202
Zeros (%)	1.1%
Negative	0
Negative (%)	0.0%
Memory size	820.4 KiB



Toggle details

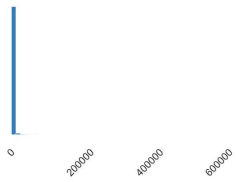
user_friends

Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
SKEWED
ZEROS

Distinct	8124
Distinct (%)	7.7%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%

Minimum	0
Maximum	642044
Zeros	2317
Zeros (%)	2.2%
Negative	0
Negative (%)	0.0%



user_favourites

Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
ZEROS

Distinct	26665
Distinct (%)	25.4%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	18519.08364

Minimum	0
Maximum	2951973
Zeros	5539
Zeros (%)	5.3%
Negative	0
Negative (%)	0.0%
Memory size	820.4 KiB

Toggle details

user_verified

Boolean

HIGH CORRELATION
HIGH CORRELATION

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	102.7 KiB

date

Categorical

HIGH CARDINALITY
UNIFORM

Distinct	81082
Distinct (%)	77.2%
Missing	0
Missing (%)	0.0%
Memory size	820.4 KiB

2022-03-09 15:40:03+00:00	24
2022-03-10 22:32:04+00:00	23
2022-03-10 16:18:04+00:00	22
2022-03-10 23:17:04+00:00	17
2022-03-11 17:00:02+00:00	17
Other values (81077)	104897

Toggle details

text

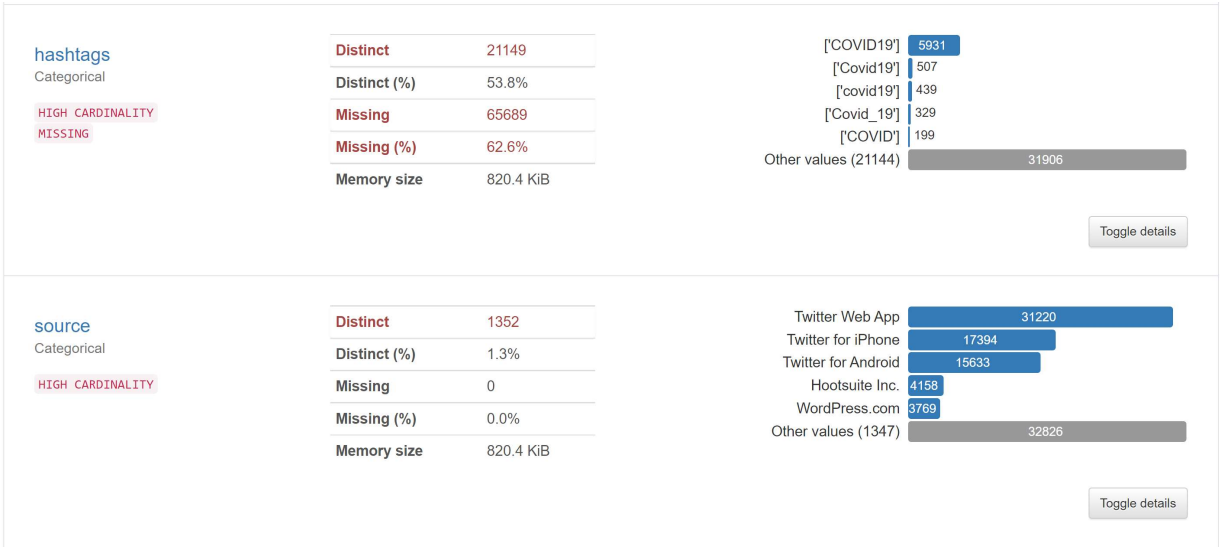
Categorical

HIGH CARDINALITY
UNIFORM

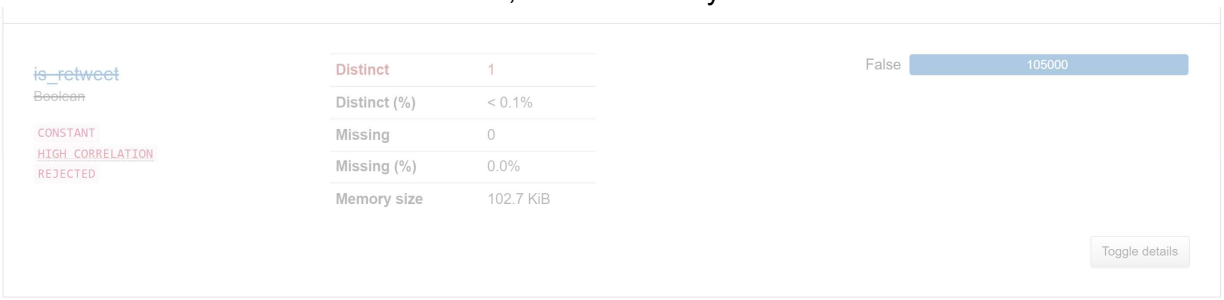
Distinct	103740
Distinct (%)	98.8%
Missing	0
Missing (%)	0.0%
Memory size	820.4 KiB

With news that @BootsUK is selling lat…	16
#ResettleEritreanRefugees from Afar&a…	10
The person below hates COVID-19.	8
After COVID-19, Optical Heights celebr…	8
KS: Rapid Test appointments available …	7
Other values (103735)	104951

Toggle details



Retweets are removed from the dataset, so there is only one distinct value in this column.



Data Visualization and Summarization

The first few records in the dataset:

	user_name	user_location	user_description	user_created	user_followers	user_friends	user_favourites	user_verified	date	text	hashtags	source	is_retweet
0	needlenose5	Milford, CT	Husband, dad to kids. Pro Trump. GPU head, Sub...	2020-03-25 21:23:13+00:00	12	86	1089	False	2022-03-07 23:59:59+00:00	@MaryLTrump @DavidCornDC Agreed, why do people...	NaN	Twitter for iPhone	False
1	workonline44	NaN	NaN	2016-07-30 03:40:00+00:00	30136	1915	62	False	2022-03-07 23:59:56+00:00	AD iHealth™ Covid-19 Antigen testing kits for ...	NaN	Twitter Web App	False
2	Uab_BabaKofi	Virginia, USA	Dad of two daughters, husband, Consulting IT S...	2016-03-05 21:10:37+00:00	9874	10842	41264	False	2022-03-07 23:59:54+00:00	Study links even mild Covid-19 to changes in t...	['DemForce']	Twitter for iPad	False
3	ActivistBowen2	Hong Kong	AMERICAN EXPAT in HONG KONG, person, blogger, ...	2011-06-08 20:17:01+00:00	4168	4997	5217	False	2022-03-07 23:59:36+00:00	,...a NON-#covid19 death ON TOP OF ALL the #co...	['covid19', 'covid19', 'HONGKONG', 'HK']	Twitter Web App	False
4	RAChampion	Sydney Australia	Digital Marketing, Brand Promotion & Social Me...	2012-06-12 12:28:09+00:00	2472	1519	4	False	2022-03-07 23:59:33+00:00	SFC urges Hong Kong businesses to update conti...	NaN	dlvr.it	False

The last few records in the dataset:

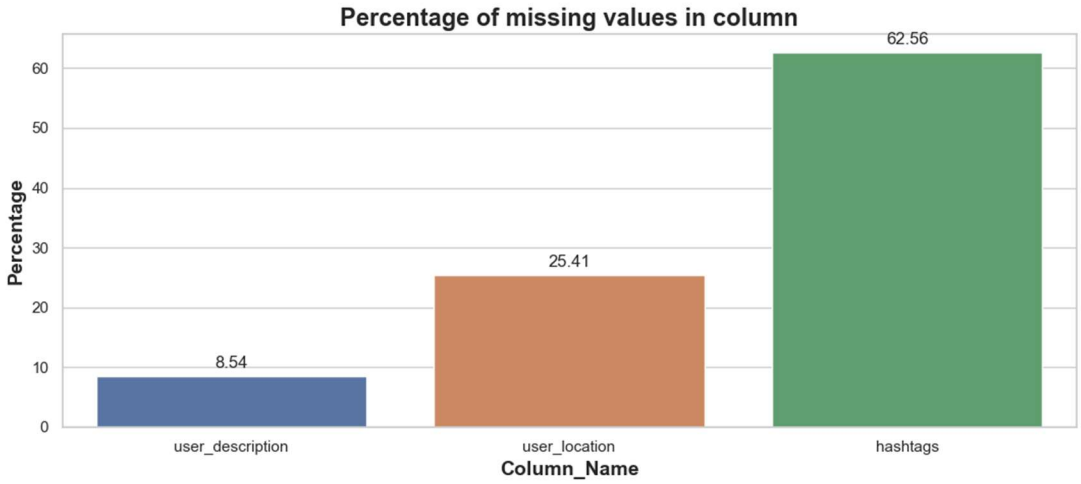
	user_name	user_location	user_description	user_created	user_followers	user_friends	user_favourites	user_verified	date	text	hashtags	source	is_retweet
104995	iTPcCares	New Delhi, India	iTPcCares is enabling the digital world with i...	2020-05-03 11:44:54+00:00	25	323	29	False	2022-03-13 10:57:17+00:00	What do you think about the 4th wave of covid-...	['4th_wave_in_India', '4th_wave_of_covid_in_in...	Twitter Web App	False
104996	MarilsaAPC	NaN	NaN	2018-09-13 20:26:14+00:00	1404	5011	120048	False	2022-03-13 10:57:17+00:00	Cervical longitudinally extensive myelitis aft...	['CovidVaccination', 'VaccineSideEffects', 'CO...	Twitter for Android	False
104997	lotuseatersnews	NaN	NaN	2022-02-14 16:57:57+00:00	1148	9	2	False	2022-03-13 10:57:15+00:00	📣 Austria suspends mandatory Covid-19 vaccinat...	NaN	Twitter Web App	False
104998	FrankieEAGB	Reality	NaN	2021-09-28 18:45:52+00:00	257	1957	8647	False	2022-03-13 10:56:55+00:00	@JohnSty83584062 God forbid that any site shou...	NaN	Twitter Web App	False
104999	56sahara	NaN	Direita \nConservador \nBolsonariano	2021-08-06 10:43:15+00:00	2652	4265	12179	False	2022-03-13 10:56:51+00:00	@carlyra @ArlineCasagran1 Covid 19 \n\n#ArmaBi...	['ArmaBiologicaChinesa', 'ArmaBiologicaAmerica...	Twitter for Android	False

Summary of the numerical values in the dataset:

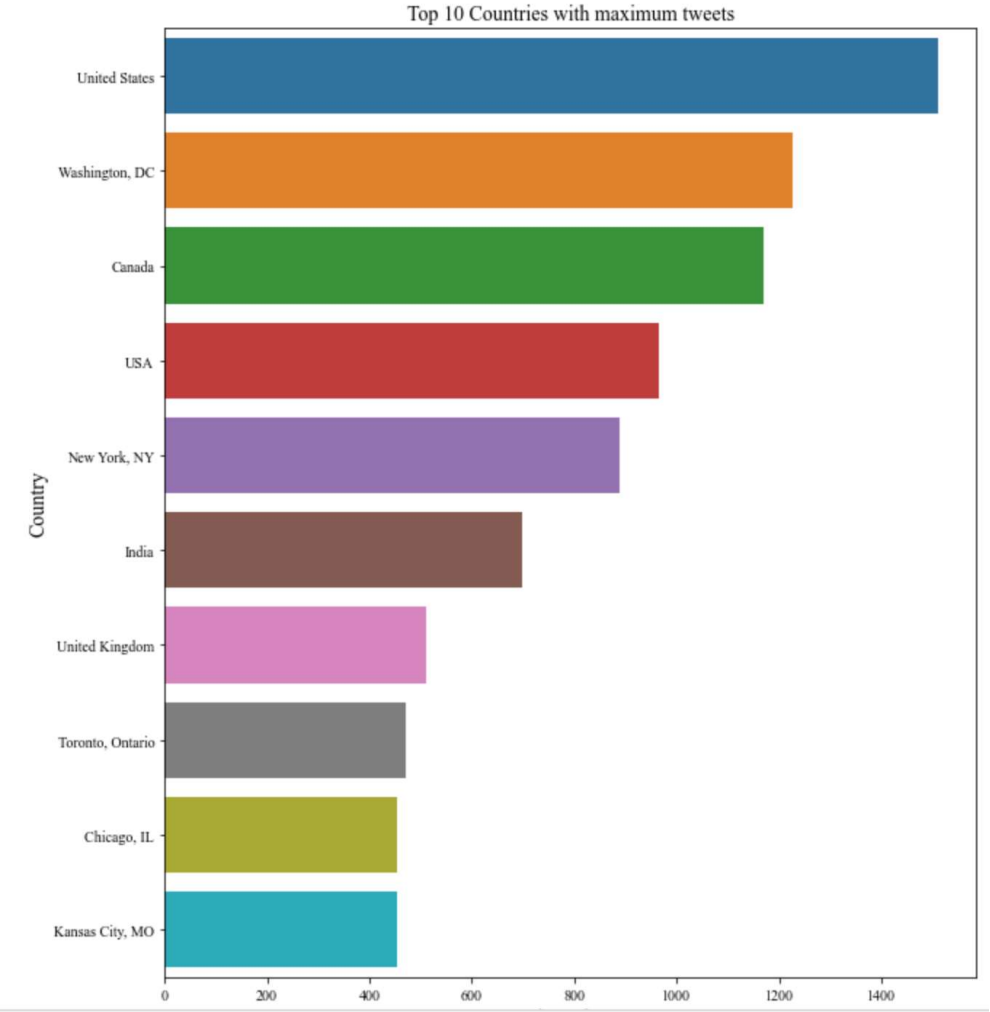
	user_followers	user_friends	user_favourites
count	105000.000000	105000.000000	105000.000000
mean	95439.040610	1997.810248	18519.083638
std	1152952.882198	9549.526646	60057.347281
min	0.000000	0.000000	0.000000
25%	160.000000	132.000000	151.000000
50%	1038.000000	543.000000	1964.000000
75%	5972.000000	1673.000000	11413.250000
max	77306922.000000	642044.000000	2951973.000000

Percentage of missing values by columns:

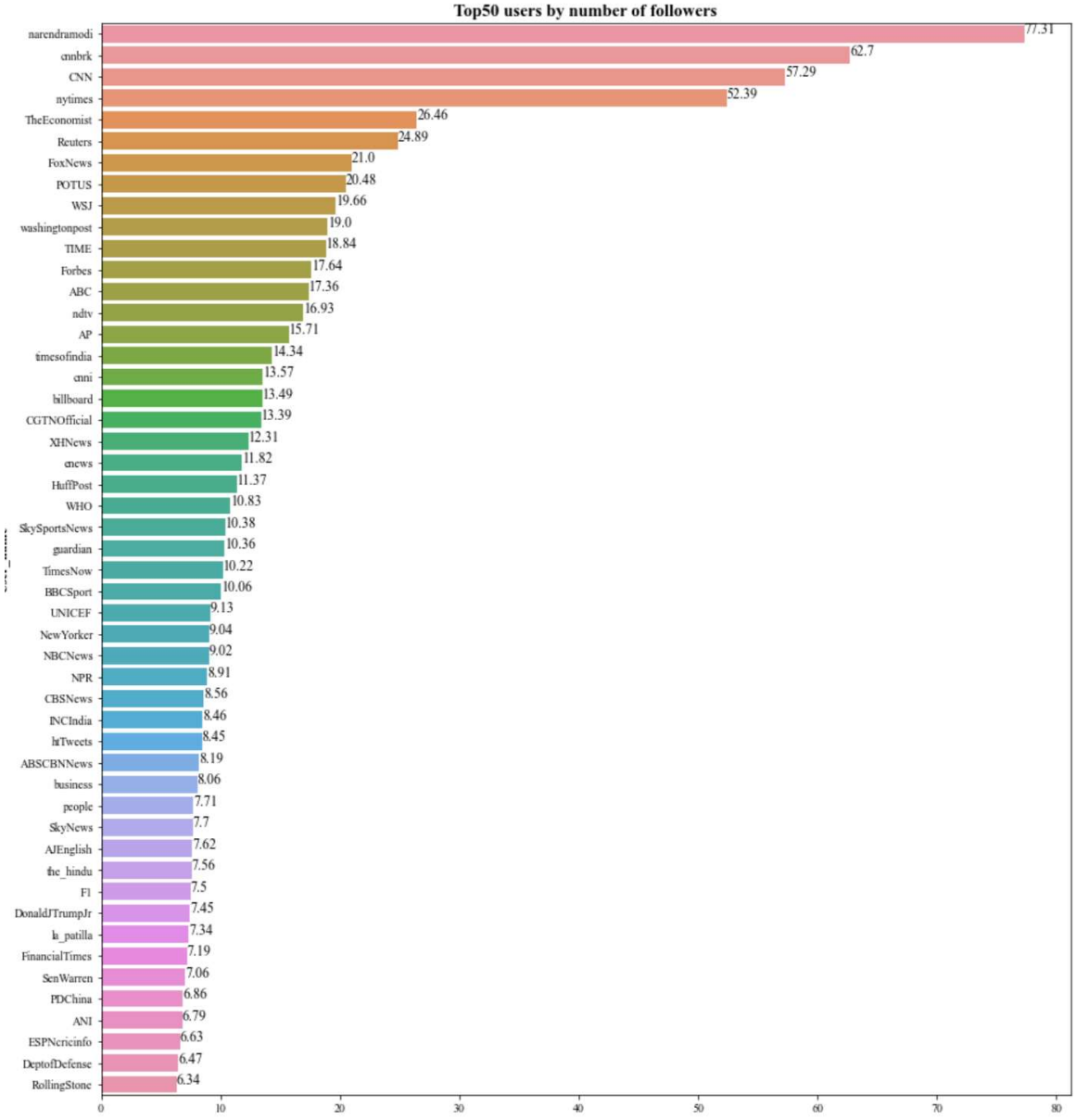
The hashtags have the maximum number of missing values.



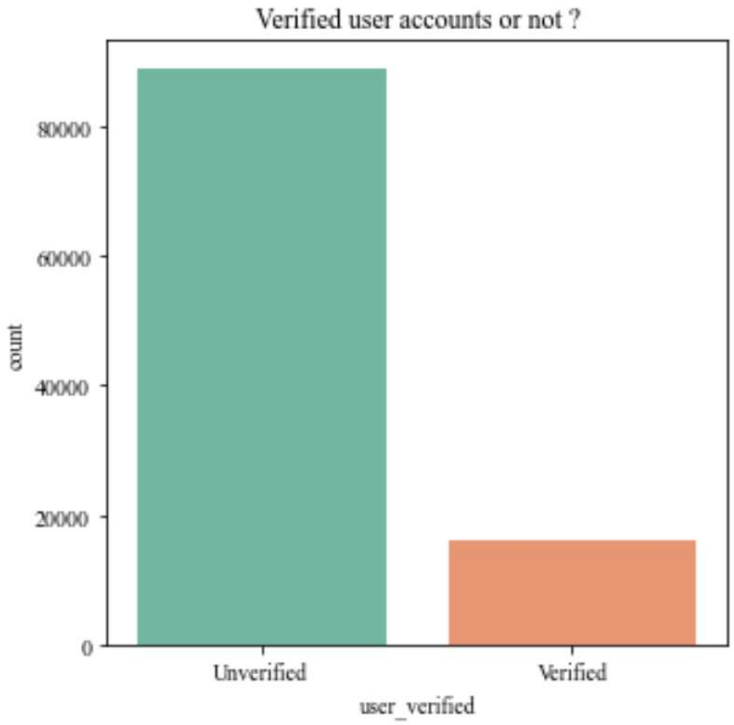
Top 10 countries by number of tweets:



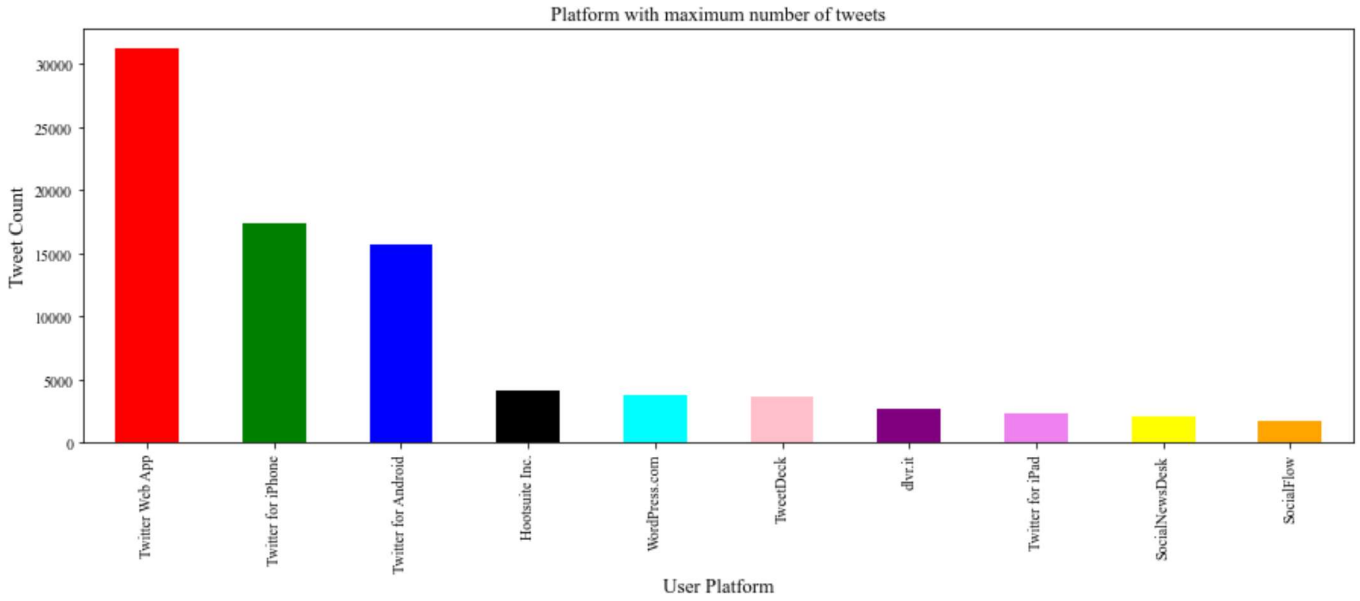
The top 50 users by number of followers:



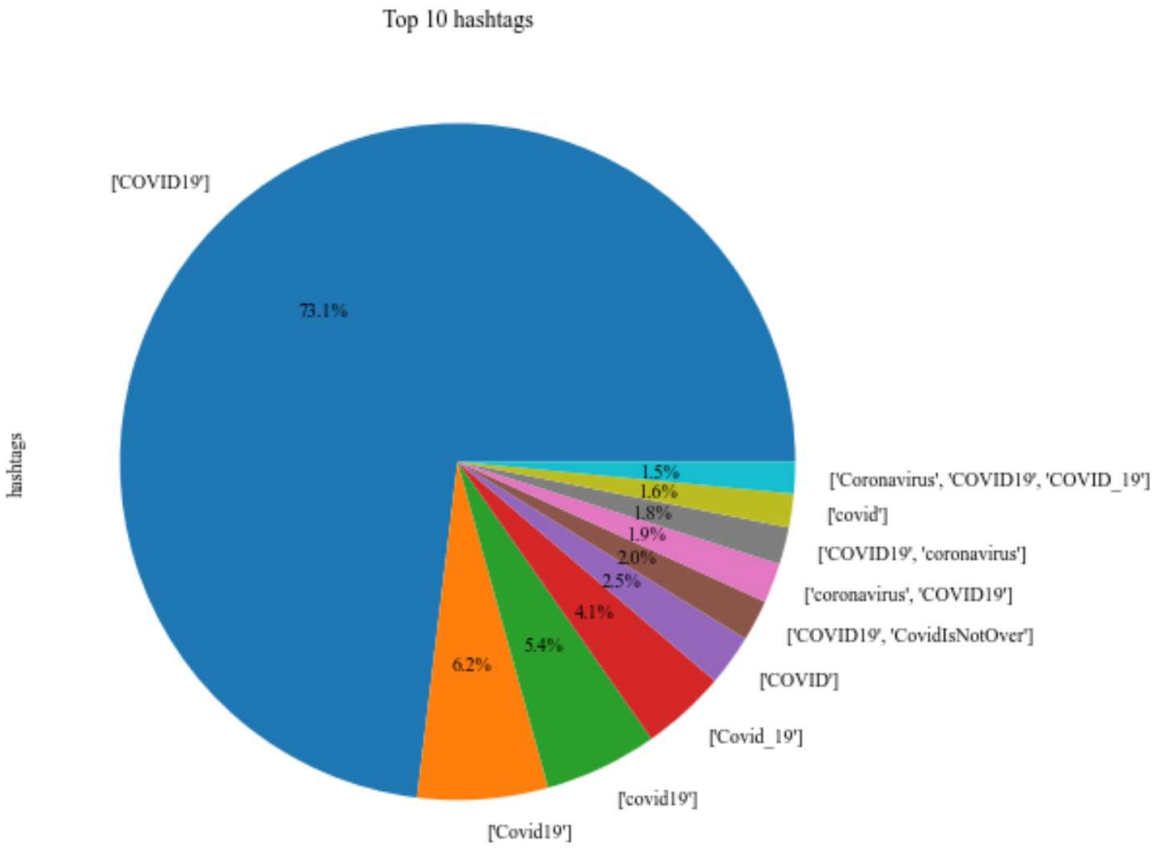
The number verified vs non verified users:



Platform with the maximum number of tweets:



Top ten Hashtags in the data:



Data Cleaning and Processing

The raw tweets consist of hashtags, @ mentions, special characters, emojis and URLs. The following steps have been done to clean the text column of the dataset. The cleaned text is added to the dataset as the clean_text column.

1. Each tweet is split into individual words
2. The words are converted to lower case
3. Contractions are replaced with their full form. For example, didn't is replaced with did not.
4. URLs are removed
5. @ Mentions are removed
6. Special characters are removed
7. Hashtags are removed

The text after the first few steps is completed:

	clean_text
0	agreed, why do people keep on supporting democrats??? from the covid-19 lies to the fake climate crisis to the fake jan 6th insurrection to the fake inflation caused by greedy corporations to the …
1	ad ihealth™ covid-19 antigen testing kits for sale
2	study links even mild covid-19 to changes in the brain - cnn
3	,...a non- death on top of all the deaths (march 7, 2022; 22:17 hkt)
4	sfc urges hong kong businesses to update continuity plans as mass covid-19 testing looms in the city
5	you've taken a few too many hits on the crack pipe. our natural immunity works better than any vaccine. my whole family had covid 19- not hospitalized! we are healthy people. no vaccine required. god bless
6	a look at covid medical waste and how to handle it \n\n
7	….information from the platforms about the major sources of covid-19 misinformation, including those that engaged in the sale of unproven covid-19 products, services and treatments.\n\n lol that would be you!
8	uncontrolled spread of can lead to the creation of a worse variant.\n\nus seems to be setting up perfect conditions for a worse variant that won't just harm our country but will harm the world.\n\nit's shameful we w...
9	does anyone know if ihealth covid-19 antigen rapid tests from usps are approved for travel back to u.s.? I am getting conflicting answers online. technically the test should be acceptable when proctored online since ...

8. The sentences are tokenized into words
9. Punctuations are removed

```
['agreed', 'why', 'do', 'people', 'keep', 'on', 'supporting', 'democrats', 'from', 'the', 'covid', 'lies', 'to', 'the', 'fake', 'climate', 'crisis', 'to', 'th
e', 'fake', 'jan', 'th', 'insurrection', 'to', 'the', 'fake', 'inflation', 'caused', 'by', 'greedy', 'corporations', 'to', 'the']
['ad', 'ihealth', 'covid', 'antigen', 'testing', 'kits', 'for', 'sale']
['study', 'links', 'even', 'mild', 'covid', 'to', 'changes', 'in', 'the', 'brain', 'cnn']
['non', 'death', 'on', 'top', 'of', 'all', 'the', 'deaths', 'march', 'hkt']
['sfc', 'urges', 'hong', 'kong', 'businesses', 'to', 'update', 'continuity', 'plans', 'as', 'mass', 'covid', 'testing', 'looms', 'in', 'the', 'city']
['you', 've', 'taken', 'few', 'too', 'many', 'hits', 'on', 'the', 'crack', 'pipe', 'our', 'natural', 'immunity', 'works', 'better', 'than', 'any', 'vaccine',
'my', 'whole', 'family', 'had', 'covid', 'not', 'hospitalized', 'we', 'are', 'healthy', 'people', 'no', 'vaccine', 'required', 'god', 'bless']
['look', 'at', 'covid', 'medical', 'waste', 'and', 'how', 'to', 'handle', 'it']
['information', 'from', 'the', 'platforms', 'about', 'the', 'major', 'sources', 'of', 'covid', 'misinformation', 'including', 'those', 'that', 'engaged', 'i
n', 'the', 'sale', 'of', 'unproven', 'covid', 'products', 'services', 'and', 'treatments', 'lol', 'that', 'would', 'be', 'you']
['uncontrolled', 'spread', 'of', 'can', 'lead', 'to', 'the', 'creation', 'of', 'worse', 'variant', 'us', 'seems', 'to', 'be', 'setting', 'up', 'perfect', 'con
ditions', 'for', 'worse', 'variant', 'that', 'won', 'just', 'harm', 'our', 'country', 'but', 'will', 'harm', 'the', 'world', 'it', 'shameful', 'we', 'won', 'e
ven', 'take', 'the', 'simple', 'step', 'to', 'continue', 'masking']
['does', 'anyone', 'know', 'if', 'ihealth', 'covid', 'antigen', 'rapid', 'tests', 'from', 'usps', 'are', 'approved', 'for', 'travel', 'back', 'to', 'am', 'get
ting', 'conflicting', 'answers', 'online', 'technically', 'the', 'test', 'should', 'be', 'acceptable', 'when', 'proctored', 'online', 'since', 'it', 'has', 'i
t', 'is', 'fda', 'approved', 'and', 'antigen', 'but', 'there', 'are', 'reports', 'it', 'has', 'it', 'is', 'not']
['fraud', 'alert', 'covid', 'scams', 'oig']
['global', 'covid', 'deaths', 'surpass', 'million']
```

10. The bigrams, trigrams are created using the tokenized words

```
['agreed', 'people', 'keep', 'supporting', 'democrats', 'covid', 'lies', 'fake', 'climate', 'crisis', 'fake', 'jan', 'th', 'insurrection', 'fake', 'inflation', 'caused', 'greed
y', 'corporations']
['ad_ihealth', 'covid', 'antigen', 'testing', 'kits', 'sale']
['study_links', 'even_mild', 'covid', 'changes', 'brain', 'cnn']
['non', 'death', 'top', 'deaths', 'march', 'hkt']
['sfc_urges', 'update', 'mass', 'covid', 'testing', 'looms', 'city']
['taken', 'many', 'hits', 'crack', 'pipe', 'works', 'better', 'vaccine', 'whole', 'family', 'covid', 'hospitalized', 'healthy', 'people', 'vaccine', 'required', 'god_bless']
['look', 'covid', 'medical', 'waste', 'handle']
['information', 'platforms', 'major', 'sources', 'covid', 'misinformation', 'including', 'engaged', 'sale', 'unproven', 'covid', 'products', 'services', 'treatments', 'lol', 'wou
ld']
['uncontrolled', 'spread', 'lead', 'creation', 'worse', 'variant', 'us', 'seems', 'setting', 'perfect', 'conditions', 'worse', 'variant', 'harm', 'country', 'harm', 'world', 'sha
meful', 'even', 'take', 'simple', 'step', 'continue', 'masking']
['anyone', 'know', 'ihealth', 'covid', 'antigen_rapid', 'tests', 'usps', 'approved', 'travel', 'back', 'getting', 'conflicting', 'answers', 'online', 'technically', 'test', 'acce
ptable', 'proctored', 'online', 'since', 'fda_approved', 'antigen', 'reports']
['fraud', 'alert', 'covid', 'scams', 'oig']
['global', 'covid', 'deaths', 'surpass_million']
['abc', 'life', 'death', 'covid']
['working', 'days', 'visa', 'amp', 'provided', 'current', 'company', 'mandatory', 'also', 'medical', 'insurance', 'covid', 'travailing', 'need', 'show', 'airports', 'authority']
['ted_cruz', 'got', 'million', 'billionaire', 'fracking', 'donors', 'last', 'covid', 'aid', 'report', 'via']
['covid', 'delta', 'three_charts']
['including', 'covid', 'shots', 'else']
['watch', 'live', 'today', 'covid', 'update']
['ron_desantis', 'talks', 'covid', 'lockdown', 'failures', 'via']
['committee', 'spokester', 'amp', 'gop', 'allies', 'go', 'bizarro', 'world']
['watch', 'live', 'covid', 'update']
['buddy', 'half', 'way', 'trumps', 'term', 'thought', 'nation', 'life', 'support', 'thought', 'dramatic', 'changes', 'could', 'help', 'nation', 'survive', 'little', 'done', 'real
ity', 'absolute', 'foreign_policy', 'abomination', 'real', 'miracle']
```

11. Lemmatization is done for only Nouns, Adverbs, Adjectives and Verbs. The stop words are removed and words less than 3 in length are removed

```
['agree', 'people', 'keep', 'support', 'democrats', 'covid', 'lie', 'fake', 'climate', 'crisis', 'fake', 'jan', 'insurrection', 'fake', 'inflation', 'cause', 'greedy', 'corporati
on']
['ad_ihealth', 'covid', 'antigen', 'testing', 'kit', 'sale']
['study_link', 'even_mild', 'covid', 'change', 'brain', 'cnn']
['non', 'death', 'top', 'death', 'march', 'hkt']
['sfc_urge', 'update', 'mass', 'covid', 'testing', 'loom', 'city']
['take', 'many', 'hit', 'crack', 'pipe', 'work', 'well', 'vaccine', 'whole', 'family', 'covid', 'hospitalize', 'healthy', 'people', 'vaccine', 'require', 'god_bless']
['look', 'covid', 'medical', 'waste', 'handle']
['information', 'platform', 'major', 'source', 'covid', 'misinformation', 'include', 'engage', 'sale', 'unproven', 'covid', 'product', 'service', 'treatment', 'lol', 'would']
['uncontrolled', 'spread', 'lead', 'creation', 'bad', 'variant', 'seem', 'set', 'perfect', 'condition', 'bad', 'variant', 'harm', 'country', 'harm', 'world', 'shameful', 'even',
'take', 'simple', 'step', 'continue', 'mask']
['anyone', 'know', 'ihealth', 'covid', 'antigen_rapid', 'test', 'usp', 'approve', 'travel', 'back', 'get', 'conflicting', 'answer', 'online', 'technically', 'test', 'acceptable',
'proctored', 'online', 'since', 'fda_approve', 'antigen', 'report']
['fraud', 'alert', 'covid', 'scam', 'oig']
['global', 'covid', 'death', 'surpass_million']
['abc', 'life', 'death', 'covid']
['work', 'day', 'visa', 'amp', 'provide', 'current', 'company', 'mandatory', 'also', 'medical', 'insurance', 'covid', 'travail', 'need', 'show', 'airport', 'authority']
['ted_cruz', 'get', 'million', 'billionaire', 'fracke', 'donor', 'last', 'covid', 'aid', 'report', 'via']
['covid', 'delta', 'three_chart']
['include', 'covid', 'shot', 'else']
['watch', 'live', 'today', 'covid', 'update']
['ron_desantis', 'talk', 'covid', 'lockdown', 'failure', 'via']
['committee', 'spokester', 'amp', 'gop', 'ally', 'bizarro', 'world']
['watch', 'live', 'covid', 'update']
['buddy', 'half', 'way', 'trump', 'term', 'think', 'nation', 'life', 'support', 'think', 'dramatic', 'change', 'could', 'help', 'nation', 'survive', 'little', 'reality', 'absolut
e', 'foreign_policy', 'abomination', 'real', 'miracle']
```

Topic Modelling

Building the Topic Model

The two main inputs to the Latent Dirichlet Allocation model are the corpus and the dictionary. These are created using the cleaned and lemmatized data.

The corpus contains a unique id for each word in the document. The produced corpus shown below is a mapping of (word_id, word_frequency). The sample corpus created:

```
[[('agree', 1),
  ('cause', 1),
  ('climate', 1),
  ('corporation', 1),
  ('covid', 1),
  ('crisis', 1),
  ('democrats', 1),
  ('fake', 3),
  ('greedy', 1),
  ('inflation', 1),
  ('insurrection', 1),
  ('jan', 1),
  ('keep', 1),
  ('lie', 1),
  ('people', 1),
  ('support', 1)],
```

The LDA model was initially run with 15 topics as the number of topics. Alpha and eta are hyperparameters that affect sparsity of the topics. Both defaults to 1.0/num_topics prior. Alpha represents document-topic density and eta represents topic-word density. Higher the value of alpha, documents are composed of more topics and lower the value of alpha, documents contain fewer topics. The alpha value is set low as the number of words in the tweets are less than in normal documents. The number of iterations to converge is set to 50. The top 15 topics keywords are identified as below:

```
[(0,
  '0.077*"covid" + 0.030*"restriction" + 0.021*"health" + 0.019*"say" + '
  '0.018*"lift" + 0.015*"public" + 0.014*"mask_mandate" + 0.013*"end" + '
  '0.011*"mandate" + 0.010*"measure"'),
 (1,
  '0.072*"covid" + 0.026*"study" + 0.021*"new" + 0.021*"infection" + '
  '0.017*"case" + 0.014*"china" + 0.014*"risk" + 0.014*"variant" + '
  '0.013*"brain" + 0.012*"omicron"'),
 (2,
  '0.053*"die" + 0.051*"covid" + 0.039*"people" + 0.037*"million" + '
  '0.023*"life" + 0.021*"death" + 0.020*"americans" + 0.018*"kill" + '
  '0.016*"many" + 0.016*"pandemic"'),
 (3,
  '0.088*"covid" + 0.075*"case" + 0.050*"death" + 0.042*"new" + 0.035*"report" '
  '+ 0.019*"update" + 0.015*"march" + 0.014*"total" + 0.013*"number" + '
  '0.011*"daily"'),
 (4,
  '0.083*"covid" + 0.035*"dose" + 0.028*"datum" + 0.025*"death" + '
  '0.016*"vaccination" + 0.016*"case" + 0.015*"current_stat" + 0.010*"show" + '
  '0.007*"administer" + 0.007*"vaccine"'),
 (5,
  '0.033*"covid" + 0.027*"amp" + 0.012*"pandemic" + 0.011*"people" + '
  '0.010*"work" + 0.009*"woman" + 0.008*"make" + 0.006*"can" + 0.006*"not" + '
  '0.006*"world"'),
 (6,
  '0.079*"year" + 0.065*"covid" + 0.058*"pandemic" + 0.045*"two" + '
  '0.017*"since" + 0.015*"today" + 0.015*"ago" + 0.015*"day" + 0.014*"first" + '
  '0.012*"last"'),
```

```
(7,
 '0.154*"test" + 0.131*"covid" + 0.114*"positive" + 0.024*"home" + '
 '0.020*"free" + 0.019*"order" + 0.015*"obama" + 0.014*"say" + '
 '0.011*"symptom" + 0.009*"negative"'),
(8,
 '0.054*"covid" + 0.018*"patient" + 0.015*"pandemic" + 0.011*"care" + '
 '0.008*"impact" + 0.008*"hospital" + 0.008*"due" + 0.007*"business" + '
 '0.007*"research" + 0.006*"medical"'),
(9,
 '0.052*"covid" + 0.033*"get" + 0.019*"people" + 0.018*"mask" + 0.013*"still" '
 '+ 0.012*"know" + 0.011*"say" + 0.011*"like" + 0.009*"wear" + 0.009*"one"'),
(10,
 '0.056*"covid" + 0.020*"school" + 0.017*"community" + 0.013*"available" + '
 '0.012*"amp" + 0.012*"health" + 0.011*"student" + 0.011*"march" + '
 '0.011*"mask" + 0.011*"vaccination"'),
(11,
 '0.076*"covid" + 0.022*"ukraine" + 0.015*"war" + 0.013*"world" + '
 '0.012*"russia" + 0.011*"china" + 0.009*"trump" + 0.009*"like" + '
 '0.008*"putin" + 0.007*"lab"'),
(12,
 '0.109*"vaccine" + 0.088*"covid" + 0.028*"child" + 0.013*"pfizer" + '
 '0.012*"dose" + 0.011*"kid" + 0.011*"florida" + 0.010*"vaccination" + '
 '0.010*"via" + 0.009*"get"'),
(13,
 '0.050*"covid" + 0.020*"get" + 0.019*"amp" + 0.016*"vaccine" + '
 '0.012*"booster" + 0.011*"clinic" + 0.010*"book" + 0.009*"vaccination" + '
 '0.008*"check" + 0.008*"new"'),
(14,
 '0.061*"covid" + 0.025*"health" + 0.020*"pandemic" + 0.011*"fund" + '
 '0.011*"billion" + 0.010*"public" + 0.009*"mental" + 0.009*"say" + '
 '0.009*"government" + 0.009*"use"')]
```

Evaluating the model

The model's perplexity and topic coherence are used as measures for a good model.

Perplexity is a measurement of how well a probability distribution or probability model predicts a sample. A low perplexity indicates the probability distribution is good at predicting the sample.

We can use the coherence score in topic modeling to measure how interpretable the topics are to humans. In this case, topics are represented as the top N words with the highest probability of belonging to that particular topic.

The perplexity and coherence score of the model are calculated.

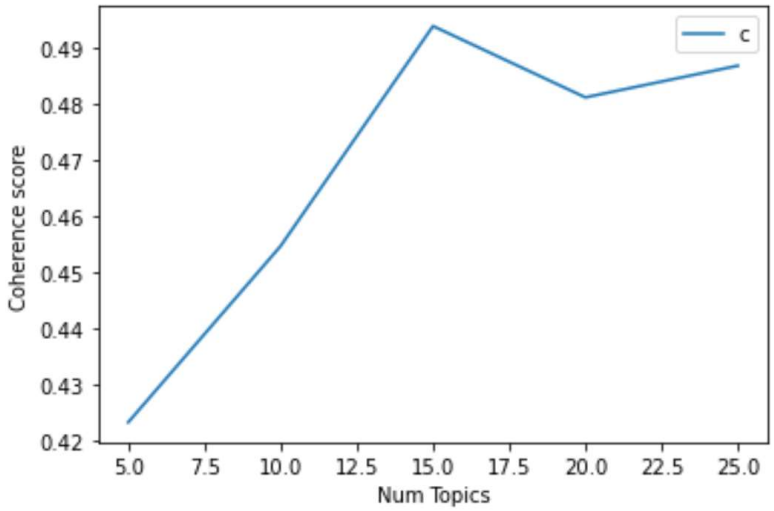
```
Perplexity:  -8.078994434360524

Coherence Score:  0.4874729225283338
```

In order to choose the optimum number of topics to build the model, the model is run with topic numbers from 5-25 in steps of 5. The coherence and the perplexity for each of the models is calculated. The model with the highest score for coherence is chosen as the optimal model.

```
Num Topics = 5  has Coherence Value of 0.4232
Num Topics = 10  has Coherence Value of 0.4547
Num Topics = 15  has Coherence Value of 0.494
Num Topics = 20  has Coherence Value of 0.4812
Num Topics = 25  has Coherence Value of 0.4869
```

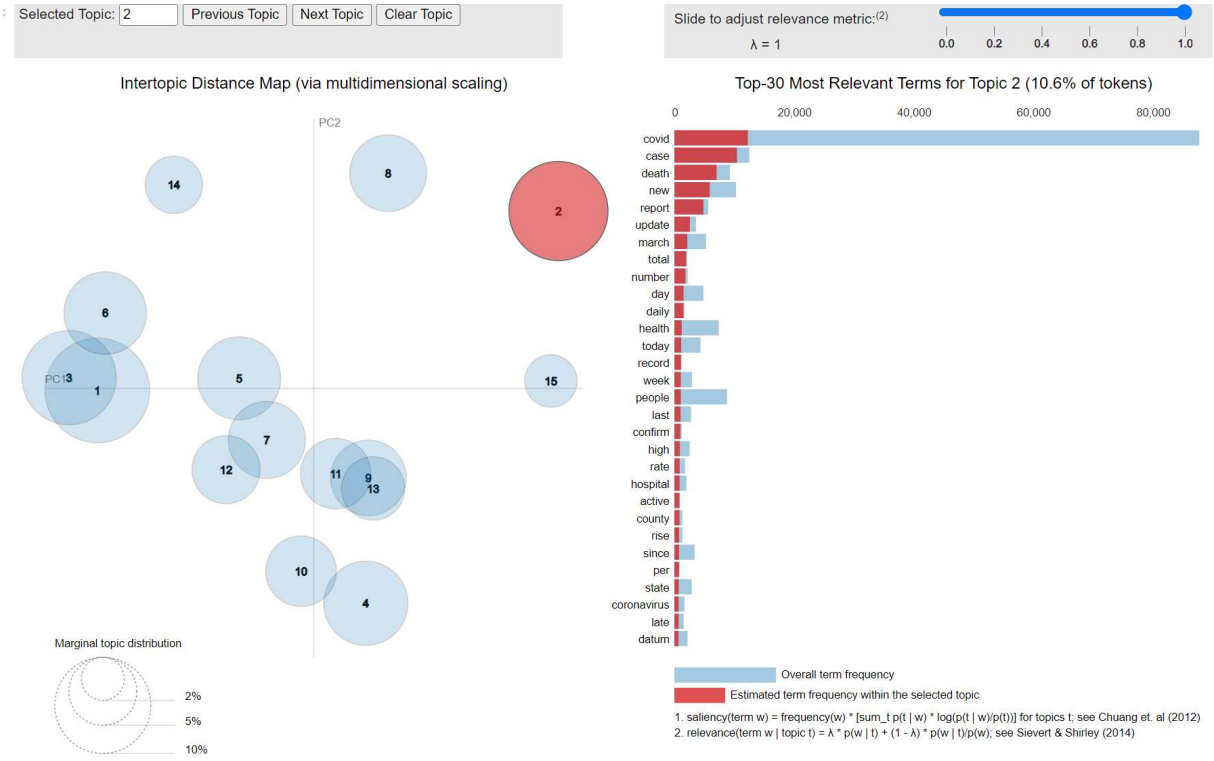
The number topics and coherence values as plotted:

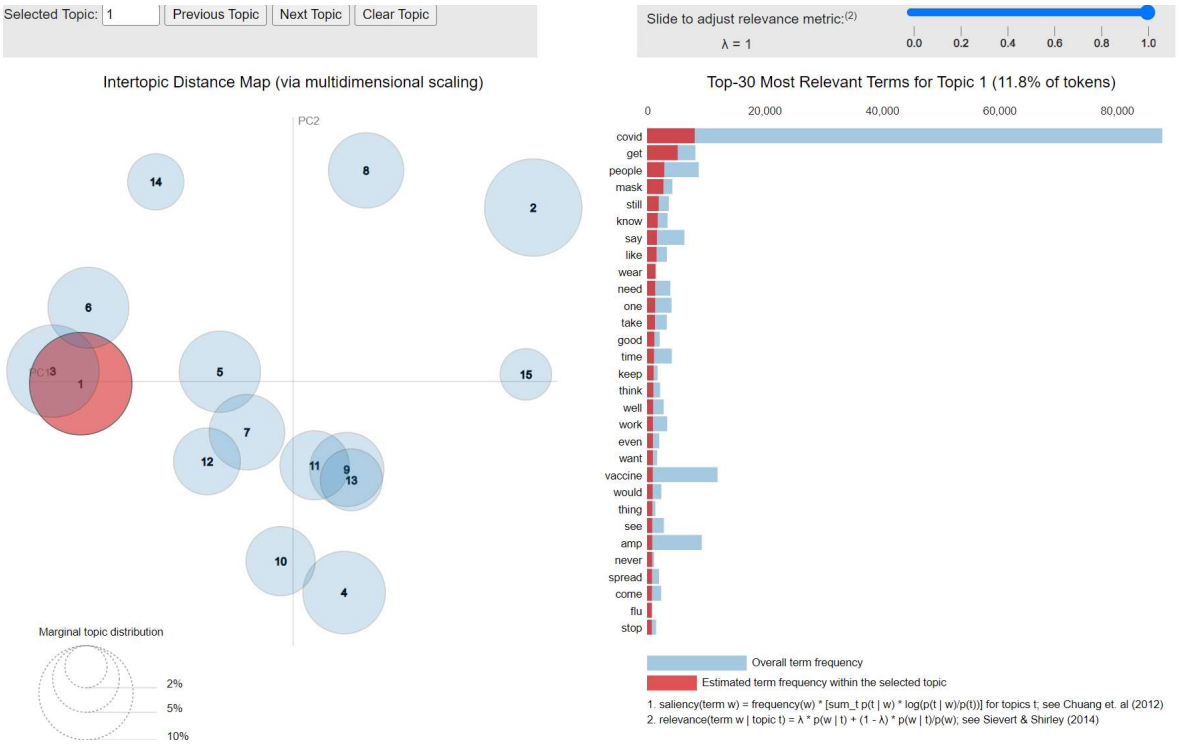


The optimal model is the model built with 15 topics as it has the highest coherence value.

Visualize the topic keywords

An interactive chart is built to visualize the topics. Each topic is represented by a bubble. The larger the bubbles and lesser the overlap the better the model is. The keywords that are most likely to be associated with that topic are also shown in the visualization.


Selected Topic: 2
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0 0.2 0.4 0.6 0.8 1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 2 (10.6% of tokens)
0 20,000 40,000 60,000 80,000
covid
case
death
new
report
update
march
total
number
day
daily
health
today
record
week
people
last
confirm
high
rate
hospital
active
county
rise
since
per
state
coronavirus
late
datum
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)


Selected Topic: 1
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0 0.2 0.4 0.6 0.8 1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 1 (11.8% of tokens)
0 20,000 40,000 60,000 80,000
covid
get
people
mask
still
know
say
like
wear
need
one
take
good
time
keep
think
well
work
even
want
vaccine
would
thing
see
amp
never
spread
come
flu
stop
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

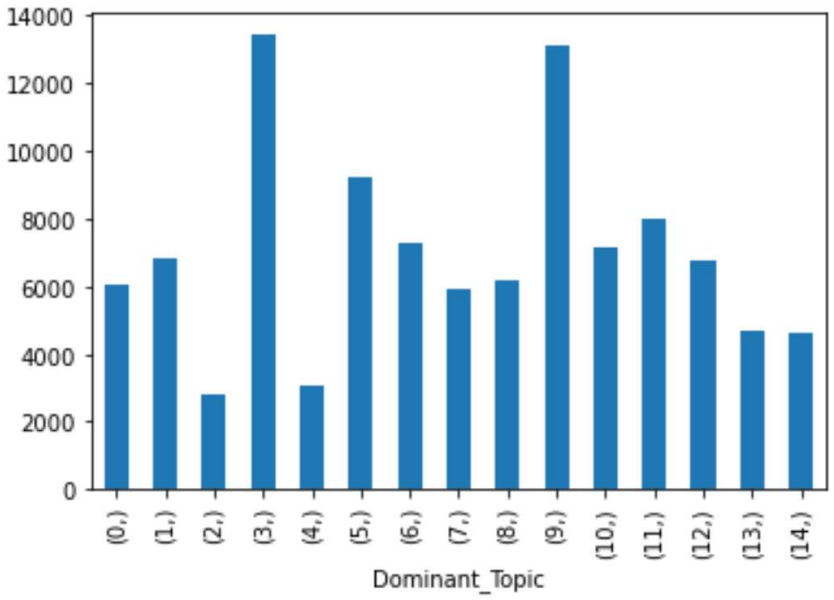
Assigning a Dominant Topic to each document in the corpus

A dominant topic is assigned to each document in the tweet corpus. One of the important applications of topic modeling is to determine what topic a given document is about.

The topic number that has the highest percentage contribution in that document is assigned as the Dominant Topic for that document.

	Dominant_Topic	Perc_Contribution	Topic_Keywords	
0	11	0.5456	covid, ukraine, war, world, russia, china, like, trump, putin, lab	@MaryLTrump @DavidCornDC Agreed, why do people keep on supporting Democrats??? From the COVID-19 lies to the the fake Jan 6th insurrection to the fake inflation caused by greedy
1	7	0.8173	test, covid, positive, home, free, order, obama, say, symptom, negative	AD iHealth™ Covid-19 Antigen testing kits for sale http
2	1	0.9977	covid, study, new, infection, case, china, risk, variant, brain, omicron	Study links even mild Covid-19 to changes in the brain - CNN #DemForce ht
3	3	0.8239	covid, case, death, new, report, update, march, total, number, day	,...a NON-#covid19 death ON TOP OF ALL the #covid19 deaths (March 7, 2022; 22:17 HKT
4	0	0.4599	covid, restriction, health, say, lift, public, end, mask_mandate, mandate, measure	SFC urges Hong Kong businesses to update continuity plans as mass Covid-19 testing looms in the city ht
...	...	...	...	
104995	9	0.7381	covid, get, people, mask, still, know, say, like, wear, need	What do you think about the 4th wave of covid-19 in India?\n\nKnow more: https://t.co/jDqGkNnzMY\n\ #4th_wave_of_covid_in_india #4th_covid_wave #Covid_case_in_India h
104996	4	0.7730	covid, dose, datum, death, case, current_stat, vaccination, show, administer, vaccine	Cervical longitudinally extensive myelitis after #CovidVaccination: \nhttps://t.co/dcXxoaid5Z\n"In summary, we emphasize ag should be aware of the possibility of development of LTEM after different COVID-19 vaccines."\n#VaccineS
			vaccine, covid, child, dose, pfizer, kid,	

Distribution of the dominant topics within the corpus is visualized as follows:



Topic 3 is the most frequent topic in this tweet dataset. The keywords for topic 3 are:

```
(3,
 '0.088*"covid" + 0.075*"case" + 0.050*"death" + 0.042*"new" + 0.035*"report" '
 '+ 0.019*"update" + 0.015*"march" + 0.014*"total" + 0.013*"number" + '
 '0.011*"day"'),
```

Sentiment Analysis

The sentiment analysis is done after data preprocessing.

1. Each tweet is split into individual words
2. The words are converted to lower case
3. Contractions are replaced with their full form. For example, didn't is replaced with did not.
4. URLs are removed
5. @ Mentions are removed
6. Special characters are removed
7. Hashtags are removed
8. The sentences are tokenized into words
9. Punctuations are removed
10. Bigrams and Trigrams are built
11. Stop words are removed

VADER Lexicon Based

The documents are assigned a Dominant Topic after the topic modelling is complete. A lexicon-based approach is used to determine the sentiments of tweets within each of the topics. The first level of classification is done as Positive, Negative or Neutral.

Valence Aware Dictionary and Sentiment Reasoner is a lexicon and rule-based sentiment analysis tool which is used to classify the sentiments in the tweets. It is sensitive to both polarity (positive/negative) and intensity (strength) of emotion.

The neg (negative), neu (neutral), and pos (positive) represent the proportion of text falling under each category and the proportion will sum up to one.

Compound score reflects the overall score. It is sum of all lexicon ratings which is normalized between -1 (most extreme negative) and +1 (most extreme positive).

The following thresholds are used for classifying the sentiments:

- Positive sentiment: compound score >= 0.05
- Neutral sentiment: (compound score > -0.05) and (compound score < 0.05)

- Negative sentiment: compound score <= -0.05

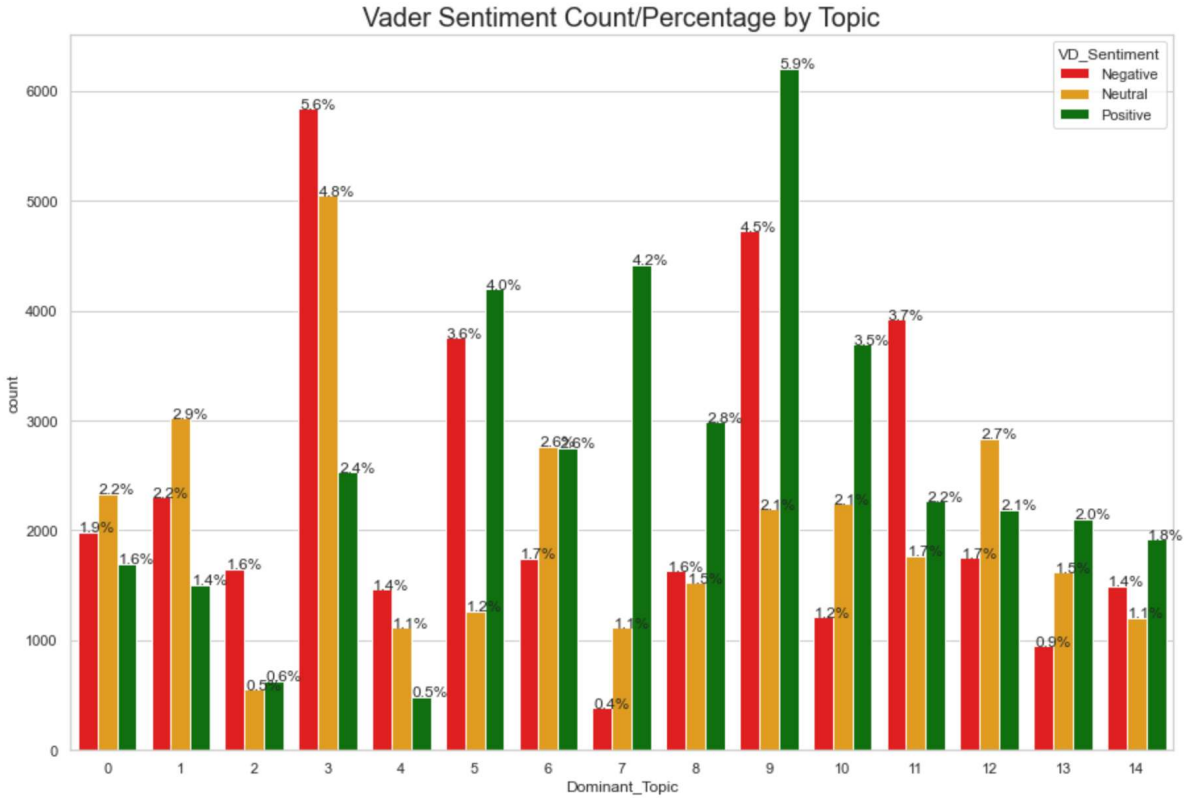
	Dominant_Topic	Perc_Contribution	Topic_Keywords	text	clean_text	VD_Scores	VD_Compound	VD_Sentiment	temp_list	TB_score
0	11	0.5456	covid, ukraine, war, world, russia, china, lik...	@MaryLTrump @DavidCornDC Agreed, why do people...	agree people keep support democrats covid lie...	{'neg': 0.492, 'neu': 0.345, 'pos': 0.163, 'co...	-0.8885	Negative	[agree, people, keep, support, democrats, covi...	-0.50000
1	7	0.8173	test, covid, positive, home, free, order, obam...	AD iHealth™ Covid-19 Antigen testing kits for ...	ad_ihealth covid antigen testing kit sale	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...	0.0000	Neutral	[ad_ihealth, covid, antigen, testing, kit, sale]	0.00000
2	1	0.9977	covid, study, new, infection, case, china, ris...	Study links even mild Covid-19 to changes in t...	study_link even_mild covid change brain cnn	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...	0.0000	Neutral	[study_link, even_mild, covid, change, brain, ...	0.00000
3	3	0.8239	covid, case, death, new, report, update, march...	,...a NON-#covid19 death ON TOP OF ALL the #co...	non death top death march hkt	{'neg': 0.619, 'neu': 0.238, 'pos': 0.143, 'co...	-0.7906	Negative	[non, death, top, death, march, hkt]	0.50000
4	0	0.4599	covid, restriction, health, say, lift, public,...	SFC urges Hong Kong businesses to update conti...	sfc_urge update mass covid testing loom city	{'neg': 0.241, 'neu': 0.759, 'pos': 0.0, 'comp...	-0.2263	Negative	[sfc_urge, update, mass, covid, testing, loom,...	0.00000

The distribution of Positive, Negative and Neutral sentiments within each topic is visualized as below.

Topic 3 = Has the highest percentage of Negative sentiments overall

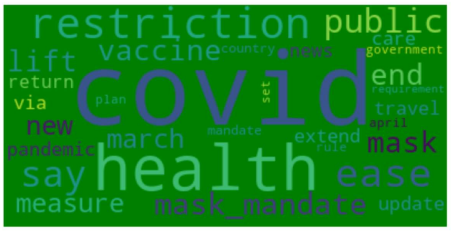
Topic 9 – Has the highest percentage of Positive sentiments overall

Topic 3 – Has the highest percentage of Neutral sentiments overall



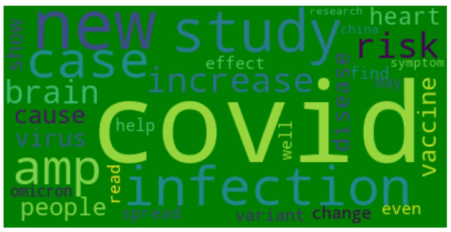
Word clouds for positive tweets within each topic:

WordCloud of Positive Sentiment Tweets for Topic 0



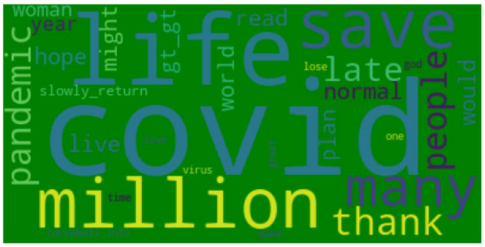
<Figure size 432x288 with 0 Axes>

WordCloud of Positive Sentiment Tweets for Topic 1



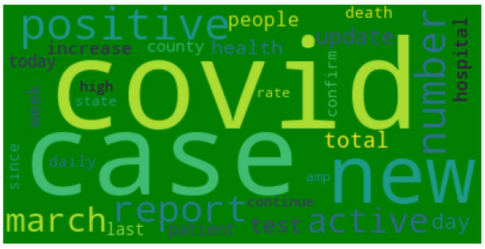
<Figure size 432x288 with 0 Axes>

WordCloud of Positive Sentiment Tweets for Topic 2



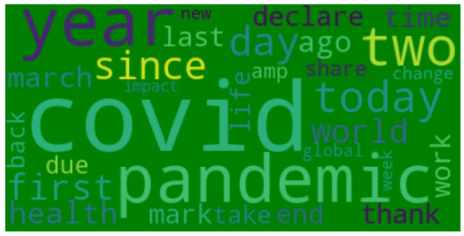
<Figure size 432x288 with 0 Axes>

WordCloud of Positive Sentiment Tweets for Topic 3



<Figure size 432x288 with 0 Axes>

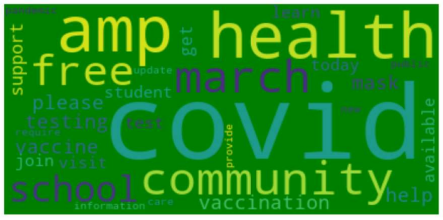
WordCloud of Positive Sentiment Tweets for Topic 6



<Figure size 432x288 with 0 Axes>

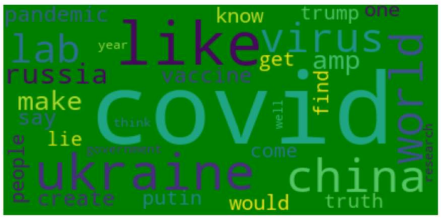
WordCloud of Positive Sentiment Tweets for Topic 7

WordCloud of Positive Sentiment Tweets for Topic 10



<Figure size 432x288 with 0 Axes>

WordCloud of Positive Sentiment Tweets for Topic 11

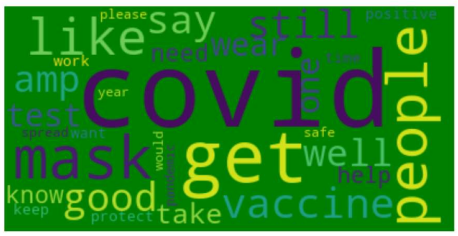


<Figure size 432x288 with 0 Axes>

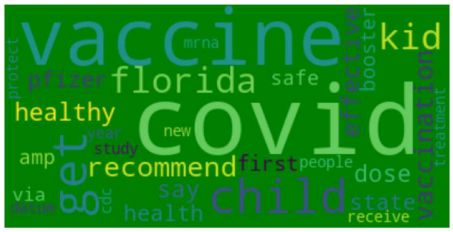
WordCloud of Positive Sentiment Tweets for Topic 8

<Figure size 432x288 with 0 Axes>

WordCloud of Positive Sentiment Tweets for Topic 9

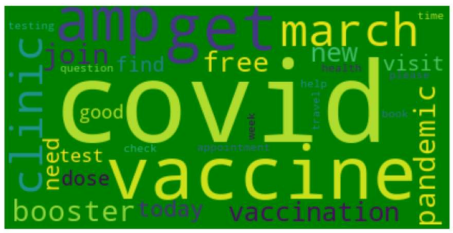


WordCloud of Positive Sentiment Tweets for Topic 12

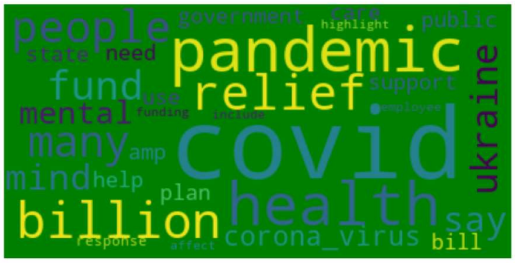


<Figure size 432x288 with 0 Axes>

WordCloud of Positive Sentiment Tweets for Topic 13



WordCloud of Positive Sentiment Tweets for Topic 14



<Figure size 432x288 with 0 Axes>

Word Cloud for Negative Sentiments within each topic

WordCloud of Negative Sentiment Tweets for Topic 0

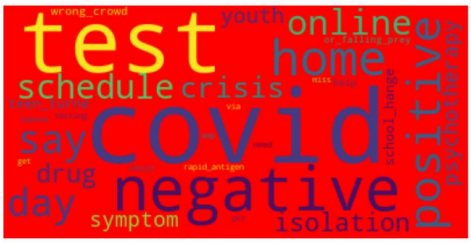
WordCloud of Negative Sentiment Tweets for Topic 1

WordCloud of Negative Sentiment Tweets for Topic 2

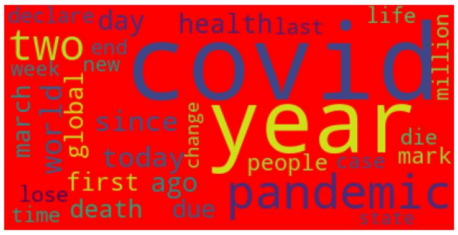
WordCloud of Negative Sentiment Tweets for Topic 3

WordCloud of Negative Sentiment Tweets for Topic 5

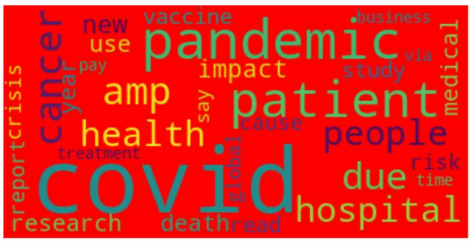
WordCloud of Negative Sentiment Tweets for Topic 7



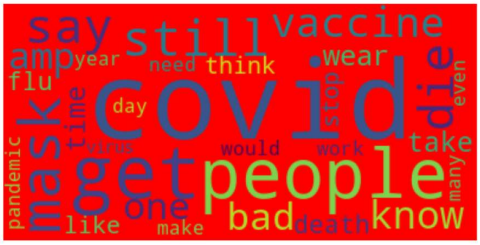
WordCloud of Negative Sentiment Tweets for Topic 6



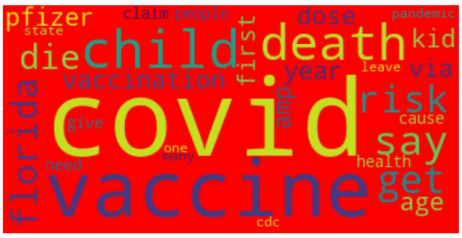
WordCloud of Negative Sentiment Tweets for Topic 8



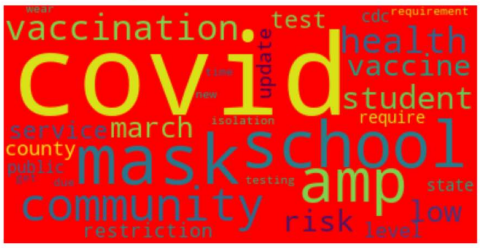
WordCloud of Negative Sentiment Tweets for Topic 9



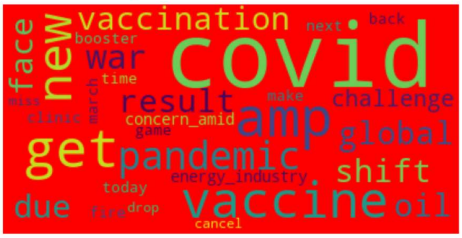
WordCloud of Negative Sentiment Tweets for Topic 12



WordCloud of Negative Sentiment Tweets for Topic 10



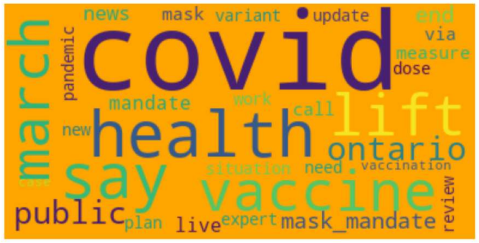
WordCloud of Negative Sentiment Tweets for Topic 13



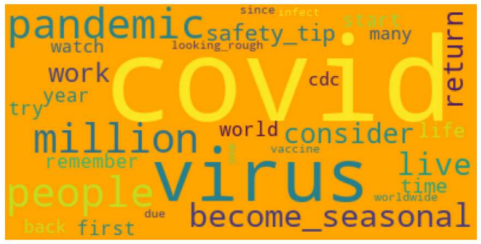
WordCloud of Negative Sentiment Tweets for Topic 14

Word Cloud for Neutral Sentiments within each topic

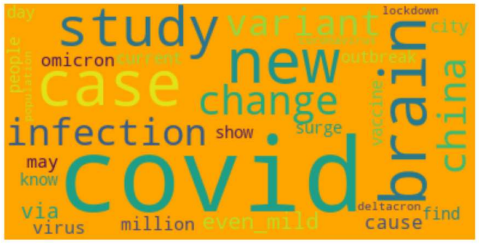
WordCloud of Neutral Sentiment Tweets for Topic 0



WordCloud of Neutral Sentiment Tweets for Topic 2

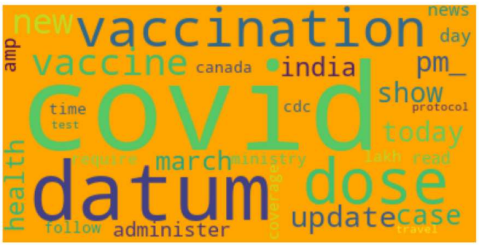


WordCloud of Neutral Sentiment Tweets for Topic 1

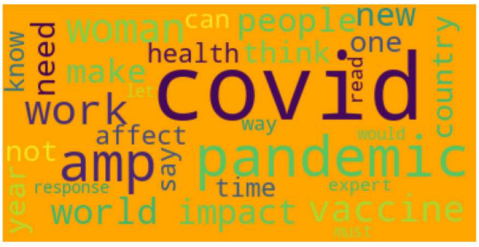


WordCloud of Neutral Sentiment Tweets for Topic 3

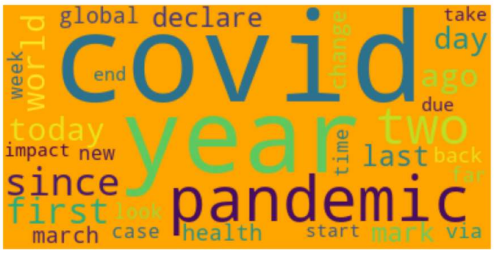
WordCloud of Neutral Sentiment Tweets for Topic 4



WordCloud of Neutral Sentiment Tweets for Topic 5



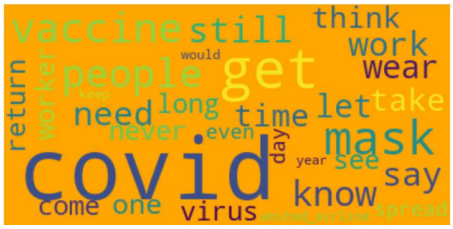
WordCloud of Neutral Sentiment Tweets for Topic 6



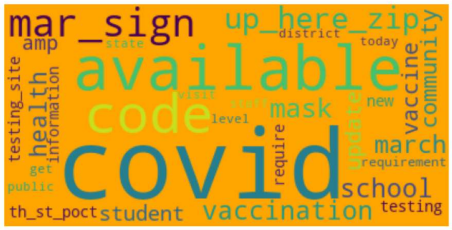
WordCloud of Neutral Sentiment Tweets for Topic 7

WordCloud of Neutral Sentiment Tweets for Topic 8

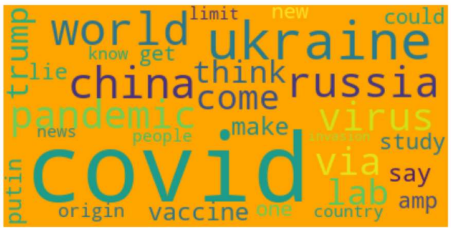
WordCloud of Neutral Sentiment Tweets for Topic 9

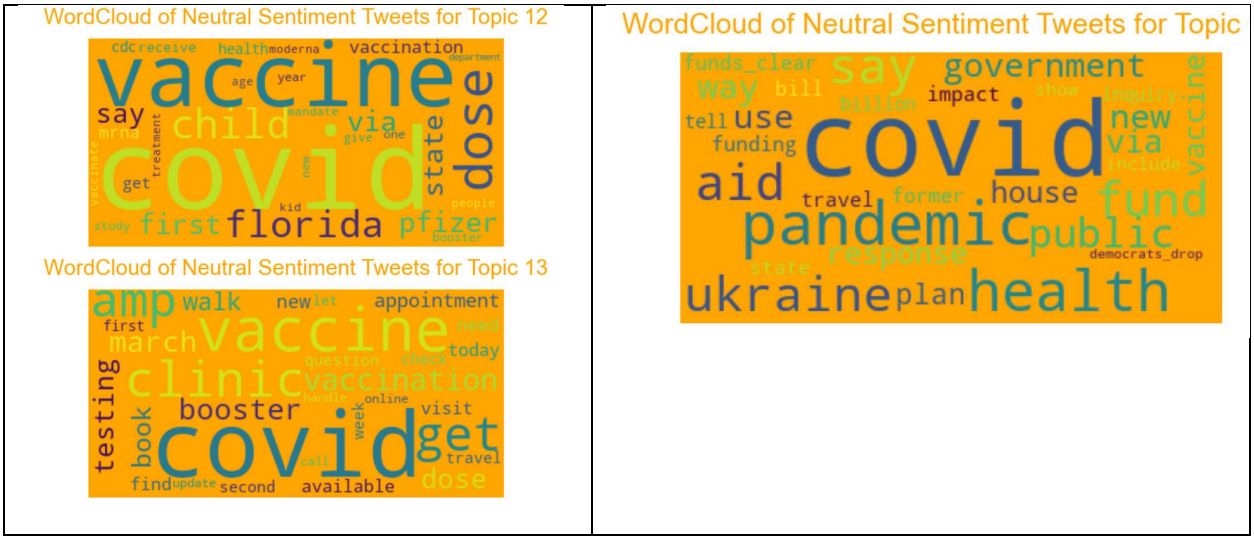


WordCloud of Neutral Sentiment Tweets for Topic 10



WordCloud of Neutral Sentiment Tweets for Topic 11





TextBlob based Sentiment Analysis

TextBlob Sentiment Analysis assigns both a polarity score (-1, +1) and a subjectivity score (0,1).

Polarity is of 'float' type and lies in the range of -1,1, where 1 means a high positive sentiment, and -1 means a high negative sentiment.

Subjectivity is also of 'float' type and lies in the range of 0,1. The value closer to 1 indicates that the sentence is mostly a public opinion and not a factual piece of information and vice versa.

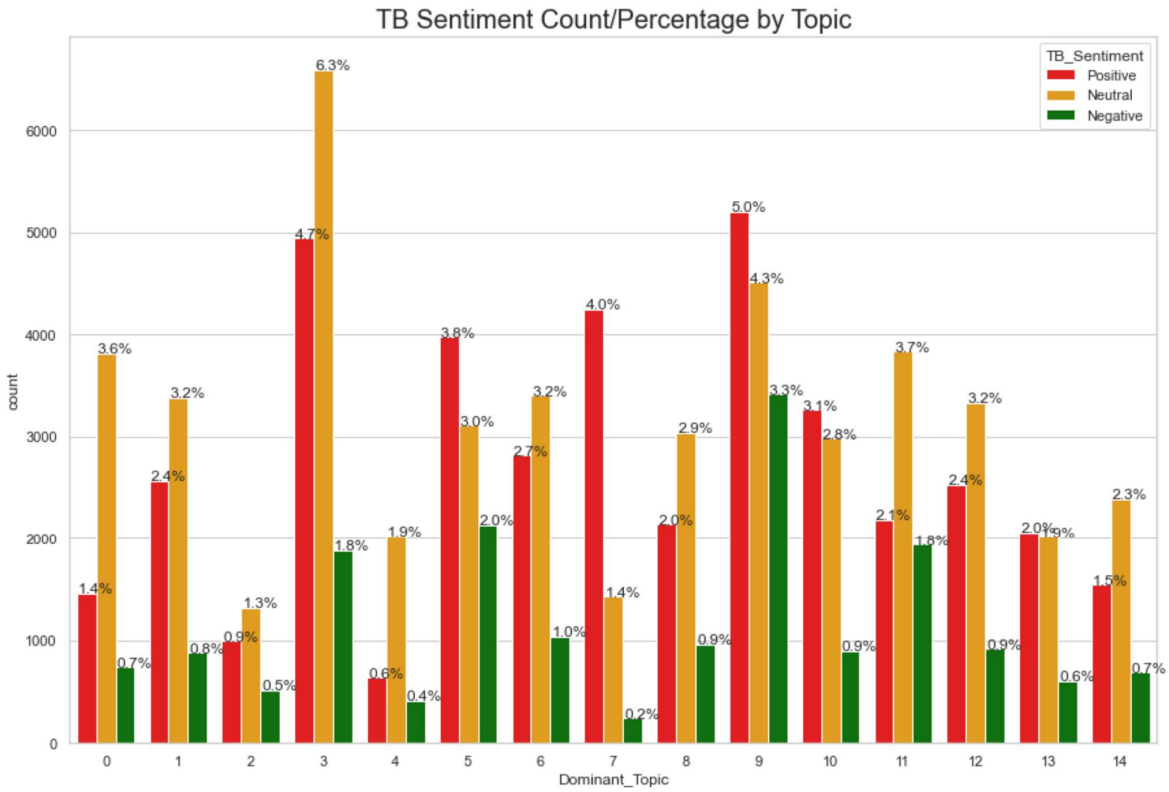
The following thresholds are used for classifying the sentiments:

- Positive sentiment: compound score >= 0.05
- Neutral sentiment: (compound score > -0.05) and (compound score < 0.05)
- Negative sentiment: compound score <= -0.05

Using the TextBlob library the sentiments within each topic are classified as follows:

	clean_text	VD_Scores	VD_Compound	VD_Sentiment	temp_list	TB_score	TB_subjectivity	TB_Subjectivity	TB_Sentiment
	agree people keep support democrats covid lie...	{'neg': 0.492, 'neu': 0.345, 'pos': 0.163, 'co...	-0.8885	Negative	[agree, people, keep, support, democrats, covi...	-0.500000	1.000000	Opinion	Negative
	ad_ihealth covid antigen testing kit sale	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...	0.0000	Neutral	[ad_ihealth, covid, antigen, testing, kit, sale]	0.000000	0.000000	Factual	Neutral
	study_link even_mild covid change brain cnn	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...	0.0000	Neutral	[study_link, even_mild, covid, change, brain, ...	0.000000	0.000000	Factual	Neutral
	non death top death march hkt	{'neg': 0.619, 'neu': 0.238, 'pos': 0.143, 'co...	-0.7906	Negative	[non, death, top, death, march, hkt]	0.500000	0.500000	Opinion	Positive
	sfc_urge update mass covid testing loom city	{'neg': 0.241, 'neu': 0.759, 'pos': 0.0, 'comp...	-0.2263	Negative	[sfc_urge, update, mass, covid, testing, loom,...	0.000000	0.000000	Factual	Neutral
	take many hit crack pipe work well vaccine wh...	{'neg': 0.0, 'neu': 0.758, 'pos': 0.242, 'comp...	0.5859	Positive	[take, many, hit, crack, pipe, work, well, vac...	0.400000	0.466667	Factual	Positive
	look covid medical waste handle	{'neg': 0.412, 'neu': 0.588, 'pos': 0.0, 'comp...	-0.4215	Negative	[look, covid, medical, waste, handle]	-0.100000	0.000000	Factual	Negative

The distribution of Positive, Negative and Neutral sentiments within each topic is visualized as below.

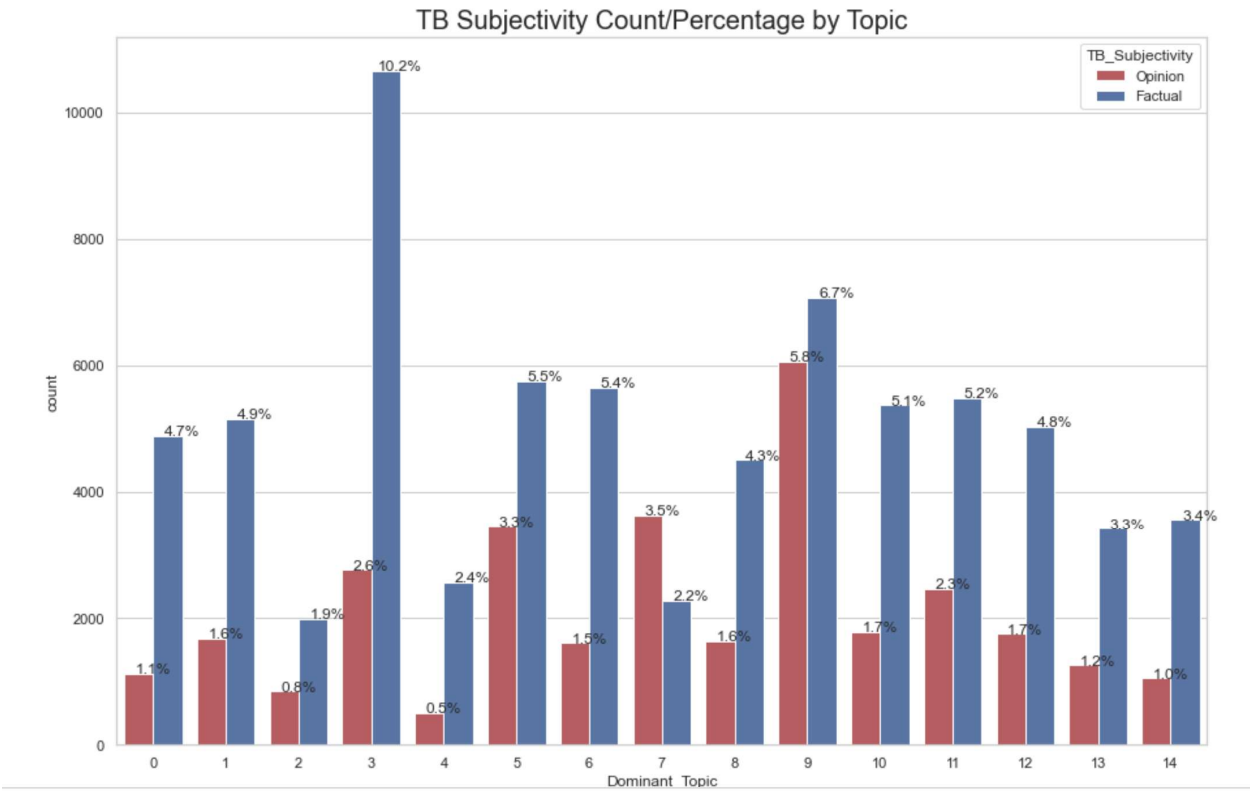


Topic 9 = Has the highest percentage of Negative sentiments overall

Topic 9 – Has the highest percentage of Positive sentiments overall

Topic 3 – Has the highest percentage of Neutral sentiments overall

The distribution of Opinion/Factual tweets (subjectivity) within each topic is visualized as below:



Topic 9 – Has the highest percentage of opinion tweets overall

Topic 3 – Has the highest percentage of Factual tweets overall

NRCLex based Sentiment Analysis

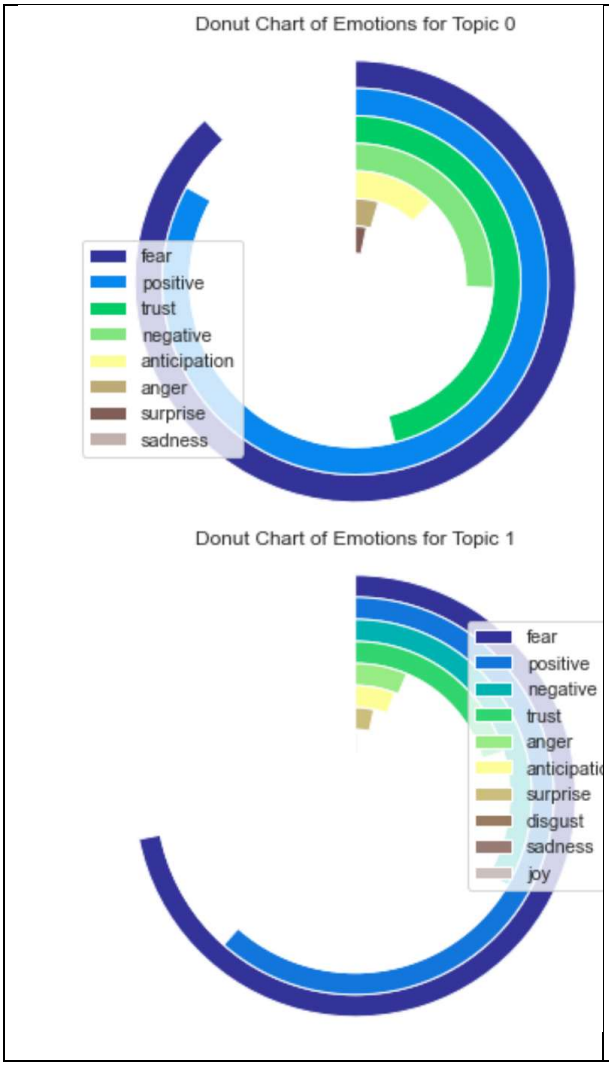
NRCLex will measure emotional affect from a body of text. The dictionary contains approximately 27,000 words, and is based on the National Research Council Canada (NRC) affect lexicon. NRCLex assigns an emotion to the document based on the words in the document.

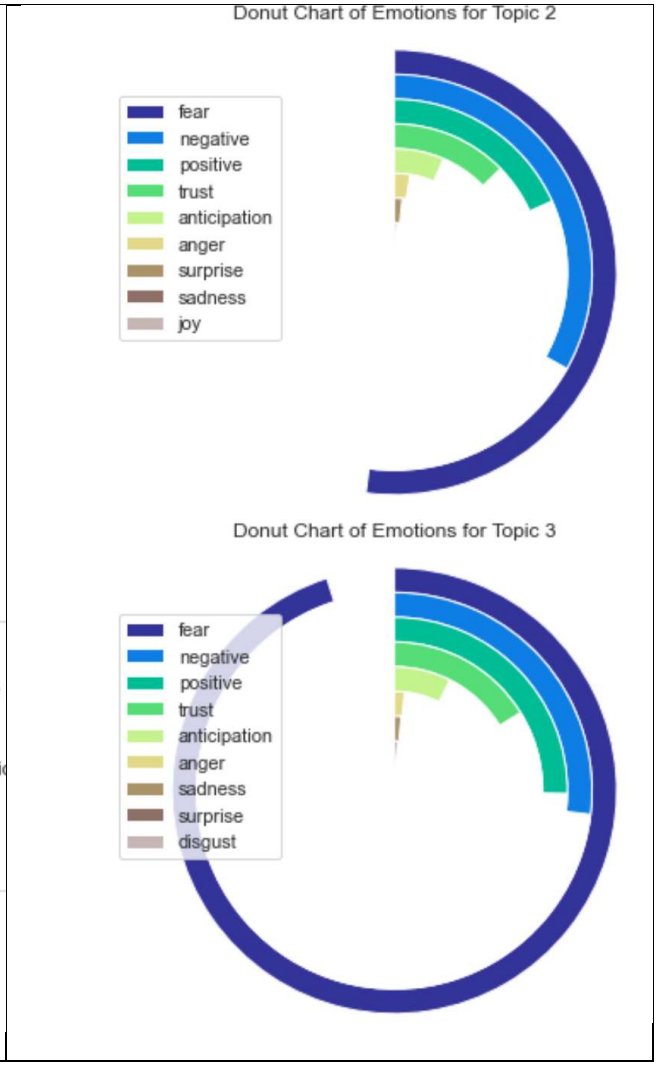
The NRC Emotion Lexicon is a list of English words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive). There is a top emotion assigned or if the value is zero, it is assigned as No Emotion.

The emotions/sentiments are assigned using the NRCLex:

text	clean_text	VD_Scores	VD_Compound	VD_Sentiment	temp_list	TB_score	TB_subjectivity	TB_Subjectivity	TB_Sentiment	Emotion
rnDC ple...	agree people keep support democrats covid lie...	{'neg': 0.492, 'neu': 0.345, 'pos': 0.163, 'co...	-0.8885	Negative	[agree, people, keep, support, democrats, covi...	-0.500000	1.000000	Opinion	Negative	negative
igen for ...	ad_ihealth covid antigen testing kit sale	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...	0.0000	Neutral	[ad_ihealth, covid, antigen, testing, kit, sale]	0.000000	0.000000	Factual	Neutral	No emotion
d-19 in t...	study_link even_mild covid change brain cnn	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...	0.0000	Neutral	[study_link, even_mild, covid, change, brain, ...	0.000000	0.000000	Factual	Neutral	fear
ı ON #co...	non death top death march hkt	{'neg': 0.619, 'neu': 0.238, 'pos': 0.143, 'co...	-0.7906	Negative	[non, death, top, death, march, hkt]	0.500000	0.500000	Opinion	Positive	anticipation
Kong onti...	sfc_urge update mass covid testing loom city	{'neg': 0.241, 'neu': 0.759, 'pos': 0.0, 'comp...	-0.2263	Negative	[sfc_urge, update, mass, covid, testing, loom,...	0.000000	0.000000	Factual	Neutral	fear
otley w t...	take many hit crack pipe work well vaccine wh...	{'neg': 0.0, 'neu': 0.758, 'pos': 0.242, 'comp...	0.5859	Positive	[take, many, hit, crack, pipe, work, well, vac...	0.400000	0.466667	Factual	Positive	positive
vaste ndl...	look covid medical waste handle	{'neg': 0.412, 'neu': 0.588, 'pos': 0.0, 'comp...	-0.4215	Negative	[look, covid, medical, waste, handle]	-0.100000	0.000000	Factual	Negative	fear
ı the maj...	information platform major source covid misin...	{'neg': 0.123, 'neu': 0.626, 'pos': 0.25, 'com...	0.3892	Positive	[information, platform, major, source, covid, ...	0.431250	0.600000	Opinion	Positive	positive

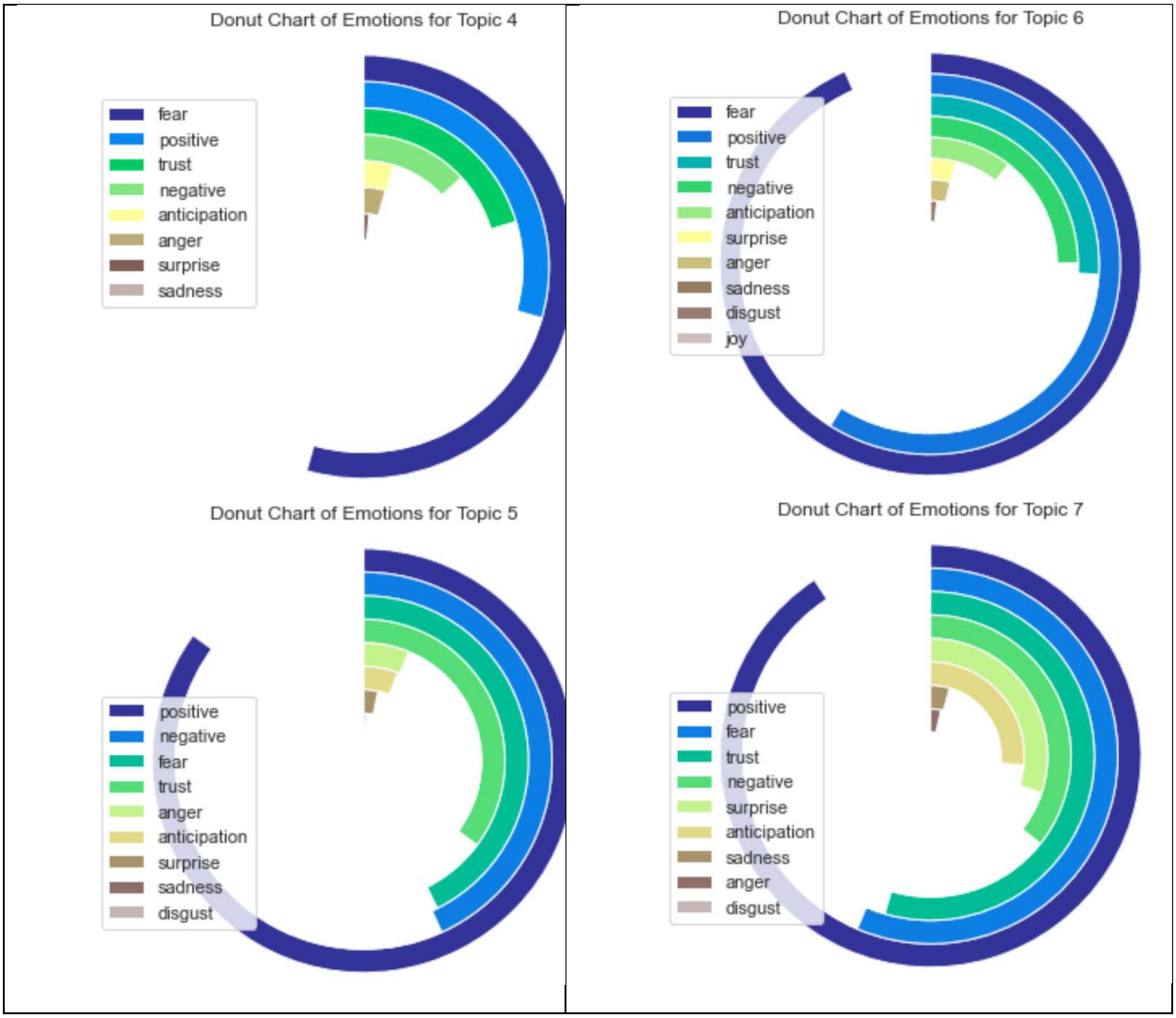
The distribution of Emotions/Sentiments of tweets within each topic is visualized as below:

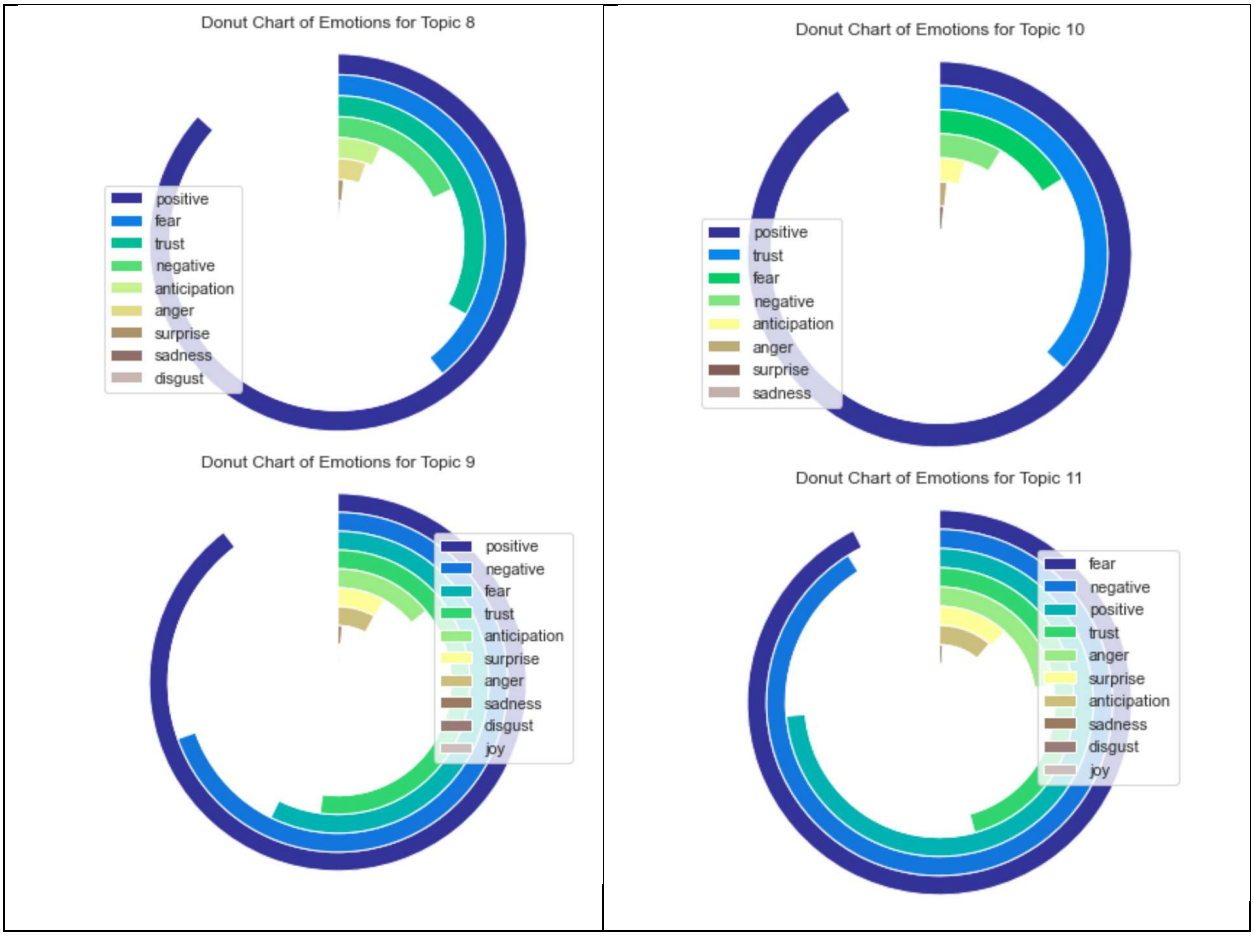

Donut Chart of Emotions for Topic 0
fear
positive
trust
negative
anticipation
anger
surprise
sadness

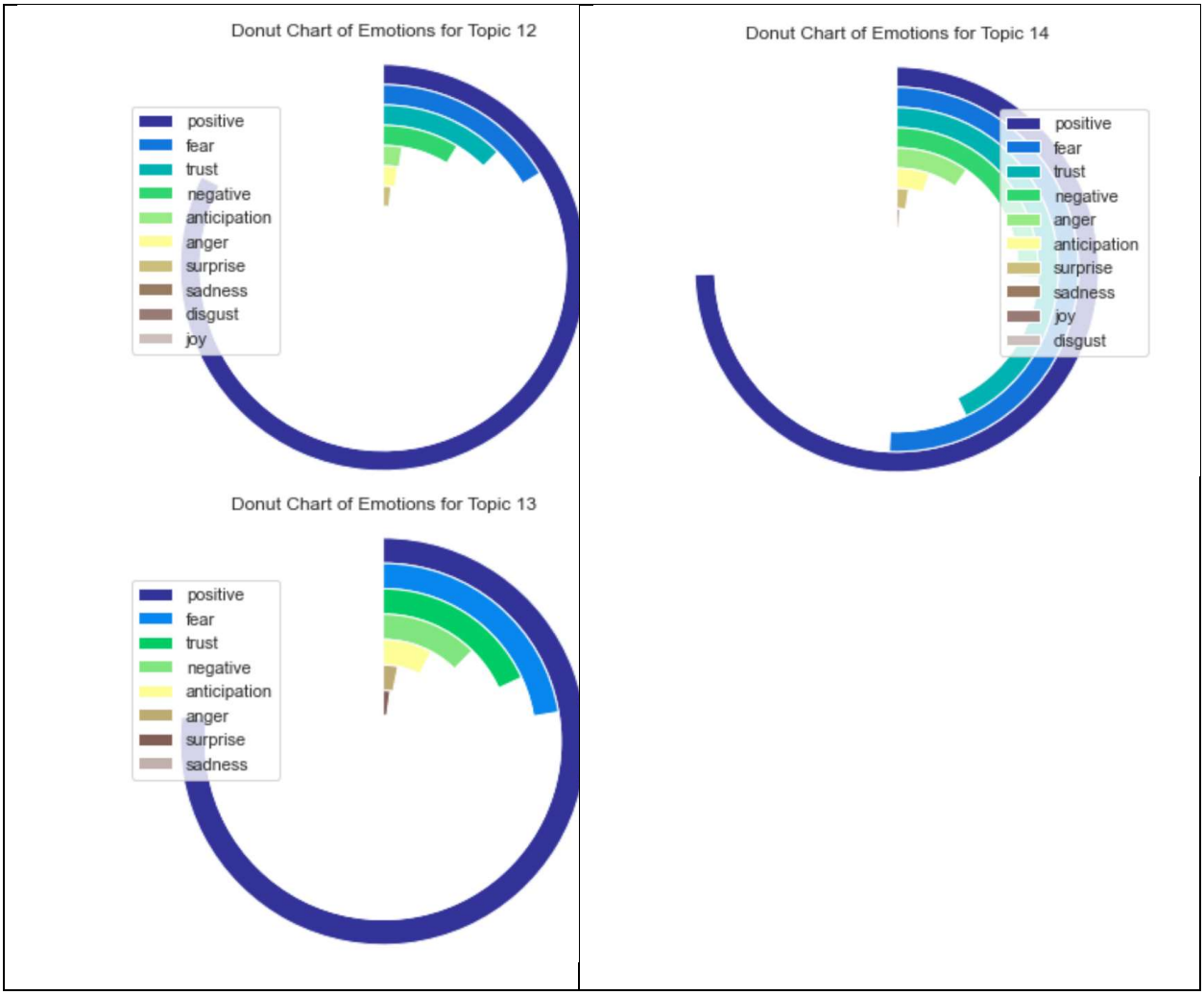

Donut Chart of Emotions for Topic 2
fear
negative
positive
trust
anticipation
anger
surprise
sadness
joy

Donut Chart of Emotions for Topic 1
fear
positive
negative
trust
anger
surprise
disgust
sadness
joy

Donut Chart of Emotions for Topic 3
fear
negative
positive
trust
anticipation
anger
sadness
surprise
disgust


Donut Chart of Emotions for Topic 4
fear
positive
trust
negative
anticipation
anger
surprise
sadness
Donut Chart of Emotions for Topic 6
fear
positive
trust
negative
anticipation
surprise
anger
sadness
disgust
joy
Donut Chart of Emotions for Topic 5
positive
negative
fear
trust
anger
anticipation
surprise
sadness
disgust
Donut Chart of Emotions for Topic 7
positive
fear
trust
negative
surprise
anticipation
sadness
anger
disgust


Donut Chart of Emotions for Topic 8
positive
fear
trust
negative
anticipation
anger
surprise
sadness
disgust
Donut Chart of Emotions for Topic 10
positive
trust
fear
negative
anticipation
anger
surprise
sadness
Donut Chart of Emotions for Topic 9
positive
negative
fear
trust
anticipation
surprise
anger
sadness
disgust
joy
Donut Chart of Emotions for Topic 11
fear
negative
positive
trust
anger
surprise
anticipation
sadness
disgust
joy



Conclusion

Project Results

The aim of the project was to identify the topics and sentiments within the topics from the microblogging data of Twitter. The following has been achieved in the project:

- The tweets have been scraped from Twitter with a particular hashtag. Data cleaning and preprocessing has been done on the dataset.
- The top topics and the keywords for each topic have been identified using the LDA topic modeling algorithm
- The topics and keywords have been visualized
- The model has been evaluated using the appropriate key performance measures

- The optimal model has been chosen and documents in the dataset have been assigned a dominant topic
- The sentiments within each topic have been identified and visualized

Future Scope

In the present study, only the LDA model was used for identifying the topics. Different topic modelling approaches can be applied to the dataset. The models can be compared to see if the key performance measures like coherence can be improved

Emoticons were not used for the classification of sentiments/emotions in this project. In the future emoticons can be included in the sentiment analysis.

The tweets can be downloaded and time series analysis of topics and sentiments can be done.

References

[1] Bing Liu. Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, May 2012.

[2] Liu B., Zhang L. (2012) A Survey of Opinion Mining and Sentiment Analysis. In: Aggarwal C., Zhai C. (eds) Mining Text Data. Springer, Boston, MA. https://doi.org/10.1007/978-1-4614-3223-4_13

[3] **Sanjiv R. Das**, **Mike Y. Chen**, (2007) Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web. Management Science 53(9):1375-1388. https://doi.org/10.1287/mnsc.1070.0704

[4] Staci M.Zavattaro[a]P. EdwardFrench[b]Somya D.Mohanty[c]A sentiment analysis of U.S. local government tweets: The connection between tone and citizen involvement. https://doi.org/10.1016/j.giq.2015.03.003

[5] Tumasjan, A., Sprenger, T., Sandner, P., & Welpe, I. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *Proceedings of the International AAAI Conference on Web and social media*, *4*(1), 178-185. Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/14009

[6] Anastasia Giachanou, Fabio Crestani, Like It or Not: A Survey of Twitter Sentiment Analysis Methods, ACM Computing SurveysVolume 49Issue 2June 2017 Article No.: 28pp 1–41 https://doi.org/10.1145/2938640

[7] Adam Bermingham, Alan. F Smeaton, classifying sentiment in microblogs: is brevity an advantage? CIKM '10: Proceedings of the 19th ACM international conference on Information and knowledge managementOctober 2010 Pages 1833–1836 https://doi.org/10.1145/1871437.1871741

[8] Saif, Hassan; Fernández, Miriam; He, Yulan and Alani, Harith (2014). On stopwords, filtering and data sparsity for sentiment analysis of Twitter. In: *LREC 2014, Ninth International Conference on Language Resources and Evaluation. Proceedings.*, pp. 810–817. **http://lrec2014.lrec-conf.org/en/**

[9] Alec Go, Richa Bhayani, and Lei Huang. 2009. *Twitter Sentiment Classification Using Distant Supervision*. Technical Report. Standford. https://www-cs-faculty.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf

[10] M. Kanakaraj and R. M. R. Guddeti, "Performance analysis of Ensemble methods on Twitter sentiment analysis using NLP techniques," *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*, 2015, pp. 169-170, Doi: 10.1109/ICOSC.2015.7050801.

[11] P. Ray and A. Chakrabarti, "Twitter sentiment analysis for product review using lexicon method," *2017 International Conference on Data Management, Analytics and Innovation (ICDMAI)*, 2017, pp. 211-216, Doi: 10.1109/ICDMAI.2017.8073512.

[12] P. Balage Filho and T. Pardo, "NILC_USP: A hybrid system for sentiment analysis in twitter messages," in Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), 2013, vol. 2, pp. 568-572. https://aclanthology.org/S13-2095.pdf

[13] Kouloumpis, E., Wilson, T., & Moore, J. (2021). Twitter Sentiment Analysis: The Good the Bad and the OMG! *Proceedings of the International AAAI Conference on Web and social media*, *5*(1), 538-541. Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/14185

[14] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in social media (LSM'11)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 30–38. https://aclanthology.org/W11-0705.pdf

[15] Abd-Alrazaq A, Alhuwail D, Househ M, Hamdi M, Shah Z, Top Concerns of Tweeters During the COVID-19 Pandemic: Infoveillance Study, J Med Internet Res 2020;22(4): e19016, Doi: 10.2196/19016

[16] Boon-Itt S, Skunkan Y Public Perception of the COVID-19 Pandemic on Twitter: Sentiment Analysis and Topic Modeling Study, JMIR Public Health Surveillance 2020;6(4): e21978; Doi: 10.2196/21978

[17] Glowacki E.M., Lazard A.J., Wilcox G.B., Mackert M., Bernhardt J.M. Identifying the public's concerns and the Centers for Disease Control and Prevention's reactions during a health crisis: An analysis of a Zika live Twitter chat. Am. J. Infect. Control, 44 (12) (2016), pp. 1709-1711, 10.1016/j.ajic.2016.05.025

[18] Morin C., Bost I., Mercier A., Dozon J.-P., Atlani-Duault In**formation circulation in times of Ebola: Twitter and the sexual transmission of Ebola by survivors;** PLoS Curr., 1 (10) (2018), 10.1371/currents.outbreaks.4e35a9446b89c1b46f8308099840d48f

[19] Miyabe M., Miura A., Aramaki E. Use trend analysis of Twitter after the great east Japan earthquake. Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work Companion, Association for Computing Machinery (2012), pp. 175-178.

[20] Lachlan K., Xu Z., Hutter E., Adam R., Spence P. A little goes a long way: Serial transmission of Twitter content associated with hurricane Irma and implications for crisis communication. J. Strategy. Innov. Sustain., 14 (1) (2019), 10.33423/jsis. v14i1.984