ANALYSIS FOR NATION'S IMAGE USING TWITTER SENTIMENT ANALYSIS

Harshavardhan Gyarala
University of North Texas
Denton, Texas, USA
11437843

Sneha Sahithi Pamarthi
University of North Texas
Denton, Texas, USA
11514108

Vaishnavi Simhachalam
University of North Texas
Denton, Texas, USA
11525465

Nikhil Polisetti
University of North Texas
Denton, Texas, USA
11546616

Vashishth Sukhadiya
University of North Texas
Denton, Texas, USA
11520555

Vamsi Ande
University of North Texas
Denton, Texas, USA
11518655

ABSTRACT

Sentiment analysis using Twitter data is a powerful tool for gaining insights into consumer opinions and preferences. By applying natural language processing and machine learning techniques to social media posts, researchers can identify and extract subjective information about specific topics or events. In this abstract, we describe a method for performing sentiment analysis on Twitter data. We begin by collecting a large dataset of tweets by using keywords which are required for three research questions to estimate nation's image and pre-processing the text to remove imbalances and prepare it for analysis. Next, we apply machine learning algorithms to classify the tweets as positive, negative, or neutral. We evaluate the performance of our approach using standard metrics and compare the results to those obtained using other methods. We also discuss the challenges and limitations of this approach, such as the inherent subjectivity of sentiment and the need for large, diverse datasets. Overall, our results demonstrate the effectiveness of sentiment analysis using Twitter data and answer the research questions with different visualization charts and rate our nation's image reputation. Highlight its potential applications in a variety of fields, including marketing, politics, and public health.

GitHub link-
https://github.com/harshavardhan2204/INFO-5810-Section-001-Project-Group-8

KEYWORDS

Recession, Mental Health, Anti-National, tweepy, twitter, cluster analysis, sentiment analysis, python, regex, pandas, rapid miner, employment,president,USA,depression, pie-chart, histogram, visualization, positive.

ASIS&TTHESAURUS

data science, Big Data, knowledge&information, datamining

1 INTRODUCTION

In this modern era, technology plays a vital role and is the key factor in producing lots of data. Now a day's internet has become the main communication platform from where people express their ideas and opinions. Social media like Twitter, Facebook, and Google are allowing people to connect with each other. These are sites where people share their ideas, express their views, talk about public issues, and a lot more than this. Some of them are useful and some are not. These sites allow people to communicate around the world.

Using sentiment analysis algorithms, social media data may be automatically examined to detect the intensity of expressed opinions. These algorithms have evolved over the past few years to look at more variables,

such as a person's sentiment toward a topic or their emotions and have even merged text analytics with other inputs like multimedia analysis or social network analysis. Sentiment analysis is contextual text mining that recognizes and extracts subjective information from the source material. Keeping an eye on online discussions enables businesses to identify the social sentiment surrounding their brand, product, or service. Recent developments in machine learning have greatly enhanced text-mining methods. The inventive application of cutting-edge artificial intelligence methods can be a useful instrument for doing in-depth study.

We have applied sentiment analysis to the data through the Twitter Platform. We have obtained various kinds of data through the Twitter API developer account by using the Python programming language. Based on the words in the context, we have classified the tweets and tried to decide the emotion of the person. All the references, methodology, data collection, and data analysis are done and explained below. For this research, we used Python programming language and rapid miner to analyze the data. We have used Python to get the data from platforms like Twitter and used Rapid Miner to implement a few machine learning algorithms as we can obtain accurate results compared to previous research.

Research questions:

1. Is recession really a factor of unemployment and inflation which affects the nation?
2. How many people in the nation are paying attention to their mental health and taking action to get out of their current conditions?
3. A country's reputation is greatly influenced by the percentage of anti-nationals living there. What proportion do they control?

Research Purpose:

The goal of this research is to analyze text data using sentiment analysis through naive Bayes and other machine learning algorithms. We also want to research how people’s mental health has changed. And our goal is to find to what extent social networking platforms such as Twitter influence current society including politics and decision-making of people. In this research, we can learn about people's sentiments and opinions regarding policies, goods, brands, and other related topics by gathering tweets and looking at the individual’s emotions expressed in them.

Data collection strategy:

We used python for extracting tweets from twitter api. As we didn’t have access we could only extract a week long tweets. Used several hashtags to extract tweets and stored them in a csv.

2 RELATED WORKS

Understanding the consumers' attitudes regarding health insurances is the focus, "Using Social Media to Identify Consumers' Sentiments Toward Attributes of Health Insurance During Enrollment Season." They used the NRC Emotion Lexicon, which offers each word's polarity as well as its accompanying emotion (anger, anticipation, disgust, fear, joy, sadness, surprise, and trust), to mine Twitter chats and analyze them using a dictionary-based technique. This study's key finding is that political preferences, prescription drug benefits, and provider networks worry consumers. Customers also have faith in medical professionals but worry about unforeseen circumstances. These findings imply that additional study is required to identify the root causes of consumer motivations in order to improve insurance products.

The Novel Coronavirus disease 2019 caused by the Sars-Cov-2 virus Has become a pandemic with a growing number of cases globally. In this duration, the people have to face challenges, although primarily infection disease with physical health implications, also affect mental health and wellbeing. Studies said that depression, anxiety disorders, substance abuse, increased suicidal tendencies, and PSTD. Twitter is the biggest social media platform for hosting an abundant number of user-generated posts. It is considered as a gold mine of data. At that time people tweeted aggressively on twitter. Most of the people started having negative tweets but with increasing time people shifted towards positive and neutral comments. In April 2020 most comments were positive and about winning against CoronaVirus. The tweets have been collected, pre-processed, and then used for text mining and sentiment analysis. The results of the study concludes that while the majority of the people throughout the world are taking a positive and hopeful approach, there are instances of fear, sadness and disgust exhibited worldwide.

The severity of both the new coronavirus outbreak has resulted in significant financial challenges, worry, anxiety, and future concerns. Social media can offer a platform for tracking the state of people’s mental well-being in local areas. Examining the evolving speech use on social networking sites can support conventional survey-based methodologies and offer fresh perspectives on the

condition of a country or region throughout a public-health emergency.

Twitter posts can reveal shifts in a community's psychological health amid a health emergency when mass polling could not be possible. Anxiety, stress, and solitude have diverged more from readings in 2019. Early detection of areas where psychological health is deteriorating can result in society treatments.

The fact that Twitter accounts must not include all demographic groups as well as the language-based calculations are dependent on a randomized 1% digital signal of tweets are both limitations of this study. Our projections haven't been verified during the assessment period during the assessment period due to a lack of public polling. We plan to verify these models with industry-recognized polling data in subsequent research.

In summation, real-time analysis of posts made on destination social media platforms can provide light on recently discovered public concerns. Early detection of regional trends can help with resource allocation, focused public health initiatives, and improved readiness for the current and upcoming public health catastrophes.

Recent studies in sociologists and other disciplines, such as medicine and entertainment, have benefited from sentiment classification, which reveals developments in human emotions, especially via social media. The approaches used for trend analysis on twitter data are the same as those used for other domains. Future research might study the application of these strategies to rich language research, such as pattern recognition and disambiguation.

The spreading of pandemic influenza makes people vulnerable and anxious on a worldwide scale. This type of situation considers conversation essential when developing better preventive actions because it does not alter the following of specific cut-off dates.

The main goal of this study is to reduce the knowledge gap that exists across data analysts and users in analyzing public behavior to combat the plague. The key elements for risk minimization and general anxiety appraisal are swift response from public health professionals and accurate data from organizations.

Sentiment classification is quite helpful in determining how people felt about a certain occurrence. Positive sentimental tweets about the epidemic are almost certain, indicating that people maintained their confidence amid a remarkable public health calamity. Positive connotations typically expressed gratitude for grassroots initiatives and frontline personnel who support humanity's most vulnerable ones.

Mainstream media exerts information hegemony by controlling what data is available to individuals and how people perceive certain situations, despite the knowledge that the internet is a significant instrument for disseminating. One of most frequently used terms on tweets were discovered using both unigrams and bigrams evaluation.

Results of confinement and similar prevention programs regarding mental health have shown that those who experienced physical isolation were more likely to experience depression, mental illnesses, post traumatic stress symptoms, confusion, and frustration due to stressors like extended confinement, fear of infestation, frustration, boredom, insufficient supplies, insufficient information, financial loss, and stigma. In many circumstances, COVID-19 has also altered fragile people's suicide behaviors. Numerous case examples of individuals who took their lives out of fear of COVID-19 are discovered based on mainstream media stories.

People who received superior psychological and interpersonal support from their interpersonal and familiar networks throughout COVID-19 experienced chronic stress during that time with less of an effect. Most of the research on COVID-19 and psychological wellbeing is merge, which could be informative in the gradual shifts in behavioural health results among the people that are impacted even though some research used comparable controls or assessed against previous mental health state of the individuals longitudinal study methods must be used to examine such adjustments.

Moreover, such research should assess the contribution of various risk variables to psycho pathogenesis in various groups in many situations over time. All sectors faced distinct issues because of COVID-19. Including social, religious, academic, vocational, economic, political, and other aspects, health, and wellbeing of human lives. The findings show that this phenomenon exists of this evaluation as well numerous elements connected with concerns with mental wellbeing suggest such those ought to be investigated.

Utilizing transdisciplinary methods, researchers from diverse fields contribute their distinctive views and resources to comprehend the socioeconomic systems that are the target of similar research endeavors and

participants in developing mental wellbeing should be included in the study of COVID-19 and related destabilizing global politics.

3 METHODOLOGY

Python is used to analyze the sentiment of user-generated tweets. Effective libraries for Python include Tweepy and TextBlob.

The official Python application for the Twitter API is called Tweepy, and it allows users to access Twitter using both OAuth and the more recent Basic Authentication mechanism. OAuth is the sole method available to access the Twitter API as Twitter no longer accepts Basic Authentication. It provides access to the thoroughly explained Twitter API. With Tweepy, you can obtain an object and employ every technique that the verified Twitter API provides. Tweets, Users, Entities, and Places are the primary Model classes in the Twitter API. In Python, it's incredibly simple to navigate through data because each access produces a response in JSON format. Using tweepy we can retrieve tweets in our wanted time frame and context.

One of the most used NLTK-based Python modules for analyzing text data is TextBlob. Almost all actions required for Basic NLP can be accomplished using it as a framework (Natural Language Processing). Sentiment Extraction and Spelling Correction are two of the more sophisticated aspects of TextBlob.

Tokenization via TextBlob allows the text blocks to be broken up into various phrases and clauses. Reading between the lines is significantly simpler as a result. In phrases, the noun is typically employed as an Entity. Additionally, it is among the most crucial NLP tools for dependency parsing. Using TextBlob, several nouns are taken out of a sentence. TextBlob can be used to tag sentences with various components of speech.

Acquiring data relating to research questions can be achieved by following methods. Large data is always imbalanced. So, Data cleaning and preprocessing plays an important role.

Using TextBlob we can filter our tweets related to covid and can acquire data related to the research. We can perform operations on this data in Rapidminer which is very essential to analyze Big data. Sentiment Analysis is the platform technique we are going to perform on all the datasets which we acquire and a classifier named Naïve Bayes is performed on this type of data to produce results. Will provide some labels relating to depression which are termed negative in sentiment analysis. Using Tokenization we can label those features and perform naïve bayes before covid timeframe and after covid timeframe. Will compare both the results and can answer our research question did covid affect people mentally or not.

A large weightage for Nation's image analysis can be given to country people reacting for social purposes. Previously build algorithms will be used to acquire related data for this research statement. User tweets who have text label features relating to human rights, environment, gender equality, religious freedom, property rights, trustworthy, well-distributed political power, racial equity, cares about animal rights, committed to climate goals and fight for justice will be Tokenized and pass through Naïve Bayes classifier to predict the output. Predicted output analysis can be done using Rapidminer and visualize the output data. If there are more positive quoted tweets then it will add huge weightage for Country's reputation.

Happy or Sad seems simple in terms of words but it is difficult to extract one's emotion behind a few lines of sentence. Happiness of the people is rated very high and resembles Happy people is equal to Happy country. This research question can predict and give accurate solutions for our research. Above used algorithms will be used to obtain required data. Using nltk we can preprocess data by removing special characters and stopwords, TFIDF vectorization is used to transform the data into a sparse matrix format. Cosine similarity function can be used to detect similar tweets regarding one's emotion and can be labeled into two vectors passing happy and sad tweets of people. We will verify Nation's Happy percentage using Data visualization features from Rapid Miner using Bar charts, Histogram and mosaic graph. Accurate output will be obtained by proceeding with the proposed methodology and can differentiate people's happiness rate before and after pandemic too.

DATA COLLECTION AND CLEANING:

Twitter provides a device with the means to collect data through its API (API). The streaming mechanism collects the input data from Tweets and performs any mandatory analyzation, percolation, or aggregation before adding the

results to a data repository. The HTTP dealing mechanism asks the storage service for the user's query's response.

HTTP uses GET method requests and, in contrast to ATOM, can produce results that have been modified using Java Structural Object Syntax. Python scripts were created to communicate with the streaming API and get associated data. These scripts gather information based on keywords and output to different entities, such as dates, location, race, languages, hyperlinks, text, and so on.

DATA COLLECTION:

Twitter is a largest social media platform to get millions of data but to acquire all the desired tweets one should have a Twitter API developer account. We can extract tweets using python library tweepy.

There are authentication keys and tokens to perform the API. Below you can find our snippet used to collect data. We used the Twitter API Elevated account, which allows us to access up to 2 million tweets from the current date, to gather data from Twitter. But data extraction takes a long time. Data on recission, mental health, depression, and data for the United States have all been gathered. Although we applied for an academic search Twitter account, it was denied. so have collected the recent data using elevated account tokens. We used tweepy library to collect our data. We pulled 1,000 tweets every 20 minutes during this time-consuming procedure, and we have 1.5–2 million tweets for our project.

```
In [2]: # Imports the needed libraries
        # You may need to install tweepy
        import tweepy
        import pandas as pd
        import time
        consumer_key = 'hk8jsV6WPXsMnXi4mwzroh81e'
        consumer_secret = 'tu46FM1pNyTP0V0Lu8N26zyz7VeXE1Vh1S8r0jnHJ3HWATPxi0'
        access_token = '1540784036074889217-whwAj71uCzpGVrMGDjx50QSt7aI1xa'
        access_token_secret = '4kpZBat8zqXUXZcSRWGlm5OeiSKAeXxrbaSwvZ3vwIUvn'
        auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
        auth.set_access_token(access_token, access_token_secret)
        api = tweepy.API(auth,wait_on_rate_limit=True)

In [10]: #for particular geo location
         places = api.geo_search(query="United States", granularity="country")
         place_id = places[0].id

In [9]: #keywords- Unitedstates,#USA,#mentalhealth,#wellbeing,#depressed,#Covid19,#recession,#layoffs etc.
        search_words = "#USA"

In [5]: created_at = []
        text = []

        Enter number of tweets you want to pull in items()

In [6]: for tweet in tweepy.Cursor(api.search,q= search_words,count=100,
                                   lang="en",
                                   since="2017-04-03").items(10):
            print (tweet.created_at, tweet.text)
            created_at.append(tweet.created_at)
            text.append(tweet.text)
```

Fig- Code snippet for Collecting data.

As it is Big data extraction took several days for collecting tweets.

Below are the search words used to collect the data from twitter.

“#USA”,”#Unitedstates”,"#covid19","#usa", "#justicearrived", "#justicedenied", "#saveplanet", "#saveearth", "#pollution", "#saveanimals", "#govegan", "#animaltesting", "#depressed", "#anxiety", "#selfcare", "#goingstrong", "#mentallystrong", "#emotionalstability", "#happy", "#sad", "#feelinglow", "#tired", "#feelingoverwhelmed", "#goodday", "#blessed", "#feelingproud", "#covid", "#lonely", "#fighter", "#covid19", "#vaccine", "#lockdown", "#positive", "#negative", "#quarantine", "#omicron", "#secondwave", "#travelrestrictions",”#recession”,”layoffs”,”inflation”.

With these search words we collected total of 2-2.5 million of tweets and separated into three csv files named mentalhealth, recession and general tweets of usa people.

DATA CLEANING:

Gathering millions of samples could lead to imbalanced data and biased data. We need to first check if there are any null values in the data and drop it if there are any.

We can import libraries like regex in python and apply to the data to remove any special characters Like #,@,! and emojis if any. Converting all the text into lower case letters is important to tackle imbalancement. Splitting the text into a single word will make our work much easier thankfully regex function has that feature to split into words.

Using NLTK we can download “stopwords” and can be applied to the data. To have unique words so that it does not affect the outcome which acts as an outlier for our research.

Below you can find a sample code-snippet which can be used for the data cleaning process.

```
In [1]:
import pandas as pd
import nltk
nltk.download('stopwords')
import re
from nltk.corpus import stopwords

[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.

In [4]:
df=pd.read_csv("uintedstates.csv")
df

In [5]:
X1=df['tweet']


# 1. Remove non-letters
X1 = X1.apply(lambda x: (re.sub("[^a-zA-Z]", " ", str(x)) ))
#X1= X1.apply(lambda x : re.sub(r'[^\w\s]', '',x))


# 2. Convert to lower case, split into individual words
X1= X1.apply(lambda x : str(x).lower())
words=X1.apply(lambda x: x.split())


#3. create stopwords using NLP
stops = set(stopwords.words("english"))


# 4. removed stop words from the dataset
meaningful_words = pd.Series(words).apply(lambda x :  [item for item in x if item not in stops])
meaningful_words=meaningful_words.str.join(" ")

In [6]:
meaningful_words

Out[6]: 0    rt diagolonmonkey interesting case made usa su...
        1    ivankatrump nothing says american patriotism c...
        2    rt theeverearn usa vs england draw everearn pa...
```

Fig- Code snippet for cleaning text data in python

Remaining data cleaning for replacing missing values and duplicate samples was done in rapid miner

EXPLAROTARY DATA ANALYSIS:

1. **Defining The Problem:**

   Identifying the problem is the first stage in any data analysis procedure. This is sometimes referred to as the "Problem Statement" in the context of data analytics.
   Creating a hypothesis and planning how to test it is necessary to define the purpose. For our research on estimating a country reputation we have created certain situations which can measure the reputation of the country. We proposed three situations which are recession and how people reacted with this situation, other is mental health which is a very important factor for any nation. And the last one includes anti national people. We gave keywords related to anti national and just wanted to make sure how many anti nationals are there because that effect the nation's image.

2. Collecting the Data:

   Once your objective has been established, you must devise a strategy for collecting and aggregating the required data. A key part of this is selecting the data you need.
   We are collecting tweets so we are expecting textual data in form of strings to collect.
   We give Keywords and search words which are required to extract tweets with those keywords in geolocation-"United States". For example, we have a situation of recission so we give search words as #recession,#layoffs,#inflation etc. We have situation for mental health so we give search words as #depression,#wellbeing,#mentalhealth etc. And lastly we have situation for anti national in united states but we just wanted to extract search words as #usa, #unitedstates .

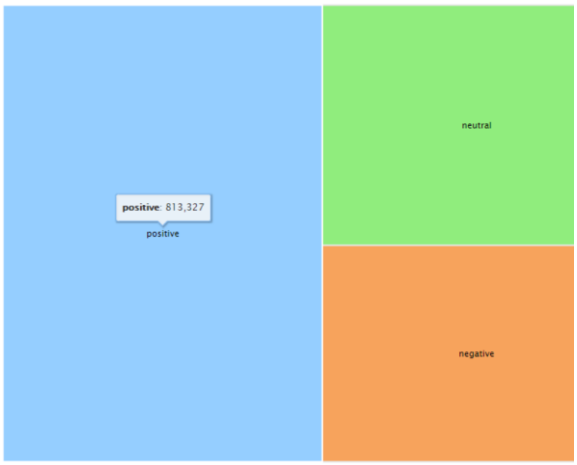
   For our research we need data at least a million tweets but we collected almost 2-2.5 million tweets for this research to be more accurate. As data quantity is directly proportional to accurate outputs.

3. **Cleaning the Data:**

   It can be difficult to manually clean up large databases. Fortunately, we have a wide range of tools at our disposal to speed up the procedure. Like tweepy in python which is used to extract tweets from twitter api. But first you need an api account to use it. Code in python to extract tweets which we already mentioned in methodology. And clean using regex function which is a regular expression in a search pattern which is used to search a certain character in strings and remove them. We used replace missing values operator in rapid miner and remove duplicates operator to clean up our data.

4. **Analyzing the Data:**
   Once the Data cleaning is done the next step is Analyzing the data. Your goal will have a significant impact on the type of data analysis you conduct.
   The goal of diagnostic analytics is to determine why something occurred. On the basis of past data, predictive analysis enables you to pinpoint potential trends. Predictive analysis is frequently

used in business to make predictions about the future, like future growth. With prescriptive analysis, you can suggest changes for the future. The process' final step in the analytics section is this
When analyzing data for #usa and #unitedstates. We ran sentiment analysis without applying filters. Majority of their score was positive which is a good thing. Total data= 1.6 million tweets



When analysed data of recission used #recission, #layoffs to pull data and performed sentiment analysis on this dataset. The below is the word chart of the sentiment analysis data and most of them are positive.Total data = 0.6 million tweets

5. **Sharing the Results**:
After the analysis the insights will be collected and the last step is to process these with the organizations or clients which can be complex since it will provide the outcomes. Sometimes we try to present these with the help of decision makers. How we present the results will influence the flow of the business. So that the organization can analyze and predict the risk factors or the things that are effecting the business. And it is crucial to find out if there are any gaps or flags that are needed to be considered in the business. There are various tools available to help to share the results and visualize them like Google Charts, Tableau, Datawrapper and Infogram. And in python we can use libraries like Plotly, Seaborn and Matplotlib.

6. **Embracing the Failures:**
The final step is Finding the failures and understanding.Because data analytics is inherently chaotic, each project will require a distinct approach. For example, while cleaning data, you can see patterns that prompt an entirely new set of inquiries. You might then need to go to step one. In addition, an exploratory analysis may bring to light a group of data points you hadn't previously considered employing. Or perhaps you discover that your key analyses' conclusions are inaccurate or misleading. This could be the result of data errors or human errors made earlier on in the process.

4 RESULTS AND DISCUSSION

i) Results on recission data. On how recession has effected people in united states are they talking about it positively or negatively. Because their sentiment will analyse the situation has taken a positive impact or not for estimating Nation's image.

Sentiment score of rescission data when selected keyword rescission. And tells maximum number of people reacted negatively which we can say it had negative impact on them.

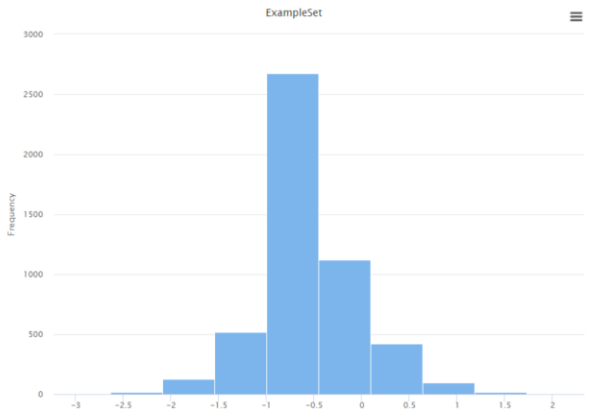


Fig- Sentiment score of people talking about recession in histogram data.

Sentiment score of recission data when selected keyword for layoff. As we can see most people was talking positive in their tweets which we can say this situation was handled pretty well in this nation.

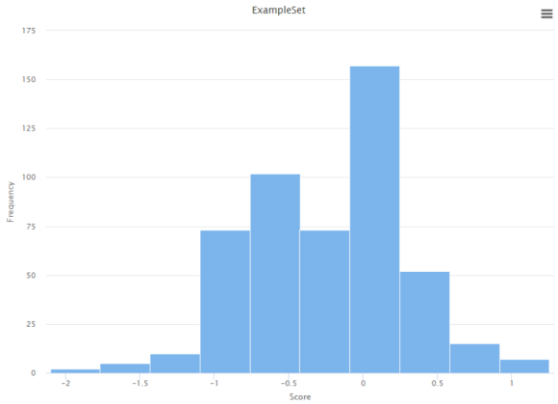


Fig- Sentiment score of people talking about layoffs in histogram data.

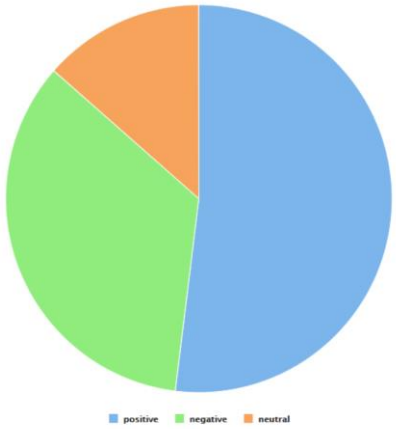


Fig- Sentiment of people talking about layoffs in pie chart.

Sentiment score of recission data when selected keyword for inflation. As we can see most people was talking negative in their tweets which we can say this situation had impacted after the recission as all the prices were high including gas, eggs etc.

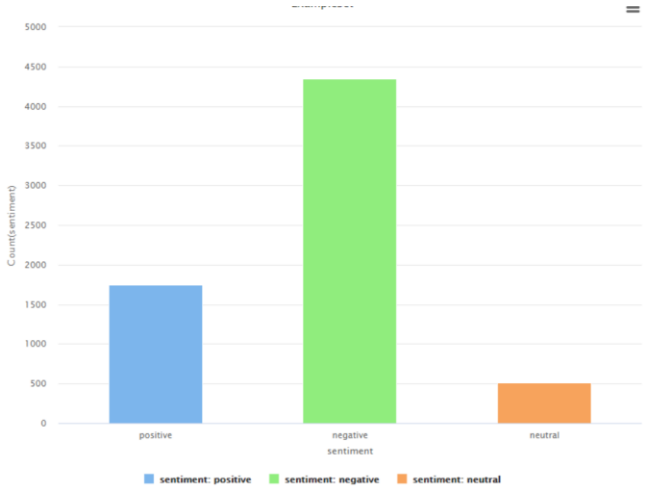


Fig- Bar chart representation of people reacting to layoffs

If unemployment is the keyword in recission data, we can see highly negative tweets for this people nation. We can say they are heavily impacted with these which effects the nation reputation

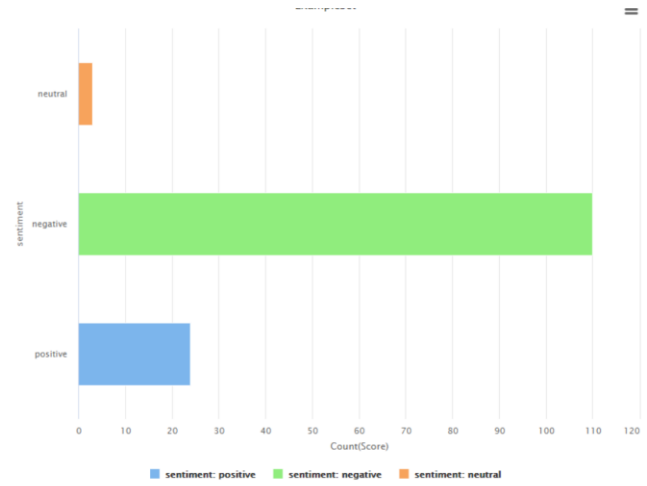


Fig- Vertical Bar graph of sentiment when unemployment is keyword

Sentiment of people when talking about jobs, and their effect of job in recession dataset has large people talking positive about it. Which is plus point for country reputation. Jobs play an important factor for people in a country.

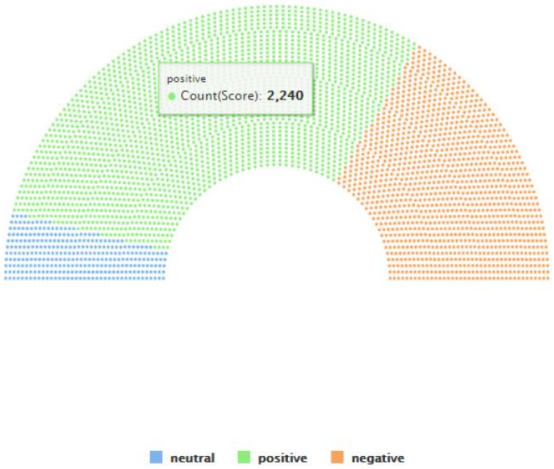
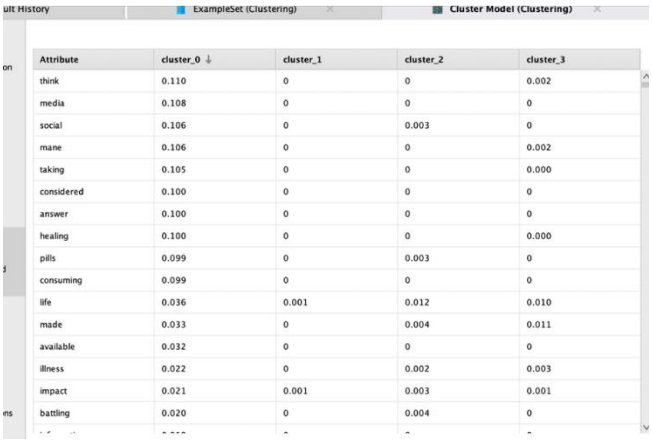


Fig- Parliament chart of people's sentiment talking about Jobs.

ii)Results on mental health data. Mental health of people play a crucial part in their lives, people are in good condition means a country is in good condition. We worked on this dataset of people in USA.

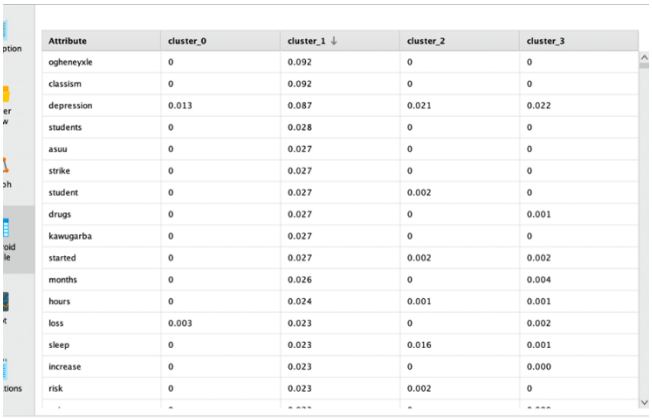
Taking care of your emotional, psychological, and social wellbeing is part of maintaining good mental health. This can involve things like obtaining regular physical activity, eating a nutritious diet, getting adequate sleep, abstaining from drugs and alcohol, and using relaxing methods like meditation or deep breathing. It may also entail participating in activities that make you happy and fulfilled as well as asking friends, family, social media or mental health specialists for assistance when necessary. Taking care of your mental health can increase your resilience and general well-being by making you feel better overall.

The terms "healing," "taking," "thinking," and "battling" in the screenshot below indicate that people are attempting to improve their condition.

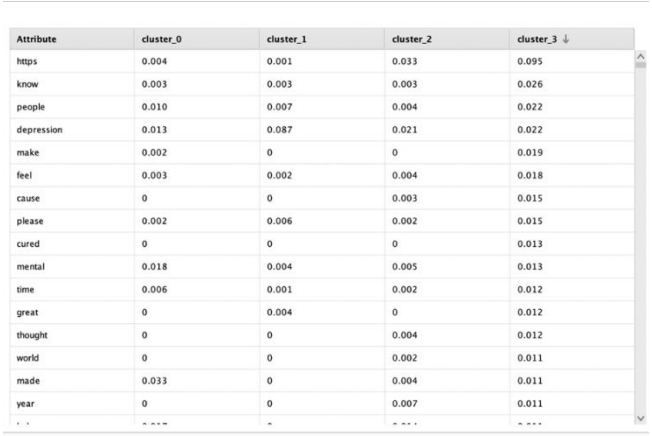


Attribute	cluster_0	cluster_1	cluster_2	cluster_3
think	0.110	0	0	0.002
media	0.108	0	0	0
social	0.106	0	0.003	0
mane	0.106	0	0	0.002
taking	0.105	0	0	0.000
considered	0.100	0	0	0
answer	0.100	0	0	0
healing	0.100	0	0	0.000
pills	0.099	0	0.003	0
consuming	0.099	0	0	0
life	0.036	0.001	0.012	0.010
made	0.033	0	0.004	0.011
available	0.032	0	0	0
illness	0.022	0	0.002	0.003
impact	0.021	0.001	0.003	0.001
battling	0.020	0	0.004	0

Fig- Cluster centroid table of mental health data



Attribute	cluster_0	cluster_1	cluster_2	cluster_3
ogheneyxle	0	0.092	0	0
classism	0	0.092	0	0
depression	0.013	0.087	0.021	0.022
students	0	0.028	0	0
asuu	0	0.027	0	0
strike	0	0.027	0	0
student	0	0.027	0.002	0
drugs	0	0.027	0	0.001
kawugarba	0	0.027	0	0
started	0	0.027	0.002	0.002
months	0	0.026	0	0.004
hours	0	0.024	0.001	0.001
loss	0.003	0.023	0	0.002
sleep	0	0.023	0.016	0.001
increase	0	0.023	0	0.000
risk	0	0.023	0.002	0

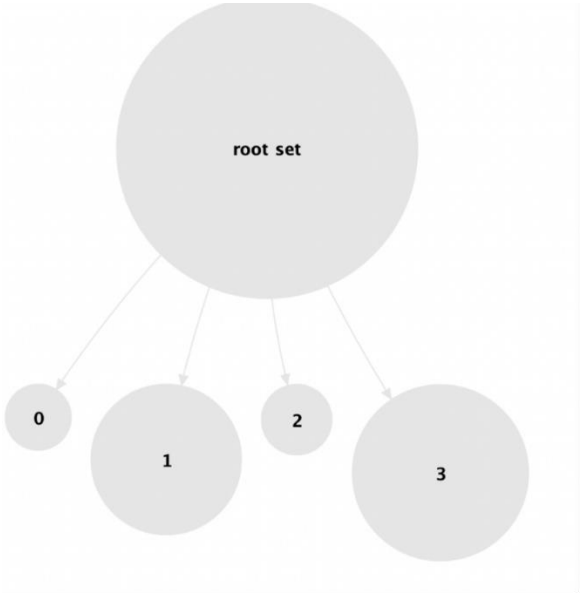


Attribute	cluster_0	cluster_1	cluster_2	cluster_3
https	0.004	0.001	0.033	0.095
know	0.003	0.003	0.003	0.026
people	0.010	0.007	0.004	0.022
depression	0.013	0.087	0.021	0.022
make	0.002	0	0	0.019
feel	0.003	0.002	0.004	0.018
cause	0	0	0.003	0.015
please	0.002	0.006	0.002	0.015
cured	0	0	0	0.013
mental	0.018	0.004	0.005	0.013
time	0.006	0.001	0.002	0.012
great	0	0.004	0	0.012
thought	0	0	0.004	0.012
world	0	0	0.002	0.011
made	0.033	0	0.004	0.011
year	0	0	0.007	0.011

Fig – centroid tables of cluster data of mental health.

The term depression also have much impact on the mental health as this is the term which is close for all the cluster centroids.

The outline of the clusters is given below



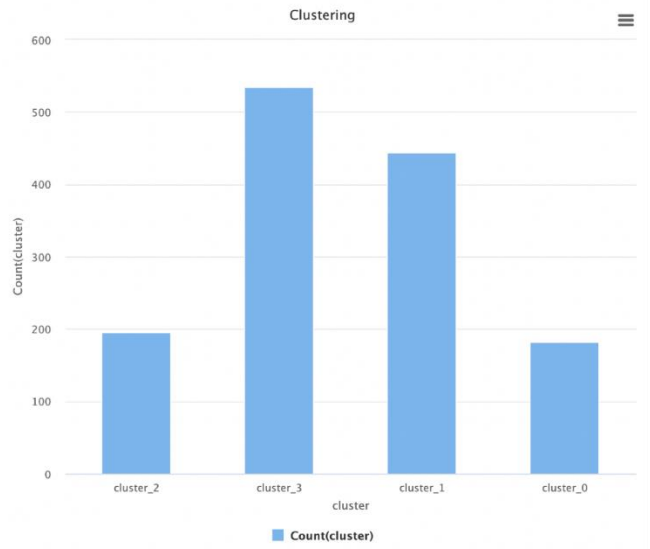


Fig- The clusters as per its size is mentioned in bar graph.

We can say that cluster 3 have morewords near to the centroid compared to others hence performance is higher for cluster-3

PerformanceVector

```
PerformanceVector:
Avg. within centroid distance: -0.950
Avg. within centroid distance_cluster_0: -0.873
Avg. within centroid distance_cluster_1: -0.954
Avg. within centroid distance_cluster_2: -0.931
Avg. within centroid distance_cluster_3: -0.981
Davies Bouldin: -7.226
```

Index	Nominal value	Absolute count	Fraction
1	ogheneyxle depressio...	264	0.053
2	mane depression thin...	205	0.041
3	mpickle regularly hour...	99	0.020
4	magic sekani beer foo...	80	0.016
5	roman baber blame c...	80	0.016
6	demonsi anxiety depr...	65	0.013
7	daddy lagos lack caus...	54	0.011
8	loeybug taking medica...	50	0.010
9	sarqasdjc remember t...	45	0.009

Close

After applying Naive Bayes to the data, we find that there is a higher likelihood of the world's depression, anger, and other words occurring in tweets, which in turn shows people's actions and mental states. the likelihood that terms like "taking medication" will appear is 0.010, indicating that some people are acting to improve their mental health.
We have taken some sample words to analyse the data.

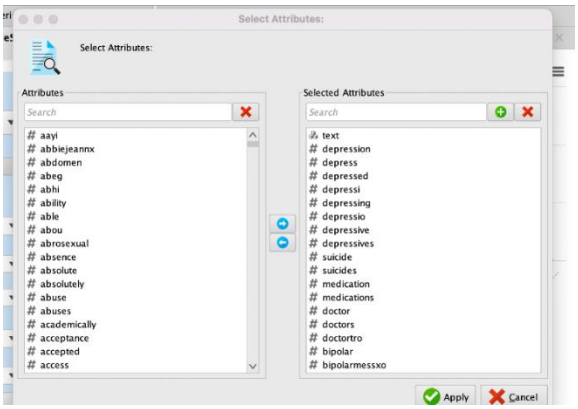
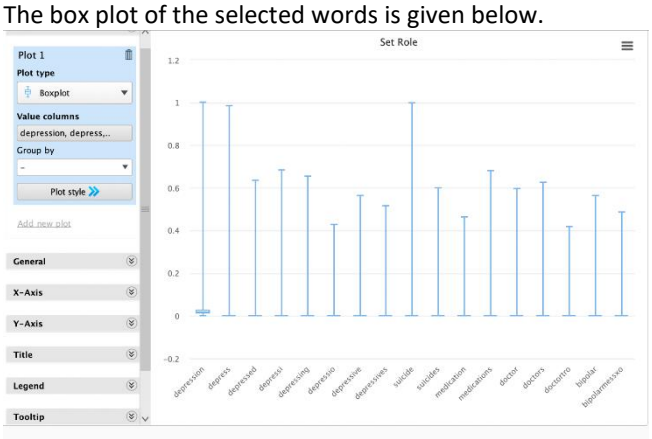


Fig- Screenshot in Rapidminer where we are selecting keywords for our mental health dataset.

The box plot of the selected words is given below.



We can infer from the plot above that not all persons who have depression are actively battling it.

iii)The last situation we are going to work on general tweets data from usa people, we will search by some important keywords and visualize the sentiment analysis and write some discussions on it.

when selected keywords for war on Ukraine and Russia and how usa people reacted. The sentiemnt score was positive based on that usa country people are not anti national in terms of war.

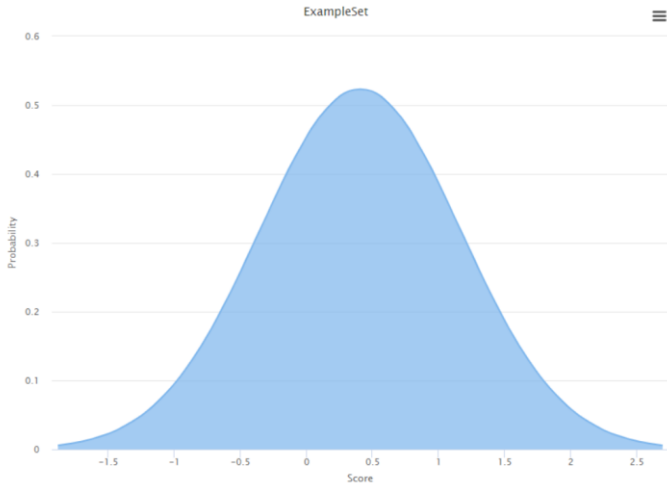


Fig- Sentiment score of people in usa for keyword- war.

When people talking about society majority of the people talk positive and it's a plus point for the nation

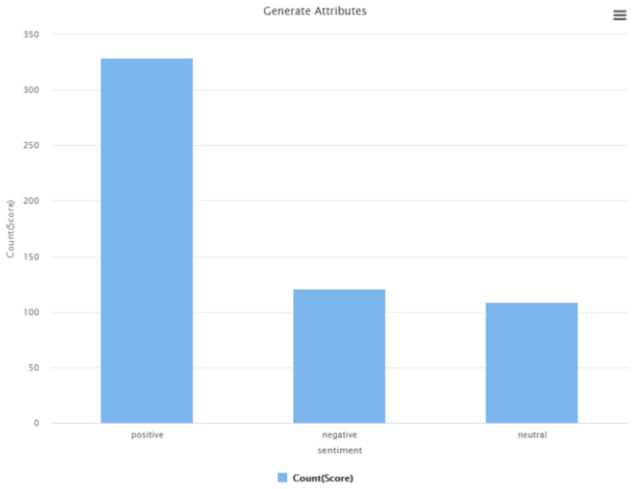


Fig- bar chart of people and their sentiment analysis of their tweets when they talk about society.

When people talk about crime majority are against to it.
As you can see there are lot negative tweets against to crime related topics, which people do not want to offer any kind of crime. We experimented with the sentiment data too as per our analysis we can say that usa people react against to any crime related topics in their country and it is a good thing for the nation image.

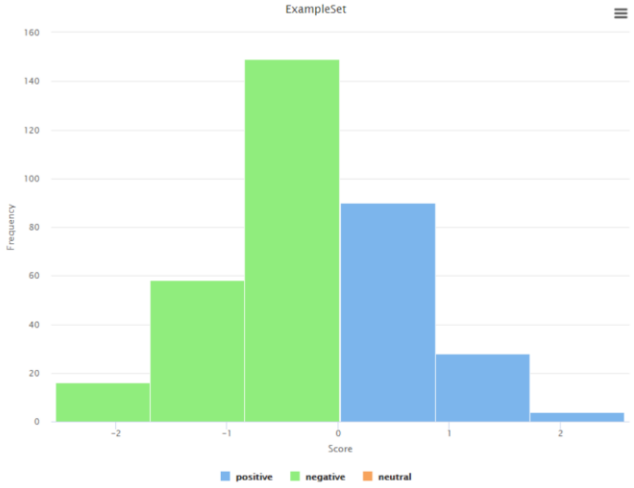


Fig- Histogram of people sentiment when keyword is crime

When people talk about their government majority are positive. Government is the key to rule the nation, and people's opinion on their government matters the most for the nation's image. If people are happy about the government then there is no doubt their reputation is good in most of the situations.

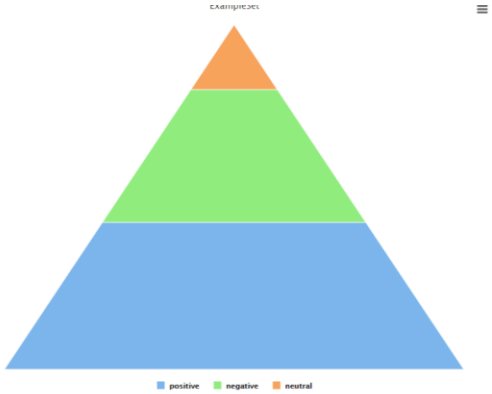


Fig- Pyramid chart of sentiment data when keyword is government.

The next keyword is president, not only government even president decisions are important and we made report on it. When people talk about their president majority are very happy.

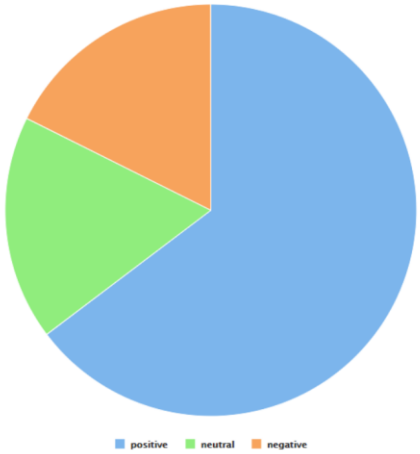


Fig- Pie chart of sentiment analysis o when keyword is president.

We just analysed people tweets on their nation which had usa and united states as keyword. We found most of their people talk about their nation positively and there very few people who are negative towards their nation we can term as antinational for reference.

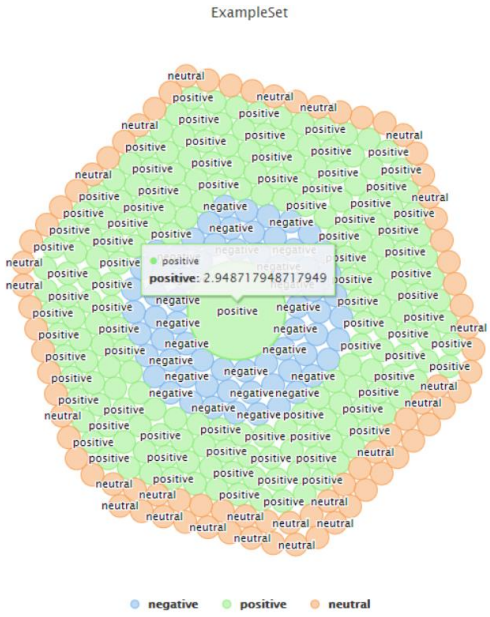


Fig - bubble plot of sentiment score when addressing nation

Rapid miner process:

Below we can find the design we used in rapidminer for addressing situations and answering research questing from data which we collected.

Some operator seems disabled because we are handling three datasets, and we did sentiment analysis and cluster analysis for different situations.
So we can enable operator in our convenience for accessing problems.

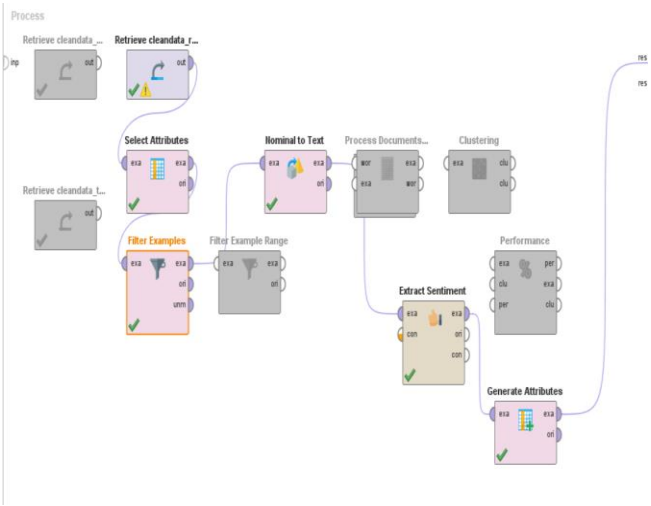


Fig- Rapid miner design used for this research.

5 CONCLUSION :

We gave detailed explanation on how we carried out this research from scratch to end and what tools we used and what data we collected is also addressed in above sections.

Results and discussion section have already seen the conclusion but to summarize our research. We can say that to measure nation's image using twitter data and applying sentiment analysis on it. We made three situations which can be summarized and give approximate answer on a country.

First scenario, where we talk about recession. Most people in the country talked negative but the effects from the recission like jobs and unemployment were dealt in positive manner. From this we can say that recession did hit the nation but people in the USA did not get effected which is good thing for Nation image.

Second Scenario, where we discussed about people's mental health conditions and how they dealt it. Most people had issues with their mental health but some are fighting but some are getting deeply affected. The people in the country are more linear towards pills and medication rather than natural treatments. Yes, people are highly effected by it and only few people are battling against it. Most people are very negative about this scenario. Hence, we can say that this scenario scored negative for the Nation image.

Third scenario, where we discussed about people being anti-national. These scenarios hold more weightage in nation's image. As in our results we discussed different keywords and measure their sentiment towards it, people talk very positive about their government and their president. Indeed, they react positively when comes for exterior matters such as war. Maximum people are not anti-national and very happy about their country. We can say that this scores a positive score for Nation's image.

Hence, we can conclude from above scenarios that two of are scenarios scored positive and only one scored negative. Where we can come to conclusion that Nation's Image is very positive and USA has good reputation among their people. Which makes a far more rank boost up in World countries reputation rank.

FUTURE ANALYSIS:

We would like to research more about different scenarios taking more surveys and getting data from widely in all possible ways. Which we can compare different scenarios in different countries and give them a rank in their reputation.

The tools we used had some issue's handing millions of samples. So, we need to do research on tools which can handle Big data effectively and reducing time consumption.

AUTHOR'S CONTRIBUTION:

As a whole group, everyone participated in all the discussions and meetings which happened over zoom.

Below are Author's contribution to the project:

HARSHAVARDHAN GYARALA - Extracted tweets for data collection using twitter API account in python, cleaned data in Python and imported files GitHub, contributed towards working on recession data analysis and writing results for that research question and contributed towards writing report, leading team, holding meeting sessions.

SNEHA SAHITHI PAMARTHI - Contributed in Data collection and gaining developer access, storing Data, contributed in doing cluster analysis of mental health scenario, written results for mental health dataset and helped in writing report.

VAISHNAVI SIMHACHALAM - Exploratory Data Analysis of cleaned data by writing essential keynotes. Which was helpful towards visualization. Worked in rapid miner for different keywords and their results. Helped in making presentation work for project.

NIKHIL POLISETTI - Creating Data visualizations and explaining them, helped towards writing Report, analysing keywords and choosing them.

VASHISTH SUKHADIYA - Contributed to Data visualizations and wrote results for general tweets dataset analysing them. Worked on anti-national scenario and helped in writing report.

VAMSI ANDE - Contributed in Exploratory Data Analysis and helped exploring research questions. Contributed towards writing report and making power point presentation.

REFERENCES:

[1] Stanford Sentiment Treebank Dataset. Available online: https://nlp.stanford.edu/sentiment/code.html/ (accessed on 11 November 2012).

[2] Pak, A.; Paroubek, P. Twitter as a corpus for sentiment analysis and opinion mining. LREc 2010, 10, 1320–1326.

[3] Alm, C.O.; Roth, D.; Sproat, R. Emotions from text: Machine learning for text-based emotion prediction. In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, BC, Canada, 6–8 October 2005; pp. 579–586.

[4] Bartlett, M.S.; Littlewort, G.; Frank, M.; Lainscsek, C.; Fasel, I.; Movellan, J. Recognizing facial expression: Machine learning and application to spontaneous behavior. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 2, pp. 568–573.

[5] Heraz, A.; Razaki, R.; Frasson, C. Using machine learning to predict learner emotional state from brainwaves. In Proceedings of the Seventh IEEE International Conference on Advanced Learning Technologies (ICALT 2007), Niigata, Japan, 18–20 July 2007; pp. 853–857 https://careerfoundry.com/en/blog/data-analytics/the-data-analysis-process-step-by-step/

[6] Twitter. (n.d.). Engineering. Twitter. Retrieved November https://blog.twitter.com/engineering/en_us

[7] Staszkiewicz, P., Chomiak-Orsa, I., & Staszkiewicz, I. (1970, January 1). Table 3 from dynamics of the covid-19 contagion and mortality: Country factors, social media, and market response evidence from a global panel analysis: Semantic scholar. undefined. Retrieved November 1, 2022, from https://www.semanticscholar.org/paper/Dynamics-of-the-COVID-19-Contagion-and-Mortality%3A-a-Staszkiewicz-Chomiak-Orsa/2e173cc2311f6f05d8ac3d44a984809c8b74ec1d/figure/5

[8] Ahmad, A. R., & murad, H. R. (2020). the impact of social media on panic during the COVID-19 pandemic in Iraqi Kurdistan Online Questionnaire study. Journal of Medical Internet Research, 22, E19556. - references - scientific research publishing. (n.d.). Retrieved November 1, 2022, from https://scirp.org/reference/referencespapers.aspx?referenceid=3326343

[9] SR;, K. A. B. A. (n.d.). Negative impact of social media panic during the COVID-19 outbreak in India. Journal of travel medicine. Retrieved November 1, 2022, from https://pubmed.ncbi.nlm.nih.gov/32307545/

[10] Yang, K.-C., Torres-Lugo, C., & Menczer, F. (2020, April 29). Prevalence of low-credibility information on Twitter during the COVID-19 Outbreak. arXiv.org. Retrieved November 1, 2022, https://arxiv.org/abs/2004.14484v1

[11] CS;, P. B. N. (n.d.). Mental health and the COVID-19 pandemic. The New England journal of medicine. Retrieved https://pubmed.ncbi.nlm.nih.gov/32283003/

[12] Brooks SK;Webster RK;Smith LE;Woodland L;Wessely S;Greenberg N;Rubin GJ; (n.d.). The psychological impact of quarantine and how to reduce it: Rapid review of the evidence. Lancet (London, England). Retrieved November 1, 2022, from https://pubmed.ncbi.nlm.nih.gov/32112714/

[13] AR;, K. (n.d.). Covid-19 in people with mental illness: Challenges and vulnerabilities. Asian journal of psychiatry. Retrieved November 1, 2022, from https://pubmed.ncbi.nlm.nih.gov/32298968/

[14] Li G;Miao J;Wang H;Xu S;Sun W;Fan Y;Zhang C;Zhu S;Zhu Z;Wang W; (n.d.). Psychological impact on women health workers involved in covid-19 outbreak in Wuhan: A cross-sectional study. Journal of neurology, neurosurgery, and psychiatry. Retrieved November 1, 2022, from https://pubmed.ncbi.nlm.nih.gov/32366684/