Report On
Statistical visualization of text mined dataset (Twitter) for data entailment and to detect global COVID hotspots.

A Project Report submitted in partial fulfillment of the requirements for the award of

Bachelor of Engineering
IN
COMPUTER SCIENCE AND ENGINEERING

Submitted by
PARITOSH SINGH
(Roll no: CO17344)

Under the supervision of
Dr. Ankit Gupta
at CHANDIGARH COLLEGE OF ENGINEERING AND TECHNOLOGY (DEGREE WING), Chandigarh Sector-26, Chandigarh. PIN-160019

CHANDIGARH COLLEGE OF ENGINEERING AND TECHNOLOGY (DEGREE WING)
Government Institute under Chandigarh (UT) Administration, Affiliated to Panjab University, Chandigarh
Sector-26, Chandigarh. PIN-160019
April 2020

Department of Computer Sc. & Engineering

Acknowledgment

I would like to express my gratitude to Dr. Ankit Gupta, Assistant Professor & Training and Placement Coordinator of the CSE Department for allowing us to make a project.

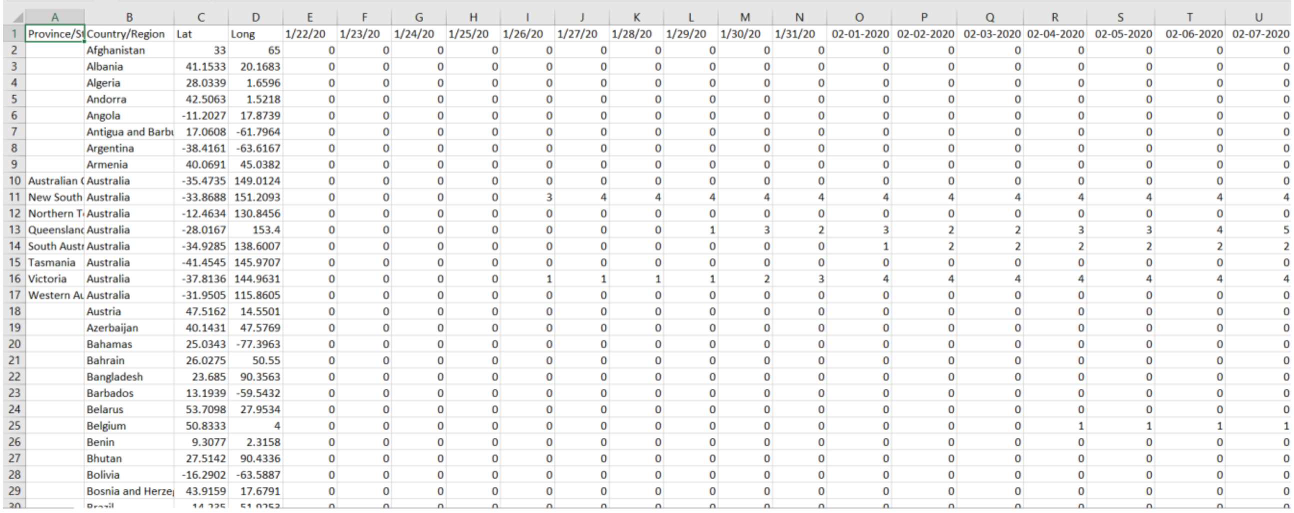
Lastly, I would also like to thank my parents who have been a constant source of inspiration and support throughout the internship and development of the project.

Contents

COVID 19 Datasets

Confirmed cases dataset:

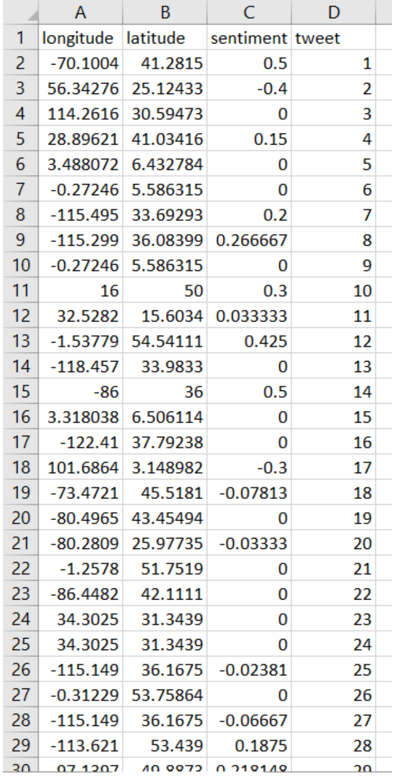
- https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Province/St	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	1/28/20	1/29/20	1/30/20	1/31/20	02-01-2020	02-02-2020	02-03-2020	02-04-2020	02-05-2020	02-06-2020	02-07-2020
2		Afghanistan	33	65	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3		Albania	41.1533	20.1683	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4		Algeria	28.0339	1.6596	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5		Andorra	42.5063	1.5218	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6		Angola	-11.2027	17.8739	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7		Antigua and Barbu	17.0608	-61.7964	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8		Argentina	-38.4161	-63.6167	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9		Armenia	40.0691	45.0382	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	Australian (	Australia	-35.4735	149.0124	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	New South	Australia	-33.8688	151.2093	0	0	0	0	3	4	4	4	4	4	4	4	4	4	4	4	4
12	Northern T	Australia	-12.4634	130.8456	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	Queenslanc	Australia	-28.0167	153.4	0	0	0	0	0	0	0	1	3	2	3	2	2	3	3	4	5
14	South Austr	Australia	-34.9285	138.6007	0	0	0	0	0	0	0	0	0	0	1	2	2	2	2	2	2
15	Tasmania	Australia	-41.4545	145.9707	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	Victoria	Australia	-37.8136	144.9631	0	0	0	0	1	1	1	1	2	3	4	4	4	4	4	4	4
17	Western Au	Australia	-31.9505	115.8605	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18		Austria	47.5162	14.5501	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19		Azerbaijan	40.1431	47.5769	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20		Bahamas	25.0343	-77.3963	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21		Bahrain	26.0275	50.55	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22		Bangladesh	23.685	90.3563	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23		Barbados	13.1939	-59.5432	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24		Belarus	53.7098	27.9534	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25		Belgium	50.8333	4	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1
26		Benin	9.3077	2.3158	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27		Bhutan	27.5142	90.4336	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28		Bolivia	-16.2902	-63.5887	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29		Bosnia and Herze	43.9159	17.6791	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30		Brazil	-14.235	-51.9253	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Sentiment Analysis dataset:

- https://ieee-dataport.org/open-access/corona-virus-covid-19-tweets-dataset

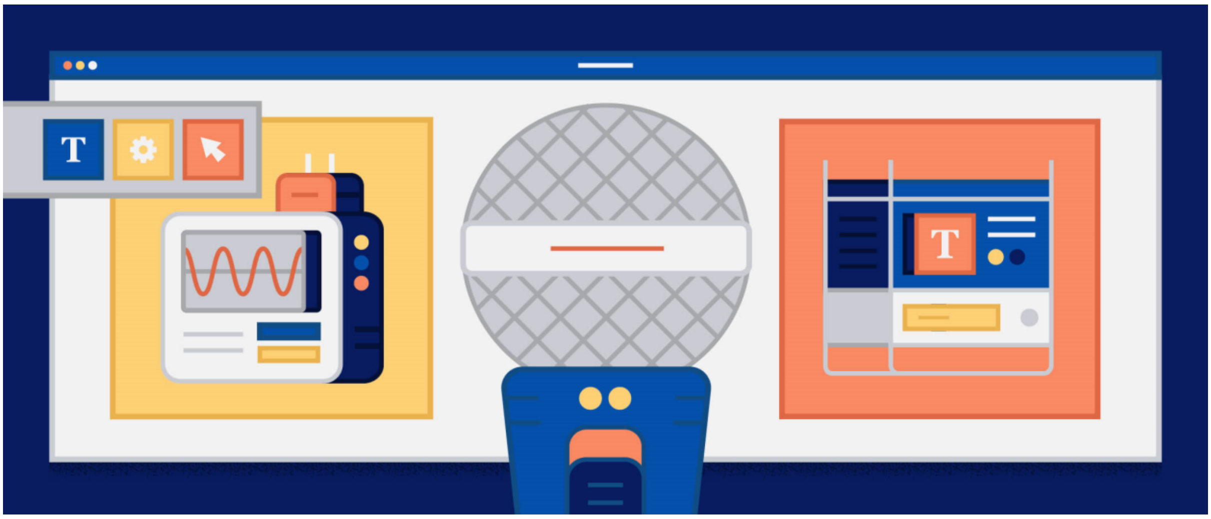


	A	B	C	D
1	longitude	latitude	sentiment	tweet
2	-70.1004	41.2815	0.5	1
3	56.34276	25.12433	-0.4	2
4	114.2616	30.59473	0	3
5	28.89621	41.03416	0.15	4
6	3.488072	6.432784	0	5
7	-0.27246	5.586315	0	6
8	-115.495	33.69293	0.2	7
9	-115.299	36.08399	0.266667	8
10	-0.27246	5.586315	0	9
11	16	50	0.3	10
12	32.5282	15.6034	0.033333	11
13	-1.53779	54.54111	0.425	12
14	-118.457	33.9833	0	13
15	-86	36	0.5	14
16	3.318038	6.506114	0	15
17	-122.41	37.79238	0	16
18	101.6864	3.148982	-0.3	17
19	-73.4721	45.5181	-0.07813	18
20	-80.4965	43.45494	0	19
21	-80.2809	25.97735	-0.03333	20
22	-1.2578	51.7519	0	21
23	-86.4482	42.1111	0	22
24	34.3025	31.3439	0	23
25	34.3025	31.3439	0	24
26	-115.149	36.1675	-0.02381	25
27	-0.31229	53.75864	0	26
28	-115.149	36.1675	-0.06667	27
29	-113.621	53.439	0.1875	28
30	97.1307	40.8873	0.218148	29

Natural language processing

Natural language processing (NLP) is a subfield of linguistics, computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyse large amounts of natural language data.

Challenges in natural language processing frequently involve speech recognition, natural language understanding, and natural language generation.



What is NLP?

01000011 01101100 01100101 01110110 01100101 01110010 01010100 01100001 01110000.

Did you get that? For those of you who can’t read binary, the direct translation is “CleverTap.”

Don’t be ashamed to admit you can’t read binary. After all, computers have difficulty understanding human speech too. When you think about the variability of the spoken word, you must consider the number of different languages, dialects, speech impediments, mispronunciations, and more.

In the English language alone, the possibilities for unique combinations of words are just shy of infinity. And with roughly 6,500 spoken languages in the world today… you do the math.

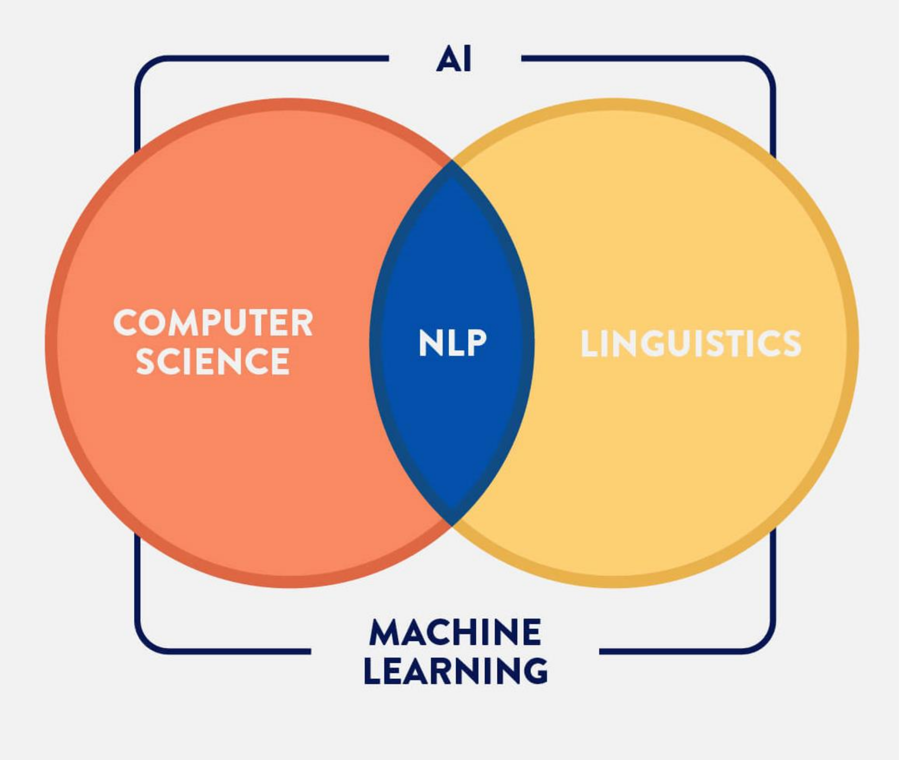
What Is Natural Language Processing?

Natural language processing (NLP) is the interdisciplinary field of computer science and linguistics, using machine learning to achieve the end goal of artificial intelligence. Simply put, it allows computers to understand human language — speech or text.

NLP is the ability to automatically receive, understand, and operate on human language in the raw written or spoken form.

Think about the communication loop between humans: a sender encodes a message through a medium (spoken or written word), and the receiver decodes the message and responds with feedback, whether it be an answer or simply an acknowledgment.

Computers must use this very same communication loop with a lot of gray area in the receiving and decoding of messages.



Why Natural Language Processing Is So Difficult

Computers are very good at processing structured data. Language, however, is about as far from “structured” as data gets.

There is a whole field of scientific study dedicated to linguistics and the attempt to make language structured. Unfortunately, in the case of real-world language, the laboratory is staffed by average people, which makes uniformity a near impossibility.

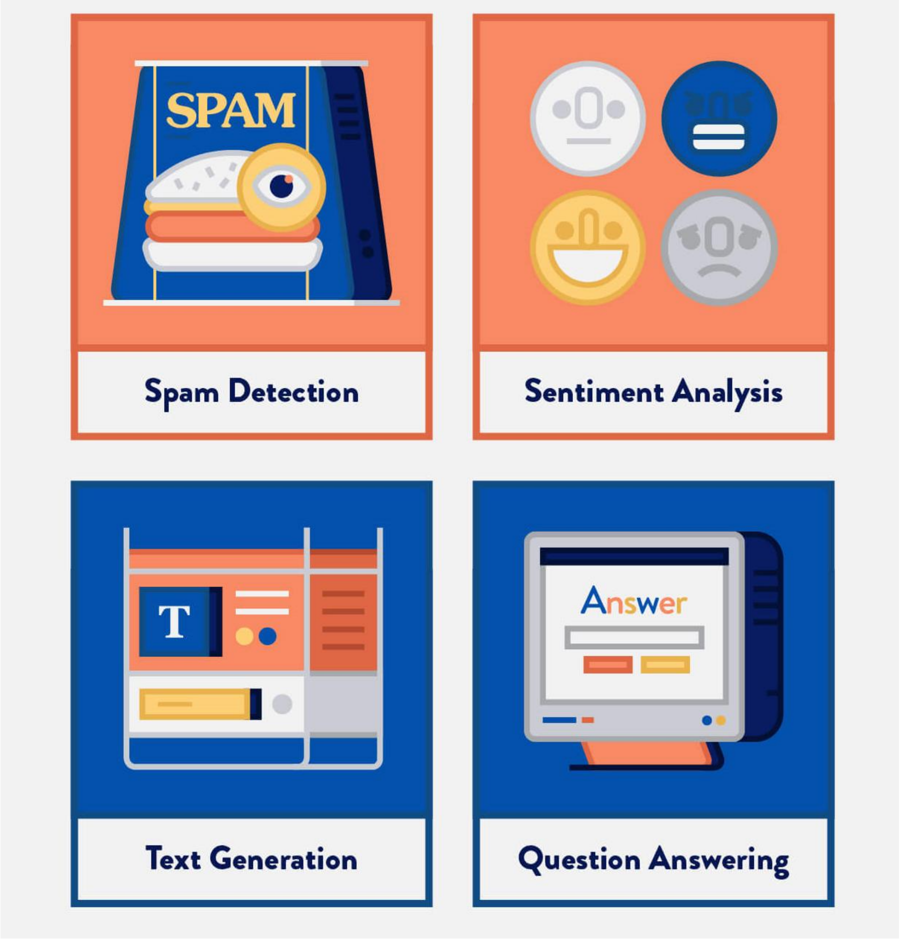
How Natural Language Processing Is Used Today

NLP is used in many different ways today, from the analysis of social media harassment to answering questions about weather forecasts and more. If you’ve ever asked Siri or Alexa a question, you have interfaced with NLP.

Some of the most basic ways in which NLP is used is for spam detection and identifying parts of speech. The spam filter on Gmail, for example, analyzes incoming emails for header information, IP addresses, and content for any signals of spam.

A more difficult use case for NLP is sentiment analysis. Analyzing the entire text for context, semantics, and pragmatics is extremely difficult. Sarcasm, for example, no matter how subtle, is understood by few readers, and even fewer computers.

As the field of study around NLP progresses, the problems being tackled have naturally increased in difficulty. OpenAI, for example, has successfully created an unsupervised model for text generation.[04] The NLP model is given a corpus of text about a given topic and is tasked with composing original prose about the subject.



Here is a summary of ways NLP can be used:

- Spam detection
- Parts of speech identification
- Sentiment analysis
- Text composition

TextBlob: Simplified Text Processing

TextBlob is a Python (2 and 3) library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.

TextBlob stands on the giant shoulders of NLTK and pattern, and plays nicely with both.

Sentiment Analysis

The **sentiment** property returns a namedtuple of the form Sentiment(polarity, subjectivity). The polarity score is a float within the range [-1.0, 1.0]. The subjectivity is a float within the range [0.0, 1.0] where 0.0 is very objective and 1.0 is very subjective.

- **sentiment**

  Return a tuple of form (polarity, subjectivity) where polarity is a float within the range [-1.0, 1.0] and subjectivity is a float within the range [0.0, 1.0] where 0.0 is very objective and 1.0 is very subjective.

Return type:	namedtuple of the form Sentiment(polarity, subjectivity)

- **sentiment_assessments**

  Return a tuple of form (polarity, subjectivity, assessments) where polarity is a float within the range [-1.0, 1.0], subjectivity is a float within the range [0.0, 1.0] where 0.0 is very objective and 1.0 is very subjective, and assessments is a list of polarity and subjectivity scores for the assessed tokens.

Return type:	namedtuple of the form ``Sentiment(polarity, subjectivity,

assessments)``

Problem Statement: Natural language processing and text mining dataset for text entailment and summarization in relational to statistical visualization.

Graphical plots and Visualizing the dataset:

COVID 19 Cases Visualization

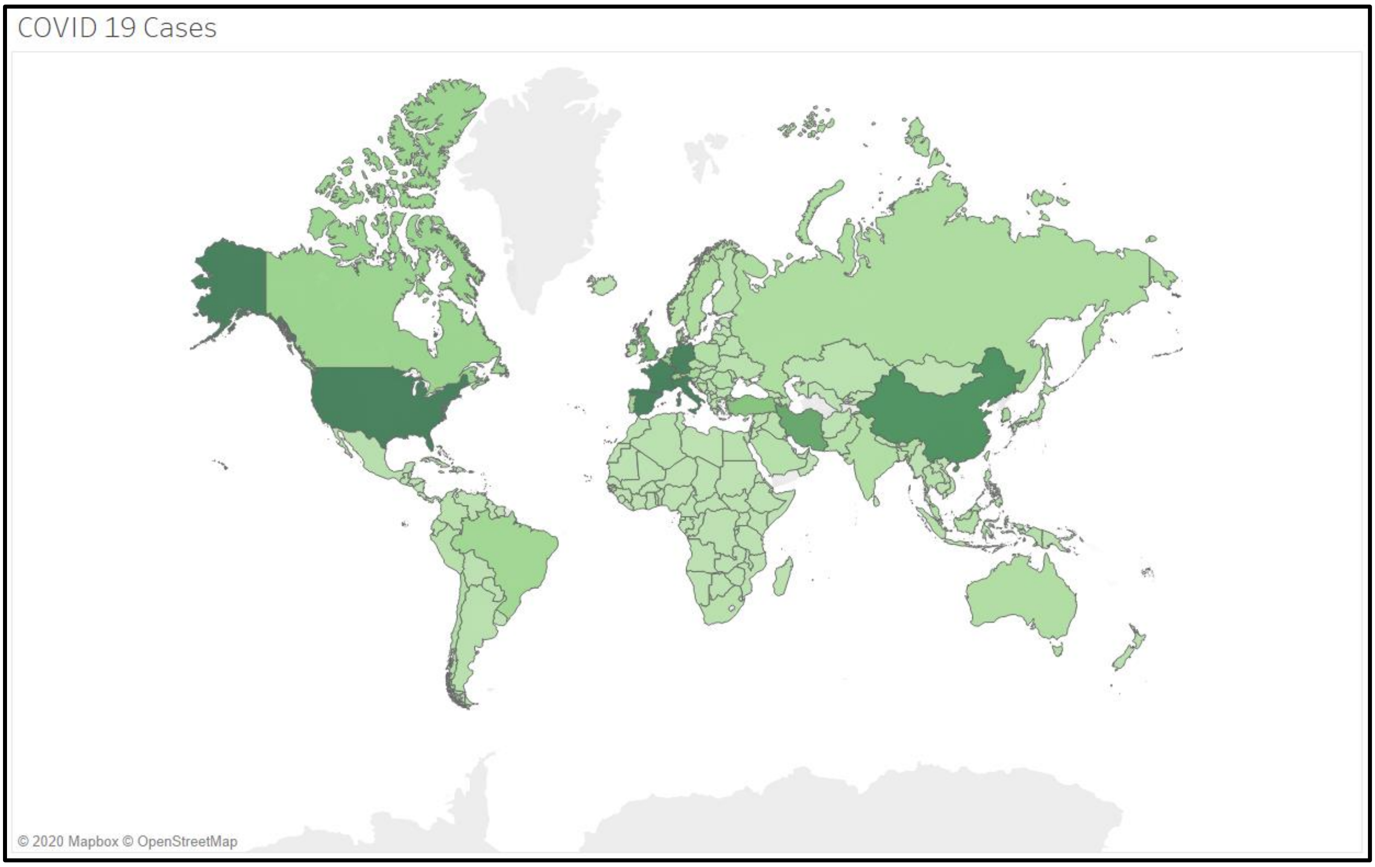
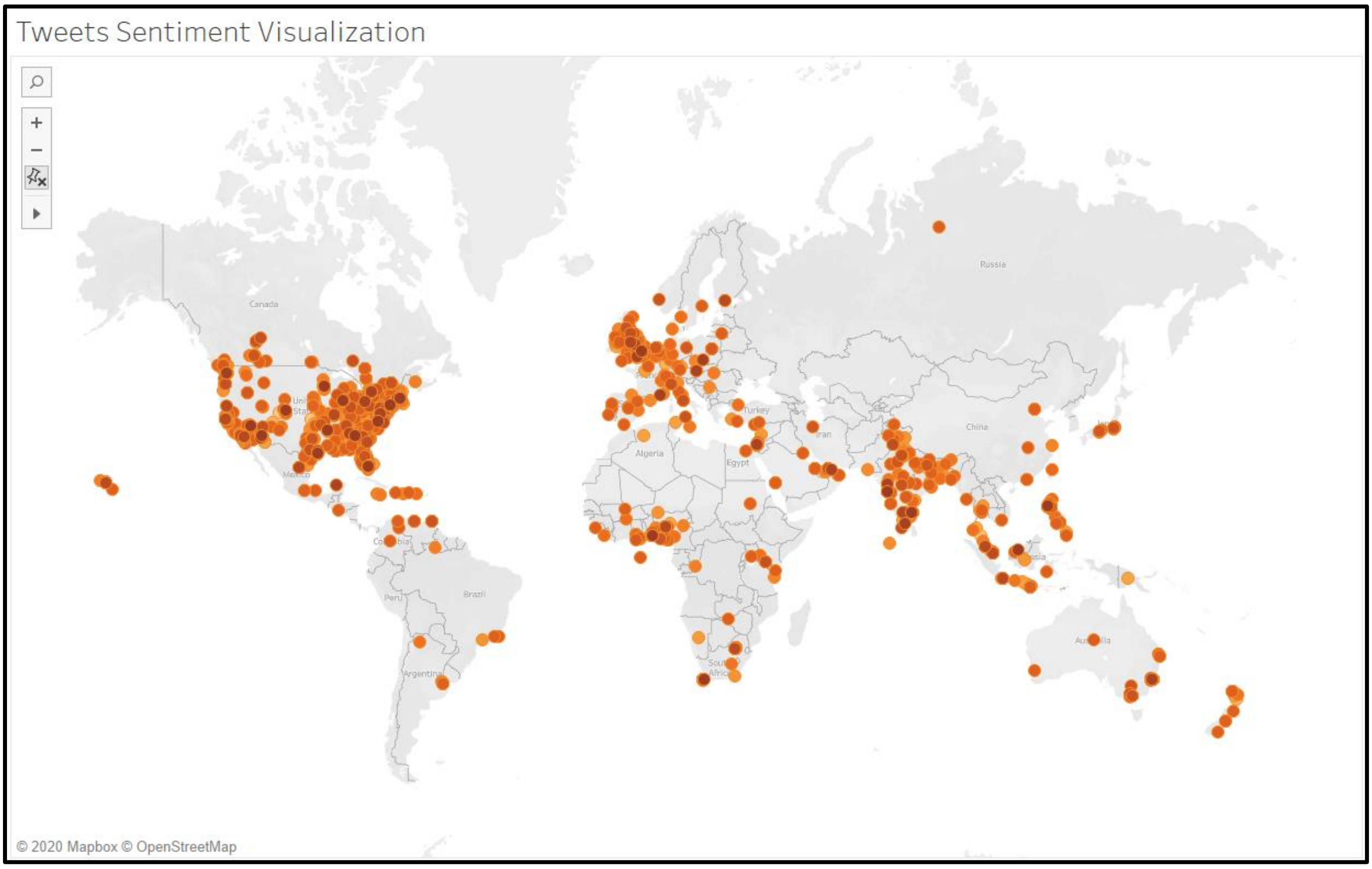


Figure : COVID 19 Cases Visualization

1. The severity of the virus throughout the globe has been mapped.
2. A world map-graph has been plotted on the global-scale to check the severity of the disease.
3. Nations in dark red show most affected nations.
4. The most affected nations are:
   a. China
   b. USA
   c. European Union
   d. Mid-West

Tweets Sentiment Visualization



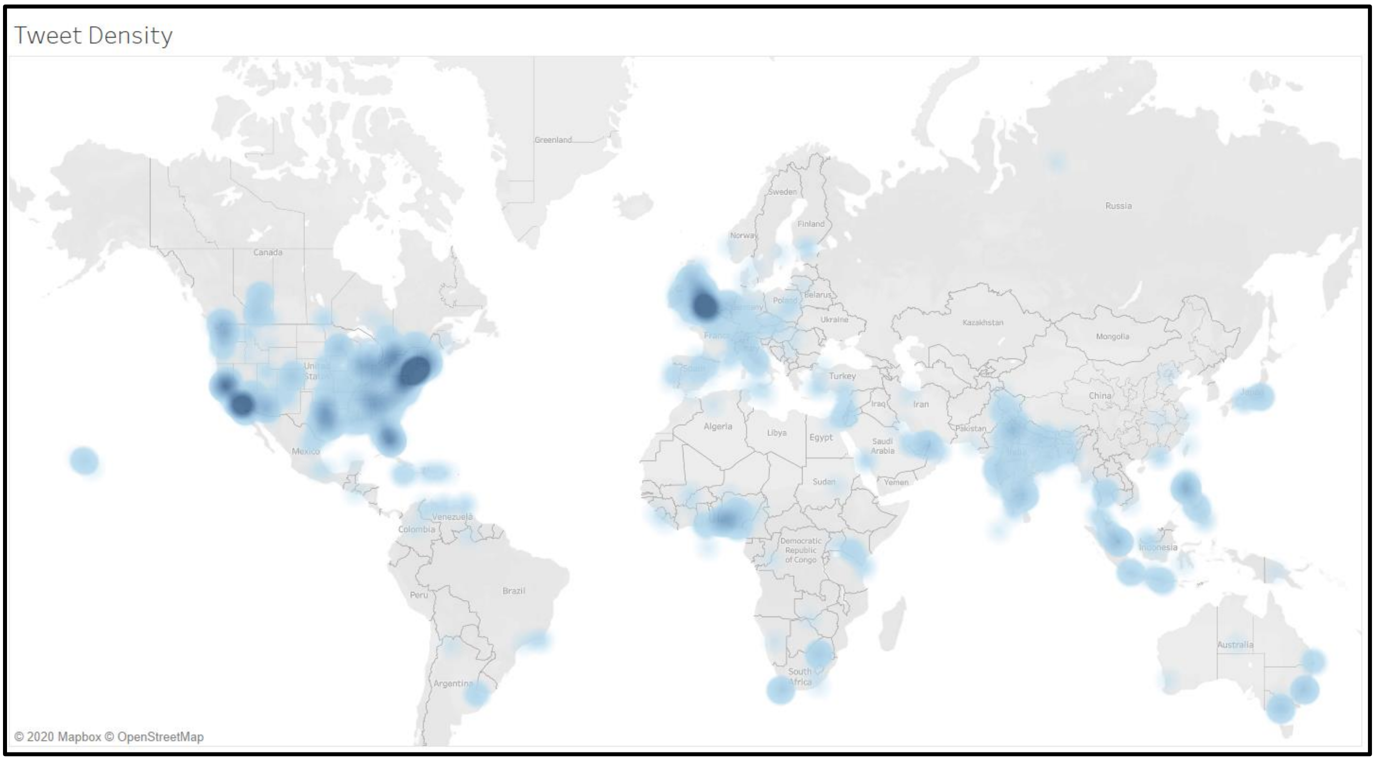
- Tweet Sentiment Visualization allows us to understand the negative or positive sentiment of tweets:

- Dark Shade: Negative Tweets
- Light Shade: Positive Tweets

- This visualization helps understand people’s opinion on COVID - 19.

- The most tweets are from nations are:
  a. USA
  b. India
  c. Europe

Tweet Density



Tweet Density describes the density of tweets about COVID 19 comes from which nations.

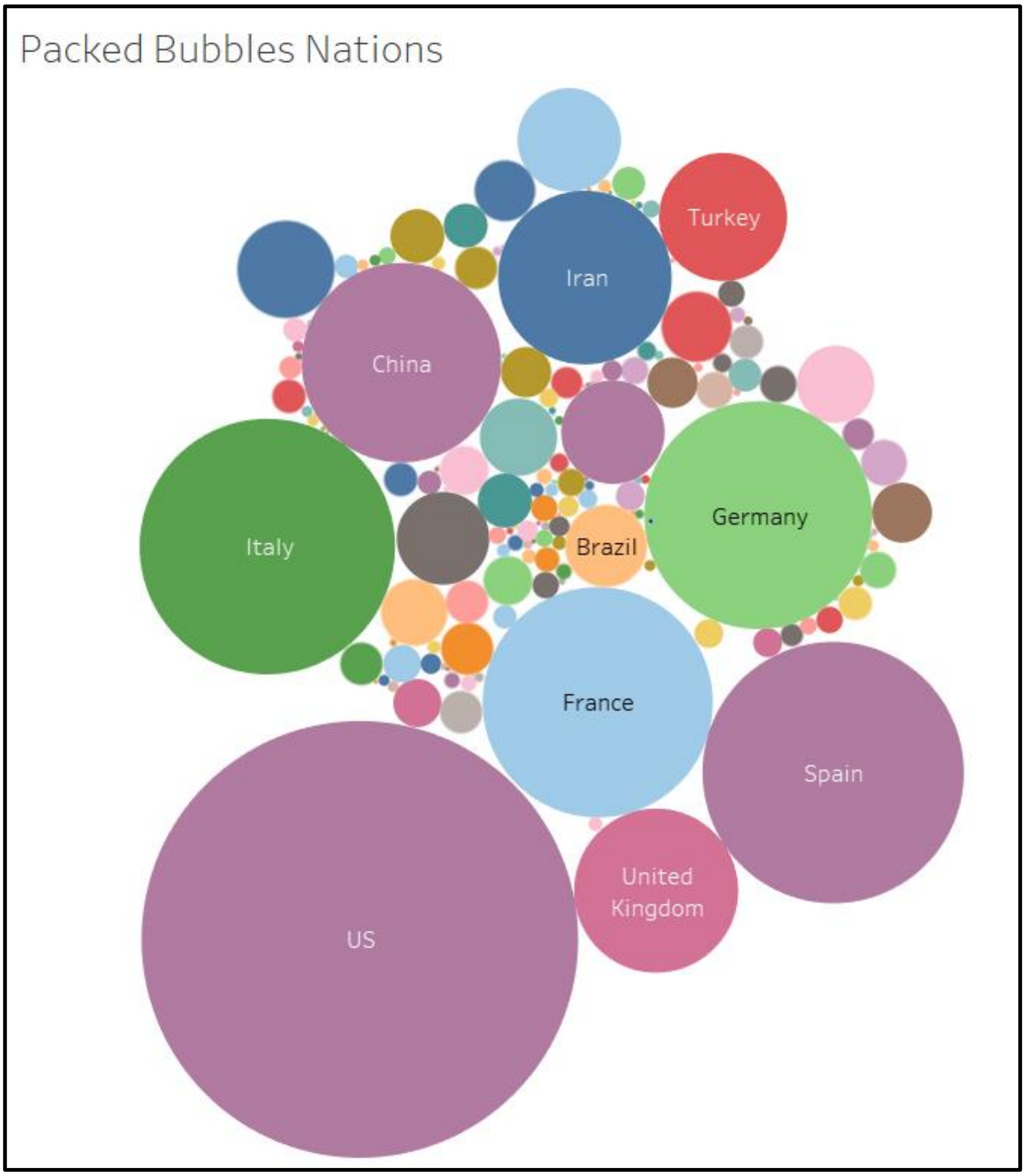
The dark blue spots show more density of tweets.
The light blue spots show sparse density of tweets.

The non-coloured spaces are due to either lack of data or no tweets.

Answers the important question do nations more affected with COVID 19 tweet more?

- The most tweets are from nations are:
  d. USA
  e. India
  f. Europe

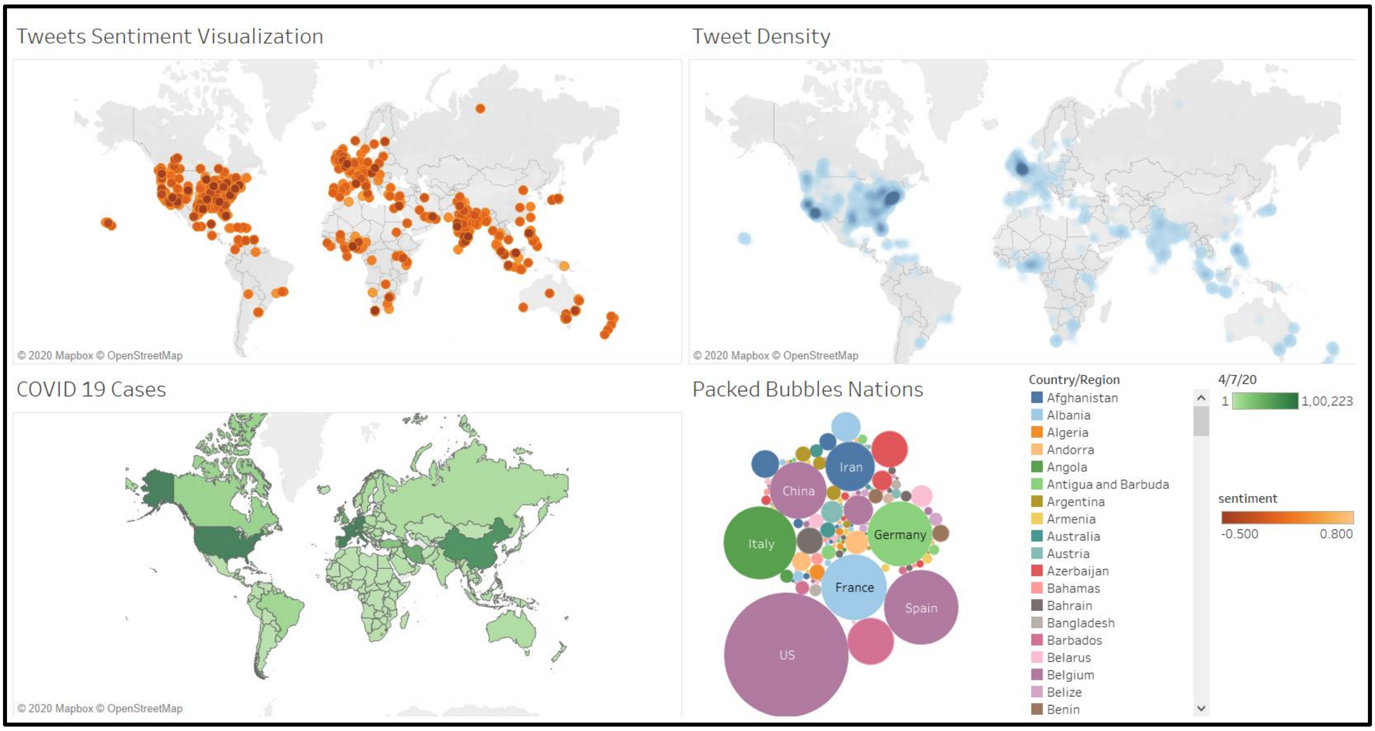
Packed Bubbles Nations



A bubble representation for nations with most confirmed cases of COVID – 19.

1. The most affected nations are:
   a. USA
   b. China
   c. European Union
   d. Mid-Wes

Dashboard



A good way to explain this dashboard is to ask the question:
Do nations more affected with COVID 19 tweet more?

We should consider the case for India:
India is not the most affected nation of COVID-19 but has substantial tweet density. This maybe probably due to high density of number of people living in India. Nothing substantial can be said as of now. Better and more comprehensive data might be needed for further conclusive answer.

But as far as the proposition: do nations more affected with COVID 19 tweet more? There is no substantial connection between tweets and COVID-19 patients.

Conclusion: Nothing conclusive can be said about the question: “Do nations more affected with COVID 19 tweet more?”.