

# **PANDEMIC EFFECTS ON SENTIMENT ANALYSED FROM TWEETS AND NEWS PAPER REPORTS**

Project Report Submitted in Partial Fulfilment of the  
Requirements for the degree of  
Master of Computer Application  
Of  
Jadavpur University  
September 2020

By  
Supriyo Das  
Master of Computer Application – III  
Class Roll Number: 001710503025  
Registration Number: 140814 of 2017-2018

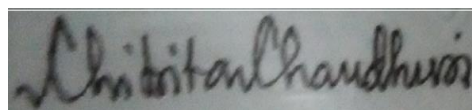
Under the guidance of  
Dr. CHITRITA CHAUDHURI  
Associate Professor

Department of Computer Science and Engineering  
Faculty of Engineering and Technology  
Jadavpur University  
Kolkata – 700032, India  
September 2020

**Department of Computer Science and Engineering**  
**Faculty of Engineering and Technology**  
**Jadavpur University**

TO WHOM IT MAY CONCERN

I hereby forward the project report entitled “*Pandemic Effects on Sentiment Analysed from Tweets and News Paper Reports*” prepared by **Supriyo Das** under my supervision to be accepted in partial fulfilment for the degree of **Master of Computer Application** in the Faculty of Engineering and Technology of Jadavpur University, Kolkata.



17/09/2020

(Dr. Chitrita Chaudhuri)

Associate Professor

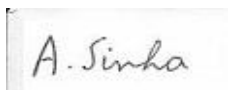
**Project Supervisor**

Dept. of Computer Science and Engineering

Jadavpur University

Kolkata – 700032

Countersigned:



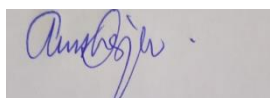
17/09/2020

Prof. Anupam Sinha

**Head**, Dept. of Computer Science and Engineering

Jadavpur University

Kolkata – 700032



18/09/2020

Prof. Abhijit Mukherjee

**Dean**, Faculty of Engineering and Technology

Jadavpur University

Kolkata – 700032

**Department of Computer Science and Engineering**  
**Faculty of Engineering and Technology**  
**Jadavpur University**

**CERTIFICATE OF APPROVAL \***

The foregoing project report is hereby approved as a creditable study of an engineering subject carried out and presented in a manner satisfactory to the degree for which it has been submitted. It is understood that, by this approval, the undersigned do not necessary endorse or approve any statement made, opinion expressed or conclusion drawn therein, but approve the project report only for the purpose for which it has been submitted.

Final Examination for  
evaluation of the project

---

(Dr.Chitrita Chaudhuri)  
Associate Professor  
**Project Supervisor**  
Dept. of Computer Science and Engineering  
Jadavpur University  
Kolkata – 700032

\*Only in case the project report is approved

(Signatures of Examiners)

## **DECLARATION OF ORIGINALITY AND COMPLIANCE OF ACADEMIC ETHICS**

I hereby declare that this project report contains literature survey and original research work by undersigned candidate, as part of my Master of Computer Application studies.

All information in this document had been obtained and presented in accordance with academic rules and ethical conduct.

I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

**NAME** : Supriyo Das

**Class Roll Number** : 001710503025

**Registration Number** : 140814 of 2017-2018

**Project Title** : Pandemic Effects on Sentiment Analysed  
from Tweets And News Paper Reports

**Signature** : 

**Date** : 17/09/2020

## **ACKNOWLEDGEMENT**

The satisfaction and euphoria that accompanies the successful completion of this task would be incomplete without the mention of the people who made it possible. Their constant guidance and encouragement crowned my effort with success.

It is a great pleasure to express my sincerest thanks to my project supervisor Dr.Chitrita Chaudhuri, Associate Professor, Department of Computer Science and Engineering, Faculty of Engineering and Technology, Jadavpur University, for her encouragement, valuable suggestion, and constant support during the course of this project.

I would like to thank all the professors of the Department of Computer Science and Engineering, Jadavpur University, Kolkata for the guidance they provided me throughout the duration of the Master of Computer Application course.

A special note of thanks goes to Prof. Anupam Sinha, Head, Department of Computer Science and Engineering, Jadavpur University.

I am also thankful to Prof. Abhijit Mukherjee, Dean, Faculty of Engineering and Technology, for providing an excellent environment for completion of this project.

I am also indebted to my co-researchers Mr. Arka Sengupta, Mr. Anupam Baidya, and Ms. Sukanya Basu for their seamless co-operation and help in completion of this project. I am thankful to my fellow classmates and my family for constant help and support.

Supriyo Das

Date: 17/09/2020

Supriyo Das

Master of Computer Application – III

Class Roll No. – 001710503025

Registration No: 140814 of 2017-2018

# Index

Chapters	Page No.
<b>1. Introduction</b>	<b>1-3</b>
<b>2. Previous Research Works</b>	<b>4-5</b>
<b>3. Basic Concepts</b>	<b>6-11</b>
3.1 Optical Character Recognition (OCR ) technology.....	6
3.2 Natural Language Processing.....	6
3.3 Sentiment Analysis.....	6-9
3.3.1 Tokenization and Part-Of-Speech(POS) Tagging.....	8-9
3.3.2 Stop Words Removal .....	9
3.4 Correlation Analysis .....	10-11
<b>4. Methodology</b>	<b>12-20</b>
4.1 Data Extraction.....	13-14
4.1.1 Tweet Extraction and Pre-processing.....	13
4.1.2 News Paper Data Extraction.....	14
4.2 Sentiment Analyzer .....	15-20
4.2.1 Stop Word Removal and Tokenization.....	15
4.2.2 POS Tagging.....	15-16
4.2.3 Selecting nouns, verbs, adjectives, adverbs as tokens.....	16-17
4.2.4 Computation of Sentiment Score .....	17
4.2.5 Calculation of Positive and Negative Scores per Text.....	17-18
4.2.6 Normalization of scores.....	18

## Index contd.

Chapters	Page No.
4.2.7 Evaluation of Resultant Score per text .....	18-19
4.2.8 Indication of sentiment of text.....	19
4.2.9 Normalized weekly sentiment score.....	19-20
4.3 Cosine similarity Calculation .....	20
 <b>5. Experimental Configuration</b>	 <b>21-24</b>
5.1 Datasets.....	21
5.2 TwitterWebsiteSearch-Master .....	22
5.3 Pytesseract.....	22
5.4 Matplotlib.....	22
5.5 Natural Language Toolkit(NLTK).....	23
5.6 SentiWordNet.....	23-24
5.7 CSV(Comma-Separated Values) File.....	24
5.8 Hardware Configurations.....	24
5.9 Software Requirements.....	24
 <b>6. Results and Inferences</b>	 <b>25-34</b>
6.1 Sentiment Score Graphs.....	25-32
A. The Weekly Line Graphs.....	25-28
B. The Overall Bar Graphs.....	29-32
6.2 Cosine Similarities and Positive to Negative Ratios.....	32-33
6.3 Overall Comparison Chart.....	33-34

## **Index** contd.

<b>Chapters</b>	<b>Page No.</b>
<b>7. Conclusion and Future Scope.....</b>	<b>35-37</b>
<b>8. Reference.....</b>	<b>38</b>

## **List of Tables**

1. Key word sets for all events from Tweets and News Data.....	21
2. NLTK tokenisation, pos tagging, stop word.....	23
3. Correlation Factors between Tweets and News Data.....	32

## **List of Figures**

1. Subjectivity Extraction.....	7
2. Sentiment Analysis.....	7
3. Stop words Removal.....	9
4. System Architecture.....	12
5. Sentiment Score Line Graphs on Public Health Awareness.....	25
6. Sentiment Score Line Graphs on Covid-19 Economy.....	26
7. Sentiment Score Line Graphs on Covid-19 Death Event.....	27
8. Sentiment Score Line Graphs on Covid-19 Education .....	28
9. Overall Sentiment Scores on Public Health Awareness.....	29
10. Overall Sentiment Scores on Covid-19 Economy.....	30
11. Overall Sentiment Scores on Covid-19 Death Event.....	31
12. Overall Sentiment Scores on Covid-19 Education Event.....	32
13. Cumulative Sentiment Counts of Tweet and News Items compared for all Events.....	33



# Chapter 1

## Introduction

Corona Virus Disease or COVID 19 is a new virus disease that originated in 2019. The virus has now spread across the world and almost all the countries are battling against this virus and are trying their best to curb the spread as much as possible. The World Health Organisation has declared it as a Pandemic and is leaving no stone unturned to control the pandemic and is awaiting a vaccine to cure it [1].

By the first week of March 2020, several countries like China, Italy, Spain, and Australia were fighting with the COVID19 pandemic by taking strict measures like nationwide lockdown or by cordoning off the areas that were suspected of having risks of community spread. Taking cues from the foreign counterparts, the government of India undertook an important decision of nationwide lockdown on March 25th for 21days from March 26th to April 14th, 2020. India, with a population of 1.3 Billion people, was at a high risk of suffering from irreversible damage, and strict measures were expected to “flatten the curve.” The Prime Minister of India announced the lockdown, but it did not come as a surprise because Indians were actually given a feel of what it had in store through a one-day curfew named as “Janata Curfew” of 14 hours on March 22nd from 7 A.M. to 9 P.M. Thus, Indians were exposed to a lockdown situation partially, and this helped in preparing mentally for the nationwide lockdown, and the announcement did not come as a shocker to them [2].

This work deals with the sentiment analysis of Indians after the lockdown announcements were made. The material used for analysis were public tweets in the social media platform Twitter and pandemic related news published in a popular widely used daily newspaper in India, The Telegraph. Tweets were studied to gauge the feelings of Indians towards the lockdown. It is well known that news in print media help to carve a niche in the public mind. So here the newspaper data related to covid-19 news were also extracted to analyse its effect on the people of India. Data were extracted using four major issues namely:

‘Public Health Awareness’, ‘Economy’, ‘Death’ and ‘Education’, from March 24th to June 30th 2020.

Social networks and Daily Newspaper were the two resources used to gather information about people’s opinions and sentiments towards the issues – mainly due to the following reasons. Firstly, in modern times, most people spend hours daily on social media to share their ideas, opinions, and reactions with others. Secondly, majority of the literate society usually also go through the morning newsprint regularly. So this project analyses the sentiments regarding four significant aspects using these two media.

The tweets related to coronavirus are being fetched using TwitterWebsiteSearch-master package and the related newspaper data using OCR technology and pytesseract. These tweets and data are then analysed using machine learning techniques. The experiments are being conducted through Python programmes on different tweets and Newspaper data, collected day wise, using NLTK library. Ultimately a lexical analyser, the SentiWordnet, exhibit the interesting results as positive, negative, and neutral sentiments utilizing different visualization tools.

It is required to find out how people’s sentiment towards an event varies with time. Thus, a temporal relationship automatically evolves. In case of Tweets, the bias of public sentiments reflected by positive comments is considered to denote optimism, whereas excessive negative comment in effect may indicate a general trend towards pessimism. Similarly, positive reports in print media is supposed to carry satisfactory hypes towards a particular event thus influencing en masse support, while negative ones may often influence general dissent towards it. It is also observed that the peak response time often coincides with the commencement of the event and decays with time, sometimes reaching a crescendo in between due to certain stray incidents. This automatically reflects the mood of the public which has been captured in the present work based on pandemic effects.

The endeavour of the present work is to develop a Knowledge Discovery tool which accepts crude Data Streams and Information packed Images at one end and predicts Pandemic effects on Public Sentiment. The system, in general, can well be tuned to predict results of other opinionated events such as Elections, Government Policies, and Economic Trends.

The rest of the document is organized as follows: Chapter 2 contains discussions on previous research work. Chapter 3 provides basic concepts on which the system is modelled. Chapter 4 illustrates the methodology used to implement the system. In Chapter 5 is provided the details of the software and hardware tools utilized. The results are discussed in Chapter 6. The conclusions drawn on the subject appear in Chapter 7. It also includes some future scope of the work. The references cited in the work are placed at the end.

# Chapter 2

## Previous Research Work

Christiana Loana Muntean et al. in their conference paper [3], discovered that subject of tweets could be found through hashtags and the terms present in tweets. In this case, distribution of hashtags in each and every tweet were taken into account. Here various datasets were collected through the Twitter Streaming API for a period of three days.

Hidenao Abe in his journal paper [4] focused on the temporal behaviour of the Twitter service known as “retweeting”. Users’ tagged retweets are affected by the content of the received tweets and their history of tweets. In order to predict such targeted tweeting behaviour of followers, a model is constructed, for which one should set up more proper features to consider the history of their tweets.

Zhao Jianqiang and Gui Xiaolin in their article [5], discussed the effects of six text pre-processing method on sentiment classification performance using feature models and classifiers on Twitter datasets. To identify the sentiment polarity, most existing approaches apply text pre-processing to reduce the amount of noise in the tweets. This improved the performance of the classifier and speeded up the classification process.

Pang B. and Lee L. have elaborated on sentiment classification of two types – one in terms of either objective or subjective - and the other categorized as positive, negative or neutral [6].

A. Pappu Rajan and S.P.Victor in [7] determined the positive or negative sentiment of text which extended to strength of polarity. This included data set collection, reading of opinion dataset, and removal of noisy data, splitting of opinion sentences into opinion word, and finally finding the positive and negative opinions. The actual number of positive and negative opinions from multiple sets are being compared here. Score of opinion is measured as the difference between the number of positive words and the number of negative words.

Shailendra Kumar Singh and Sanchita Paul in [8] stated that in sentiment analysis process, negation words and negative prefixes have potential to reverse the sentiment of sentences. Part-of-Speech (POS) tagging information

and opinion words and phrases are used for sentiment extraction. The opinion words and opinion phrases are used to extract positive / negative sentiments. There are two approaches – one lexicon-based and the other statistical-based.

Anusha K S and Radhika A D in [9] discussed about the sentiment analysis of twitter data. This involves data collection, data pre-processing, feature extraction, sentiment analysis, and ultimately polarity classification into positive, negative and neutral.

# Chapter 3

## Basic Concepts

### 3.1 Optical Character Recognition (OCR) technology

Optical Character Recognition, or OCR, is a technology that enables one to convert different types of documents, such as scanned paper documents, PDF files or images captured by a digital camera into editable and searchable data.

Assuming there is a paper document – such as a magazine article, brochure, or a PDF contract obtained via email - a scanner is not sufficient to make this information available for editing on an editor software. All a scanner can do is create an image or a snapshot of the document that is nothing more than a collection of black and white or colour dots, known as a raster image. In order to extract and repurpose text from scanned documents, camera images or image-only PDFs, what is needed is an OCR software that would single out letters on the image, put them into words and then - words into sentences, thus enabling one to access and edit the content of the original document.

### 3.2 Natural Language Processing (NLP)

NLP involves computational treatment of human language. It teaches computer the method to understand and generate human language.

NLP is utilized for massive management of textual information from multiple sources that is required for human usage.

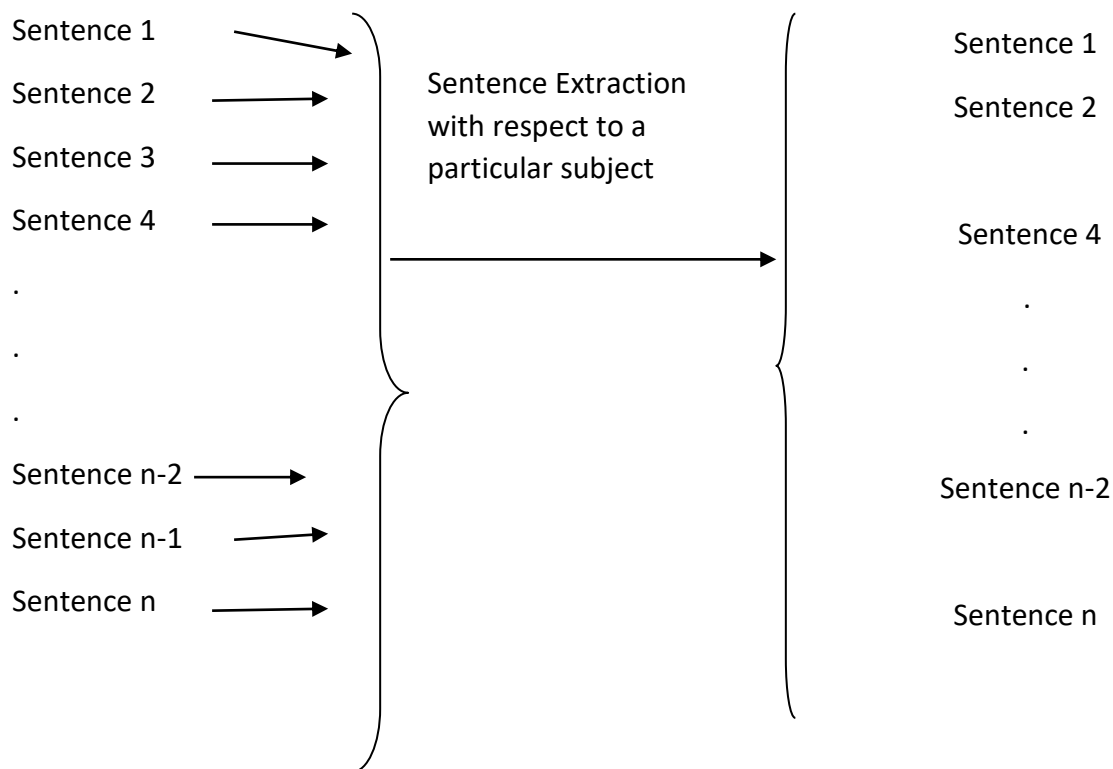
NLP may well be used for text classification which is an essential part in many applications, such as web searching, information filtering and sentiment analysis.

### 3.3 Sentiment Analysis

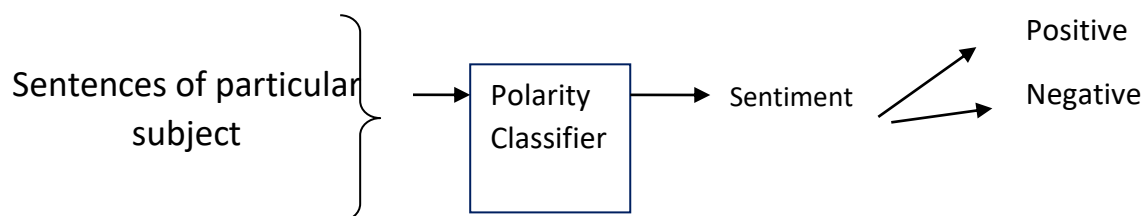
Sentiment Analysis is the computational study of people's opinions, appraisals, and emotions toward entities, events and their attributes.

Sentiment Analysis involves subjectivity analysis of a statement and then emotion identification that get expressed through the statement.

The process is depicted in the following figures 1 and 2



**Fig.1 Subjectivity Extraction**



**Fig.2 Sentiment Analysis**

Subjective Sentences express people's beliefs.

Components needed for identifying sentiments:

- Text containing the attitudes (sentence or entire document)
- Emotional expressions (eg. Positive, Negative)

The actual task of sentiment analysis can be broken up into several subtasks, of which two major ones are discussed next.

### 3.3.1 Tokenization and Part-Of-Speech (POS) Tagging

Tokens are individual words in a text and Tokenization is the process of breaking up into its individual words. Next, the POS Tagger software assigns a specific part of speech to each word in a text using the Penn Treebank tag set.

- CC coordinating conjunction
- CD cardinal digit
- DT determiner
- EX existential there (like: “there is” ... think of it like “there exists”)
- FW foreign word
- IN preposition/subordinating conjunction
- JJ adjective ‘big’
- JJR adjective, comparative ‘bigger’
- JJS adjective, superlative ‘biggest’
- LS list marker 1)
- MD modal could, will
- NN noun, singular ‘desk’
- NNS noun plural ‘desks’
- NNP proper noun, singular ‘Harrison’
- NNPS proper noun, plural ‘Americans’
- PDT predeterminer ‘all the kids’
- POS possessive ending parent’s
- PRP personal pronoun I, he, she
- PRP\$ possessive pronoun my, his, hers
- RB adverb very, silently,
- RBR adverb, comparative better
- RBS adverb, superlative best
- RP particle give up
- TO, to go ‘to’ the store.
- UH interjection
- VB verb, base form take
- VBD verb, past tense took
- VBG verb, gerund/present participle taking
- VBN verb, past participle taken
- VBP verb, sing. present, non-3d take
- VBZ verb, 3rd person sing. present takes
- WDT wh-determiner which
- WP wh-pronoun who, what
- WP\$ possessive wh-pronoun whose
- WRB wh-abverb where, when



Thus, POS-tagging is also known as grammatical tagging or word-category disambiguation. It may be described as the process of marking up a word in a text to correspond to a particular part of speech, based on its relationship with adjacent and related words in the text.

Following is an example of a parsed text and its POS-taggings:

**Text:** I would like to inform you that I performed badly in the Mathematics test

**Tokens:**{“I”, “would”, “like”, “to”, “inform”, “you”, “that”, “I”, “performed”, “badly”, “in”, “the”, “Mathematics” “test”}

**POSTags:** [(‘I’, ‘PRP’), (‘would’, ‘MD’), (‘like’, ‘VB’), (‘to’, ‘TO’), (‘inform’, ‘VB’), (‘you’, ‘PRP’), (‘that’, ‘IN’), (‘I’, ‘PRP’), (‘performed’, ‘VBD’), (‘badly’, ‘RB’), (‘in’, ‘IN’), (‘the’, ‘DT’), (‘Mathematics’, ‘NNS’), (‘test’, ‘NN’)]

### 3.3.2 Stop words Removal

Stop words usually refer to the most common words in a language. They would appear to be of little value in text. Using a stop list significantly reduces the number of words that a system has to use for some analysis. Stop words are removed to reduce the noise of textual data.

Sample text with Stop Words	Without Stop Words
GeeksforGeeks – A Computer Science Portal for Geeks	GeeksforGeeks , Computer Science, Portal ,Geeks
Can listening be exhausting?	Listening, Exhausting
I like reading, so I read	Like, Reading, read

Fig. 3 Stop words Removal [14 ]

### 3.4 Correlation Analysis

J. Han et al. in [11] mention that frequent pattern mining is required to find interesting associations between recurring relationships in large transactional or relational data sets. They further point out that while mining frequent item sets, problem arises when a huge number of item sets are generated satisfying the minimum support threshold. To weed out the uninteresting rules, they prescribe correlation measures providing additional statistical significance.

C. C. Aggarwal in [12] stated that the words are typically correlated with one another in a large corpus of documents. Number of principal components is much smaller than the feature space. This necessitates finding word correlations. Document frequency is used to filter out irrelevant features. Very infrequent terms contribute the least to the similarity calculations. A term-document matrix may be viewed with the (i, j)th entry as the frequency of the jth term in the ith document. Bursty features can be identified depending on the underlying frequency. A pair of documents can have a relation if their cosine similarity is above a user-defined threshold. Considering  $A = (A_1 \dots A_n)$  and  $B = (B_1 \dots B_n)$  as the normalized frequency term vector in two different documents A and B, the cosine similarity between the two documents can be defined by the following Equation in (i).

$$\mathbf{A} \cdot \mathbf{B} = \|\mathbf{A}\| \|\mathbf{B}\| \cos \theta$$

Given two vectors of attributes, A and B, the cosine similarity,

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}, \dots (i)$$

where  $A_i$  and  $B_i$  are components of vector A and B respectively.

$\|\mathbf{A}\|$  is the Euclidean norm of vector  $A = (A_1, A_2, \dots, A_n)$ , defined as  $(A_1^2 + A_2^2 + \dots + A_n^2)^{1/2}$ . Conceptually, it is the length of the vector. Similarly,

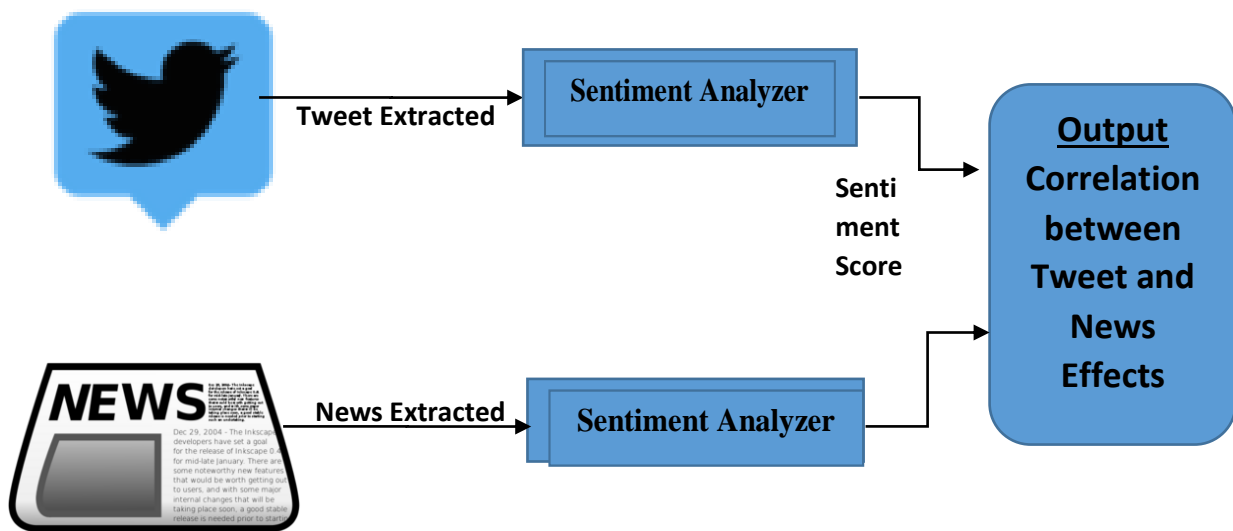
$||B||$  is the Euclidean norm of vector B. The measure computes the cosine of the angle between vectors A and B. A cosine value of 0 means that the two vectors are at 90 degrees to each other (orthogonal) and have no match. The closer the cosine value is to 1, the smaller the angle and the greater the match between vectors.

In the present context, the correlation between the two input segments, one from the tweeted texts and the other from the print media, can be established by using the above principles. The exact procedure is described at the end of the next chapter captioned Methodology.

# Chapter 4

## Methodology

The system accepts input from two separate sources - the tweets from the Twitter, and newsprint media data from e-newspaper The Telegraph as depicted in the System Architecture drawing in Figure 4 below.



**Fig. 4 System Architecture**

The Architecture diagram also exhibits the role of a Sentiment Analyser in deciphering positive and negative scores from each type of input media text, which are ultimately used to assess correlation effects. The actual functioning of the Sentiment Analyser would be further elaborated hereafter in subsection 4.2.

Presently the detailed process of data extraction from both of the input media is discussed below.

## 4.1 Data Extraction

Tweets of a particular event for a specific time range are collected using the package TwitterWebsiteSearch-master and Newspaper data for day wise from the popular Indian e-newspaper The Telegraph are collected using OCR technology with pytesseract of python and also fetching the particular event using some probabilistic approach with some set of keyword.

Event selections were achieved by utilising a probabilistic approach which helped to detect maximum association with a particular event, each maintaining a separate set of keyword repertoire for the purpose.

### 4.1.1 Tweet Extraction and Pre-processing

Tweets on a specific event (obtained from the relevant key word set for that event as tabulated in the next chapter) and within a certain time period are collected. Next, those containing emoji's, '?' and non-English characters are removed. The tweet text is produced along with Tweet Id and creation time.

Algorithm: Tweet\_ Extraction and Pre-processing

Input:

1. Set of predefined keyword for each Event
2. Time period, Language

Output: Tweet ID, Tweet Created At, Tweet Text

Method:

- [1] Import the package TwitterWebsiteSearch-master
- [2] Extract all the tweets for the given event
- [3] Convert all tweet data into lowercase
- [4] Remove punctuation, URL ,numbers, emoji from the data
- [5] tweet\_id=tweet['id\_str']
- [6]tweet\_createdat=tweet['created\_at']
- [7]tweet\_text=tweet['text']
- [8]Record tweet\_id, tweet\_createdat, tweet\_text in TweetDetails.txt

### 4.1.2 News Paper Data Extraction

News on a specific day are collected from e-newspaper portal ,The Telegraph. A probabilistic approach to fetch the news related to each event, in turn, is applied here with the help of a set of keywords formed earlier with samples from both datasets picked out manually. A threshold value is also set up for each event to facilitate the process of classification in choosing the correct event from the news-data.

Algorithm: Newspaperdata\_Extraction

Input:

1. URL of chosen e-newspaper
2. Set of predefined keyword for each Event
3. Time period
4. Probabilistic threshold value for each Event

Output: Event-wise News Text

Method:

- [1] Import the following packages:  
                pytesseract, requests, BeautifulSoup, datetime, Image.
- [2] Fetch all pages of news in image format.
- [3] For each page
- [4]     For each image of the article
- [5]         Convert the news image to text format.
- [6]         For each event
- [7]             Count= No. of key word present in this text.
- [8]             P\_value=count/length(predefined keyword set)
- [9]             If P\_value >= Probabilistic threshold value
- [10]            Then
- [11]                 Record text news data for the event
- [12]         End For
- [13]     End For
- [14] End For

## 4.2 Sentiment Analyzer

Sentiment Analysis from text involves several interim steps:-

1. Stop word removal and Tokenization
2. POS Tagging
3. Selecting nouns, verbs, adjectives, adverbs as tokens
4. Computation of sentiment scores(positive,negative)
5. Calculation of Positive and Negative Scores per Text
6. Normalization of Positive and Negative scores of a Text
7. Evaluation of Resultant Score per text
8. Indication of sentiment of text
9. Normalized weekly sentiment Score

Each of these steps are described in detail in the following subsections.

### 4.2.1 Stop Word Removal and Tokenization

In this section, Stop words are removed from Tweet Text and News Text

Algorithm: Stopword\_Removal

Input: Tweet Text or News Text

Output: Tweet Text without stop words or News Text without stop word

Method:

- [1] Download NLTK
- [2] Import stopwords module from nltk
- [3] stop\_words = set(stopwords.words('english'))  
//Consider English stopwords
- [4] Tokenize the Text into words
- [5] Remove stop words

### 4.2.2 POS Tagging

All words (except stop words) in tweet text or news text are assigned with their part of speech

Algorithm: POS Tagging

Input: Text words without Stop Words

Output: Words with their parts of speech

Method:

[1] Download NLTK

[2] Tag each word of a text with its respective part of speech

#### **4.2.3 Selecting nouns, verbs, adjectives, adverbs as tokens**

Here nouns, verbs, adverbs, adjectives present in a tweet are taken into account for sentiment analysis.

Algorithm: Generation of tokens

Input: Words with their pos tagging's

Output: Nouns, Verbs, Adverbs, Adjectives

Method:

[1] If postag(word)=='NNP'/'NNS'/'NN'/'NNPS'

[2] Then

[3]     word='noun'

[4] End If

[5] If postag(word)=='JJ'/'JJR'/'JJS'

[6] Then

[7]     word='adjective'

[8] End If

[9] If postag(word)=='RB'/'RBR'/'RBS'

[10] Then

[11]     word= 'adverb'

[12] End



[13] If postag(word)=='VB'/'VBD'/'VBN'/'VBP'/'VBG'/'VBZ'

[14] Then

[15]       word= 'verb'

[16] End If

#### **4.2.4 Computation of Sentiment Score**

Words of a text are matched with the words with hash in SentiWordNet. If pos tagging of the word according to NLTK and SentiWordNet are matched then the positive and negative scores of each word are taken.

Algorithm: Generation of Sentiment Score

Input: SentiWordNet, Words with their pos tagging's

Output: Positive and Negative Scores of Word(s)

Method:

[1] Check if a word in text is present with hashtag in

      SentiWordNet

[2] If the pos tagging's of the word in SentiWordNet and that

      according to NLTK match or not

[3] If match occurs take positive and negative scores for further

      analysis

#### **4.2.5 Calculation of Positive and Negative Scores per Text**

The positive and negative score of all the words are added to find the Total of Positive and Negative Sentiment Scores (Positive and Negative) of a text.

Algorithm: Generation of Positive and Negative Scores of a tweet

Input: Average Positive and Average Negative Scores of all Words in a text.

Output: Total Positive and Total Negative Scores of all the words in a text.

Method:

- [1] For each word in the text
- [2]     Add Positive Score to Total Positive Score
- [3]     Add Negative Score to Total Negative Score
- [4] End For

#### **4.2.6 Normalization of scores**

The positive and negative Sentiment scores of a text are normalized by dividing the total by the number of words in the text.

Algorithm : Normalized Positive and Negative Score per Text

Input: Total Positive and Total Negative Score of text.

Output: Normalized Positive and Normalized Negative Sentiment Score  
of the text.

Method:

- [1] Count total number of words in text and assign in  
word\_countintext
- [2] Divide both Total Positive and Total Negative Score of text by  
word\_countintext
- [3] Assign the results as Normalized Positive and Normalized  
Negative Sentiment Score

#### **4.2.7 Evaluation of Resultant Score per text**

The difference between the Normalized positive and Normalized negative sentiment score are calculated

Algorithm: Resultant Score per text

Input: Normalized Positive Sentiment and Normalized Negative  
Sentiment score of the text.

Output: Final\_sentiment\_value of text

Method:

$$[1] \text{ Final\_sentiment\_value} = \text{Normalized Positive Sentiment Score} - \text{Normalized Negative Sentiment Score}$$

#### **4.2.8 Indication of sentiment of text**

Finding out whether text contain positive / negative sentiment.

Algorithm: Ascertain the sentiment polarity of text

Input: Final\_sentiment\_value

Output: Sentiment Polarity (Positive/Negative)

Method:

[1] If Final\_sentimentvalue>0

[2] Then

[3]       Sentiment Polarity = Positive

[4] End If

[5] If Final\_sentiment\_value<0

[6] Then

[7]       Sentiment Polarity = Negative

[8]End If

#### **4.2.9 Normalized weekly sentiment score**

Normalized Positive/Negative Sentiment Scores of all text are taken for analysis.

Algorithm: Generation of Normalized weekly sentiment score

Input: Normalized Positive and Negative Sentiment scores per day.

Output: 1) Normalized Positive Sentiment score per week

2) Normalized Negative Sentiment score per week.

Method:

```
[1] Initialization:
[2]   Normalized_P_Score_Week:=0
[3]   Normalized_N_score_Week:=0
[4] For each Week
[5]   Normalized_P_Score_Week += Normalized_P_Score_Day
[6]   Normalized_N_Score_Week += Normalized_N_Score_Day
[7] End For
[8] Normalized_P_Score_Week = Normalized_P_Score_Week / 7
[9] Normalized_N_Score_Week = Normalized_N_Score_Week / 7
```

### 4.3 Cosine similarity Calculation

Correlation between two different data source of each event is found by using the cosine similarity between the tweeted texts and the print media text for that event. The cosine similarity is measured based on Eq. (i) provided in section 3.4 of the previous chapter. The vector elements for each data source comprised of the total sentiment score of a days text – the total is calculated considering both positive and negative sentiment scores together. Thus, as discussed in the above mentioned section of the previous chapter,

- Element  $A_i$  represents total sentiment score for tweet text of  $i^{\text{th}}$  day.
- Element  $B_i$  represents similar score for print media texts of  $i^{\text{th}}$  day.
- Variable  $i$  varies from 1 to  $n$ , where  $n$  is the number of days considered.

The actual cosine similarity values for each event is calculated and tabulated within chapter 6 captioned Results and Inferences.

# Chapter 5

## Experimental Configuration

### 5.1 Datasets

A. Tweet Data of Twitter between March 24, 2020 and June 30, 2020

B. News Data of The Telegraph from March 24, 2020 to June 30, 2020

C. **Table 1: Key word sets for all events from Tweets and News Data**

Sl. No.	Tweets and News on event	Key word sets
1.	Public Health Awareness	'social distancing ', 'quarantine', 'isolation ', 'community spread ', 'lockdown', 'W.H.O guideline', 'guideline', 'awareness ', 'covid-19 spread', 'Sanitization', 'PPE N95 mask', 'confirmed case', 'prevent', 'stay safe from corona', 'hygiene maintain', 'trauma', 'child affected'
2.	Economy	'economy effect', 'covid-19 economy', 'lockdown economy', 'industry', 'corporate sector effect ', 'shutdown economy', 'economic growth', 'companies', 'market', 'jobless in corona time', 'economic fallout', 'manufacturing sector', 'rural agriculture economy', 'RBI', 'unemployment in corona time', 'farmer ', 'capital investment in corona time', 'atmanirbhar bharat package'
3.	Death	'covid-19 death', 'corona death', 'co-morbidities death ', 'death rate in corona'
4.	Education	'Impact of Coronavirus on Education', 'school college off', 'education ', 'lockdown education', 'covid-19 education', 'school', 'college education', 'examination', 'exam postpone', 'university education ', 'final year exam', 'education session ', 'lockdown education', 'online education', 'online class ', 'online child education'

## 5.2 TwitterWebsiteSearch-master

Twitter website search master is a package to extract tweets older than 7 days from Twittercom.search without using Twitter API.

Here language of tweet to be extracted and the event whose tweets are extracted are placed in particular commands under this package and the tweets get automatically downloaded for further offline uses.

Tweet attributes such as id\_str, created\_at and text of a tweet are recorded using TwitterWebsiteSearch-master.

## 5.3 Pytesseract

Pytesseract is an optical character recognition (OCR) tool for python. That is, it will recognize and “read” the text embedded in images.

Pytesseract is a wrapper for Google’s Tesseract-OCR Engine. It is also useful as a stand-alone invocation script to tesseract, as it can read all image types supported by the Pillow and Leptonica imaging libraries, including jpeg, png, gif, bmp, tiff, and others. Additionally, if used as a script, Pytesseract will print the recognized text instead of writing it to a file.

Google Tesseract OCR Engine is a Machine Learning model developed by Google. Image of characters are trained by Google for this particular model. As we have used scanned article which is basically image of printed data, so it is used as test data for this model and it is acceptable worldwide for its performance and accuracy and for our case it has observed that we have achieved an acceptable accuracy for later work.

## 5.4 Matplotlib

Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002.

One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

## 5.5 Natural Language Toolkit (NLTK):

NLTK consists of NLP libraries and programs through which various works can be done like tokenisation, pos tagging, stop word removal and many others.

A sentence or data can be split into words using the method `nltk.word_tokenize(sentence)`.

Words in a sentence can be assigned with their parts of speech using `nltk.pos_tag(word)`.

NLTK has a module of stop words to detect stop words from sentences. This can be done using the command

```
stop_words = set(stopwords.words('english'))
```

after importing stopwords module of nltk

**Table 2: NLTK tokenisation, pos tagging, stop word**

Sentence	NLTK Tokenization	NLTK POS Tag	NLTK Stopwords
I love to dance	['I', 'love', 'to', 'dance']	[('I', 'PRP'), ('love', 'VBP'), ('to', 'TO'), ('dance', 'VB')]	to

## 5.6 SentiWordNet

SentiWordNet is a lexical resource in which each WordNet synset is associated to three numerical scores `Obj(s)`, `Pos(s)` and `Neg(s)`, describing how objective, positive, and negative the terms contained in the synset are.

A typical use of SentiWordNet is to enrich the text representation in opinion mining (OM) applications, adding information on the sentiment-related properties of the terms in text. OM is a recent subdiscipline at the crossroads of information retrieval and computational linguistics which is concerned not with the topic a document is about, but with the opinion it expresses. OM has a rich set of applications, ranging from tracking users' opinions about products or about political candidates as expressed in online forums, to customer relationship management.

In order to aid the extraction of opinions from text, recent research has tried to automatically determine the 'PN-polarity' of subjective terms, i.e. identify whether a term that is a marker of opinionated content has a positive or a negative connotation.

SentiWordNet is the first lexical resource which provide such specific level of detail (the word sense represented by a synset) and such broad coverage (all the 115,000+ WordNet synsets).

The method used to develop SentiWordNet is based on the quantitative analysis of the glosses associated to synsets, and on the use of the resulting vectorial term representations for semi-supervised synset classification.

A word is present for a number of times with it's positive and negative scores in SentiWordNet. This scores vary depending on the situation when the word is used. SentiWordNet also assigns score to words depending on their parts of speech.

## **5.7 CSV(Comma-Separated Values) File**

A CSV File is a comma-separated values file, which allows data to be saved in a tabular format. CSVs look like a garden-variety spreadsheet but with a .csv extension. CSV files can be used with most any spreadsheet program, such as Microsoft Excel or Google Spreadsheets.

CSV Module implements classes to read and write tabular data in CSV format.

## **5.8 Hardware Configurations**

- i. System: hpTM 15-AY542TU
- ii. Processor: Intel® Core™ i3-6006U CPU @ 2.00GHz  
RAM: 4.00 GB
- iii. System type: Ubuntu 16.04, 64-bit Operating System, x64-based

## **5.9 Software Requirements**

- i. Python of version greater than 2.6
- ii. Integrated Development Environments like Spyder of Python, where Python Programming is done



# Chapter 6

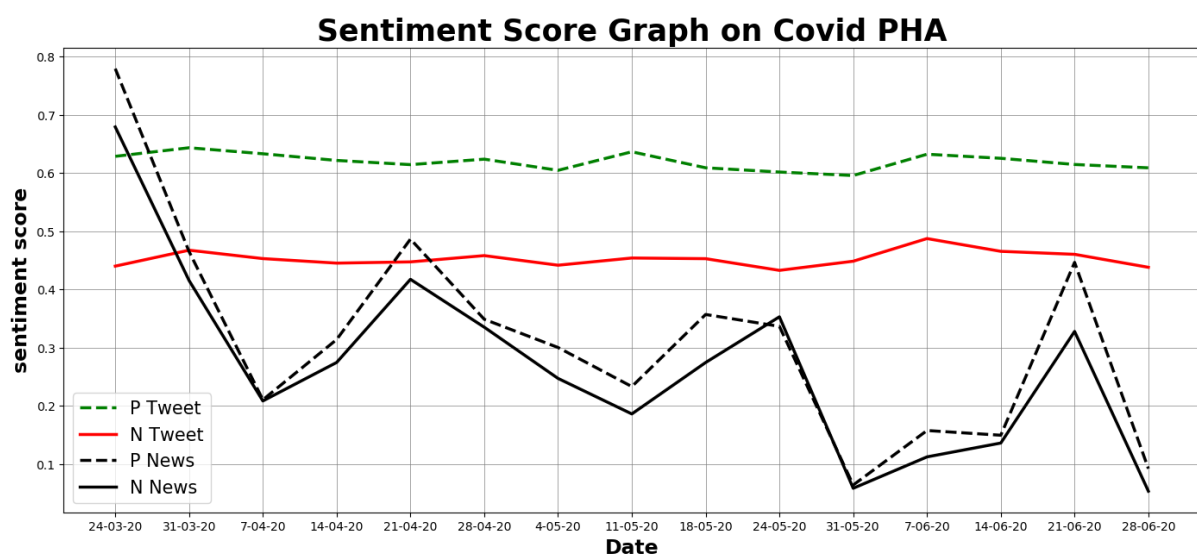
## Results and Inferences

### 6.1 Sentiment Score Graphs

#### A. The Weekly Line Graphs

Following are four Sentiment Score Line graphs drawn by plotting and joining weekly normalized sentiment scores derived corresponding to the four major events 'Covid-19 Public health Awareness' , 'Covid-19 Economy' , 'Covid-19 Death' , 'Covid-19 Education' between 24<sup>th</sup> March 2020 and 30<sup>th</sup> June 2020. The tweets were collected on a day wise basis, with a total tally of 45571, 32478, 7309, 16511 respectively per event. The Newspaper reports were taken likewise from the popular Indian daily "The Telegraph" related to the same events and during the same period to analyse how print media influences the public sentiment and to what extent.

#### i) Covid-19 Public Health Awareness (Covid-19 PHA)



**Fig.5: Sentiment Score Line Graphs on Public Health Awareness**

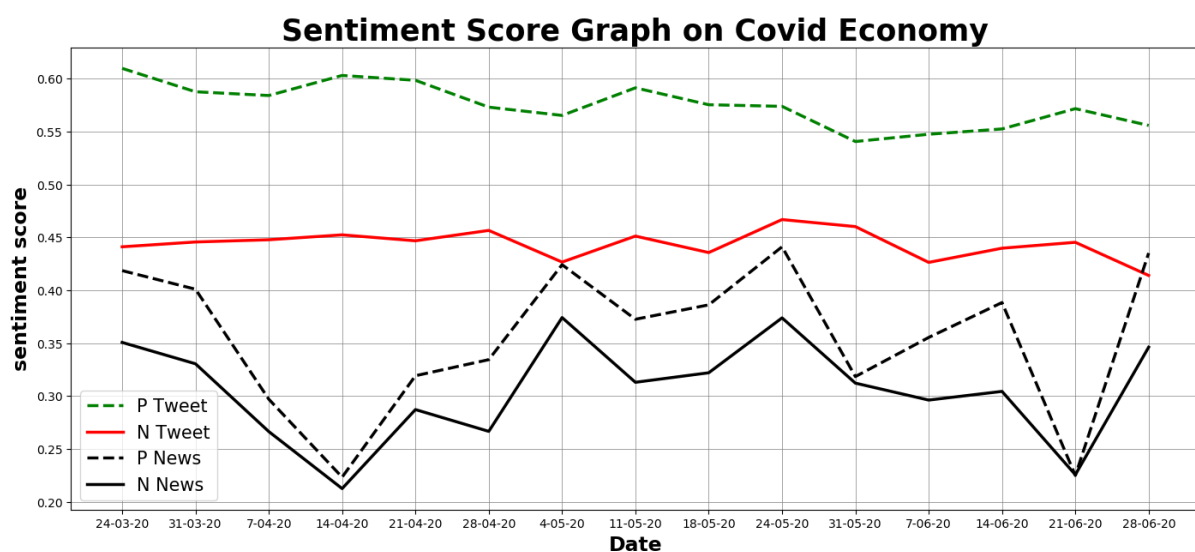
Fig 5 above apparently shows that the positive sentiment score is predominantly more prominent than the negative sentiment score, although followed in almost a parallel track by the latter, for both data sources during the whole period of observation. It is further observed that the print media curves offer a neck-to-neck race indicating that the policy of the media agents

is always towards feeding the public mind with balanced doses of positive and negative news!

For this event, the hype at the beginning of the newsprint data curve shows a steady decline with time, barring a few occasional spikes triggered off by sudden sensational news items. Interestingly, both the positive as well as the negative score curves for the public opinion remains relatively flat and undisturbed by media hypes and the optimists largely outperform the pessimists all through in terms of PHA!

Overall, the print media curve has a general exponential decaying trend with time for the PHA event, which is totally at variance with the social media curve with its regular almost horizontal nature for this event. We can safely deduce that correlation is not too strongly positive, but nor is it at all negative.

## ii) Covid-19 Economy



**Fig.6: Sentiment Score Line Graphs on Covid-19 Economy**

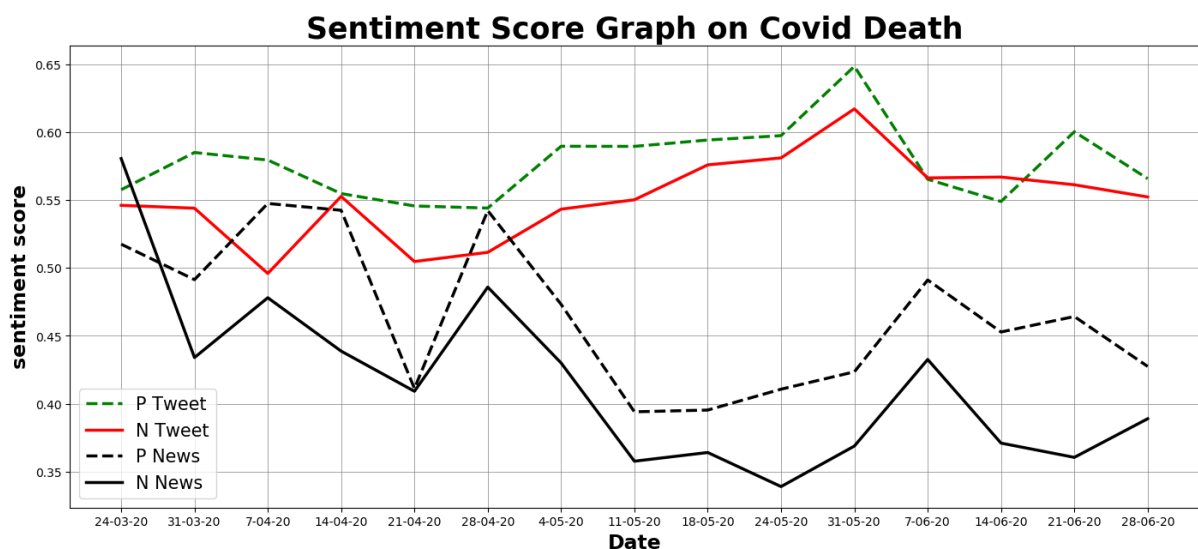
The main salient features remain the same for this event too as is adequately reflected by Figure 6 above – both inputs indicating more positivity, following almost parallel tracks, with the news media keeping a far lesser margin in between. Like in the case of PHA, the media curves blow hot and cold almost simultaneously, albeit a bit more on the optimistic side with several small crescendos attained at regular intervals. These hikes in positive trends in the print media may have been invoked by policies providing packages such as the

Pradhan Mantri Gareeb Kalyan Yojana (PMGKY) to distribute free food grains amongst a large majority of the population at the beginning of the lockdown period and extending it around June end.

Still, public opinions, in this case too, mostly steer clear of the media stings, and remain relatively on the higher side of sentiment score exclusively, with a definite optimistic trend. Yet, the two sets of curves for the two forms of inputs follow a generally undulating nature. Although it is definitely more discernible for the print media at all times.

Thus, all said and done, a far stronger correlation can be inferred in case of Covid-19 Economic effects as compared to the first event related to PHA, due to well reported economic crisis such as massive loss of employment and business caused by extended lockdown periods. Similarly, the different unlock phase's heralded definite changes in the economic scenario as reflected by the sharper peaks in the news media curves. These were generally followed by milder ups in the public opinion curves for this particular event.

### iii) Covid-19 Death



**Fig.7: Sentiment Score Line Graphs on Covid-19 Death Event**

Out of all the Public Sentiment graphs, the ones related to Death, depicted above in Figure 7, show the maximum ups and downs, and understandably so! The negative feelings expressed in tweets and news do not entirely follow the optimism lead like in the previous two cases, and at points even exceeds it.

Hence, the correlation between the two media seem to be certainly greater for the Covid-19 Death items. However, people are supposedly maintaining their sanctity in spite of all these, as may be deduced by the generally higher positive curve almost all throughout.

#### iv) Covid-19 Education

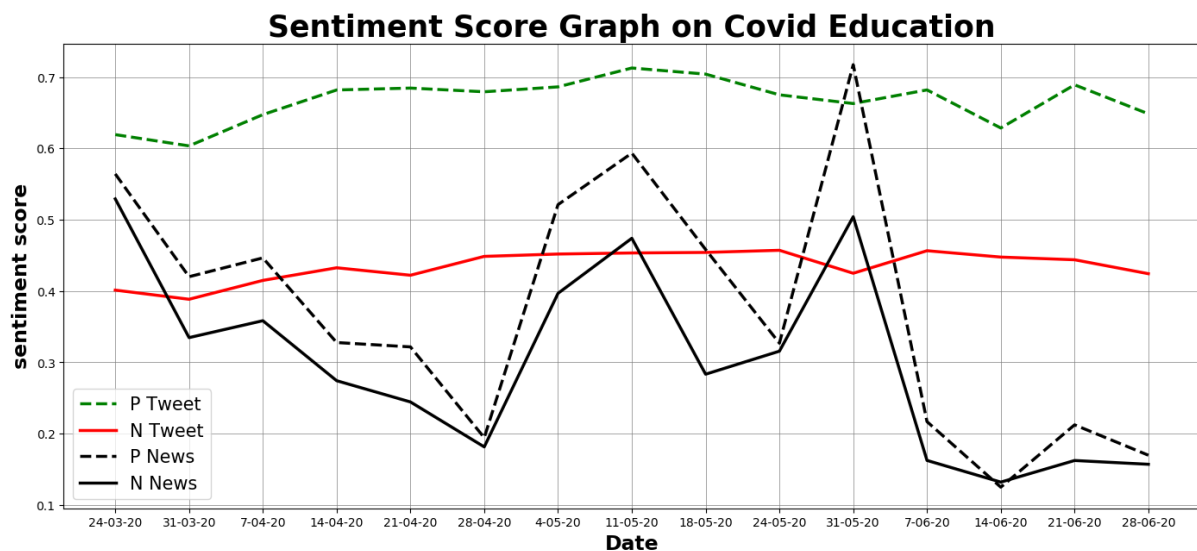


Fig. 8: Sentiment Score Line Graphs on Covid-19 Education

The fourth topic on Education seem to have affected news media sentiments most sporadically, while public sentiments retain its sangfroid nature with a large degree of positivity shown towards this topic. For the news media, Fig. 8 above displays at least two prominent spikes midway during the lockdown period, besides the initial burst common to all topics dealt by the media. These seem to correspond to important decisions taken by the Government and the Education Policy Makers regarding the fate of outgoing students of the year from various courses, as well as announcements of on-line classes generally.

But, the phlegmatic crowd of our country are apparently not much affected by these occasional bursts of enthusiasm on the part of news media! So, the correlation may be said to be much less than that for either Economy or Death scenarios. However, online education seem to have gone down well with our netizens – as reflected by the persistently huge positive dominance of the public sentiment curve in this particular case.

## B. The Overall Bar Graphs

The corresponding event graphs representing consolidated positive and negative sentiment scores for the entire period with both type of inputs are depicted in four more charts between Figures 9 and 12 next. These are projected as bar graphs to facilitate comparison between total positive and negative sentiments associated with each event both for public sentiment as well as news media.

The actual positive and negative score totals appear in the table? Presented in the next section, albeit as positive and negative percentages for both media.

The following graphs generally display the tendency of news media to preserve a balance between positive and negative bytes, with a slight bias on the positive side. The public sentiment also seems generally to veer predominantly towards optimism, which is a great sign for the health of the society.

### i) Covid-19 Public Health Awareness (Covid-19 PHA)

Overall Sentiment Score bar graphs depicted in Figure 9 below reflect the above observations summarily for the first event – Public Health Awareness.

## Overall Sentiment Score on Covid PHA

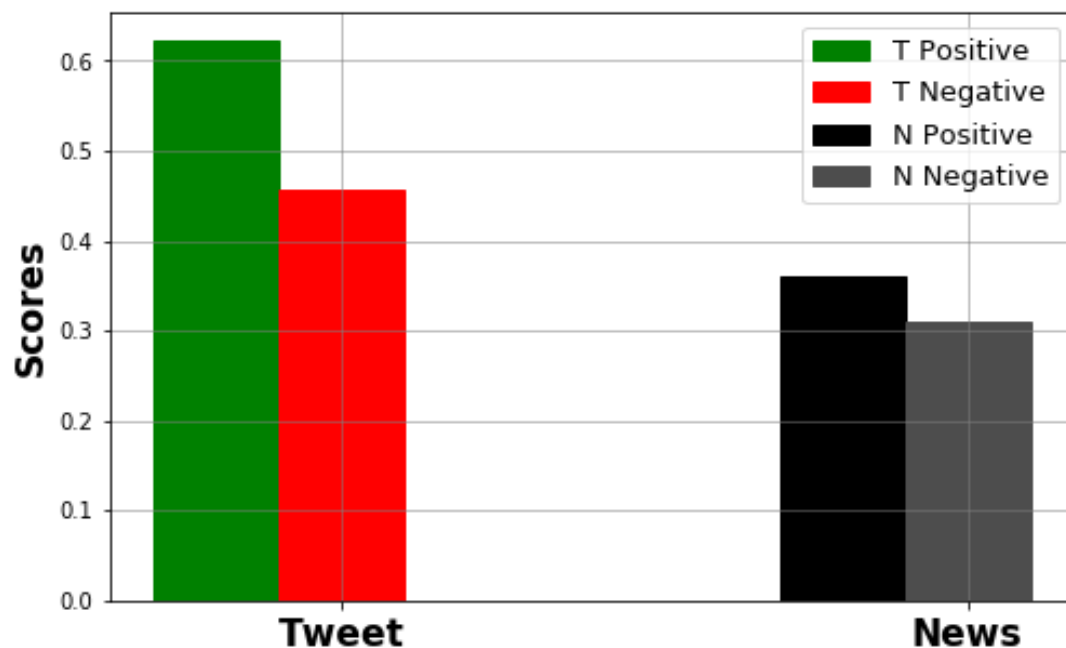


Fig.9: Overall Sentiment Scores on Public Health Awareness

## ii) Covid-19 Economy

In the overall sentiment bar graphs shown in Figure 10 below, the trend for marginal supremacy on the positive side by the newspaper media text and a much higher one for the tweet data is also maintained in the Economic Scenario.

### Overall Sentiment Score on Covid Economy

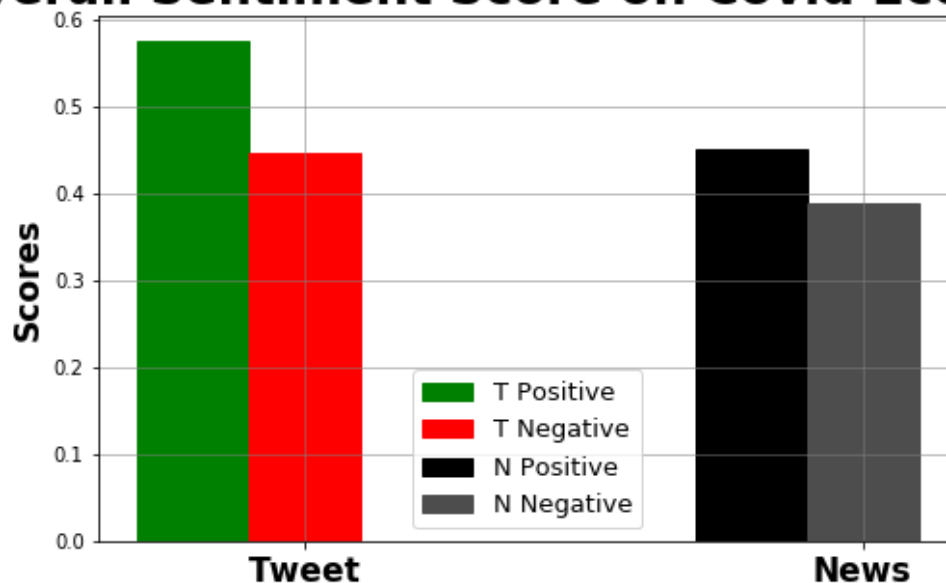


Fig.10: Overall Sentiment Scores on Covid-19 Economy

## iii) Covid-19 Death

The only situation where the gap between the positive and negative public sentiment is precariously compromised is in case of Covid-19 Death tolls. On hindsight, the contradiction and confusion offered by spurious news on the media may have been largely responsible for the panic attack reflected in the following Figure 11.

## Overall Sentiment Score on Covid Death

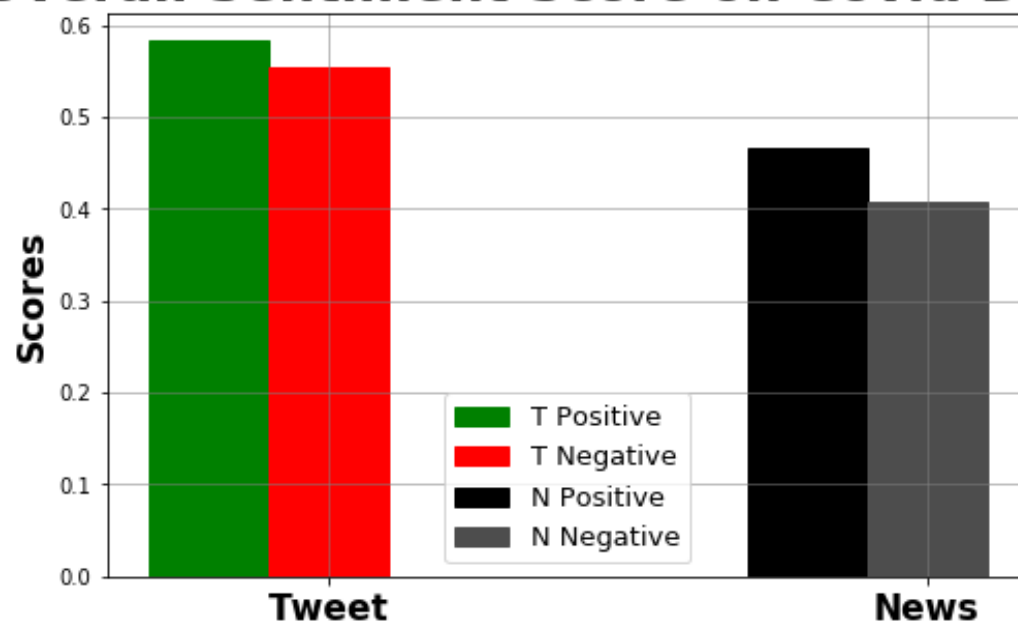


Fig.11: Overall Sentiment Scores on Covid-19 Death Event

### iv) Covid-19 Education

A reasonable jump in the positive count aggregate for Education - the last event considered in this research work – reflects that the part of society, who uses social media artefacts such as twitter, are quite comfortable with the online courses being offered as alternative education platforms. The newspaper media maintain its overall combination strategy, as reflected in following figure 12.

## Overall Sentiment Score on Covid Education

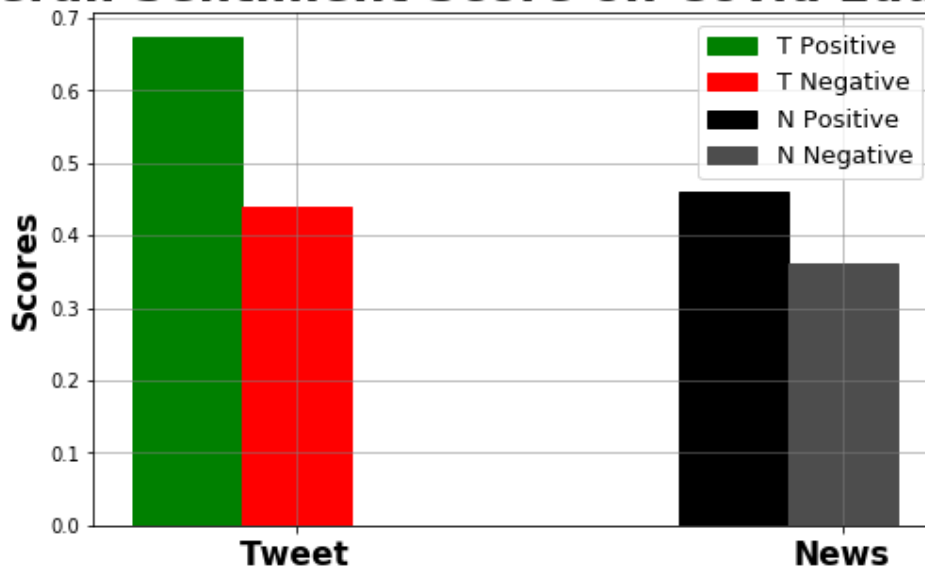


Fig. 12: Overall Sentiment Scores on Covid-19 Education Event

### 6.2 Cosine Similarities and Positive to Negative Ratios

Real life data captured from the two types of media were tested for correlation on a day-to-day basis by forming document vectors using event related keywords. Accordingly their cosine similarity measures have been recorded in the following Table 3, along with the total positive and negative count percentage and their ratios for both types of input. In this context, it may be pointed out that the actual count values have already been displayed in bar graph format in the previous section for event-wise visual comparison purpose.

Table 3: Correlation Factors between Tweets and News Data

Sl. N o.	Documents (Tweet & Newspaper Data)	Cosine Similarity rounded to 2 places of decimal	Positive Count%		Negative Count%		+ve:-ve Ratio	
			Tweet	News	Tweet	News	Tweet	News
1	On PHA	0.64	57.69	53.81	42.30	46.18	1.36	1.16
2	On Economy	0.92	56.34	53.76	43.65	46.23	1.29	1.16
3	On Death	0.89	51.25	53.35	48.74	46.64	1.05	1.14
4	On Education	0.70	60.25	55.92	39.47	44.07	1.52	1.27

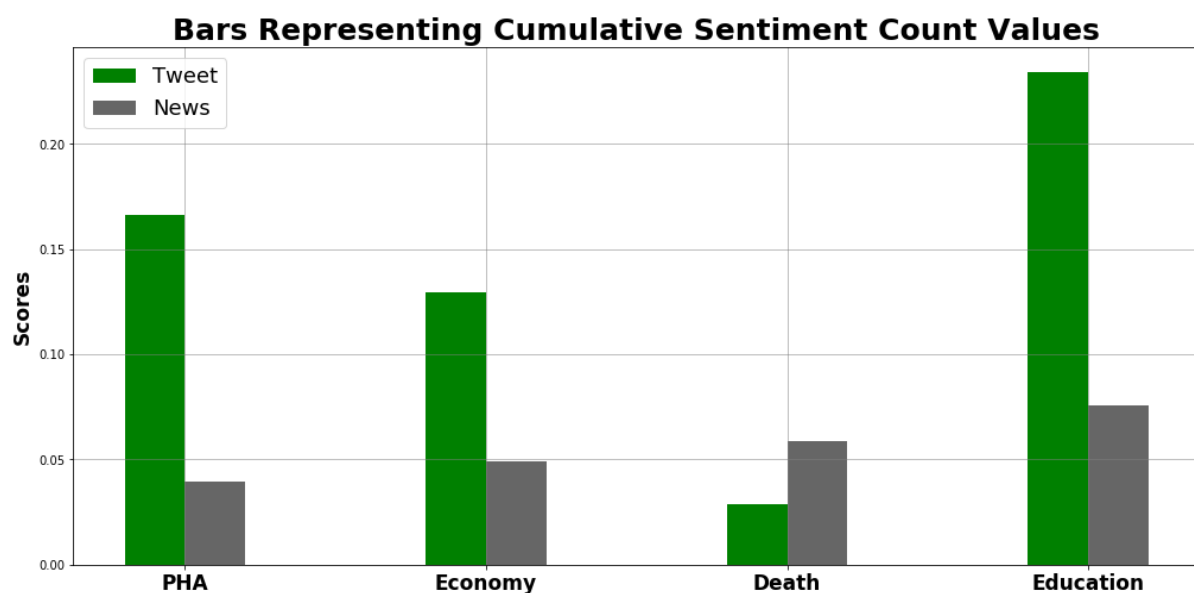


According to the cosine similarity measure results, two of the topics show high correlation between Public sentiment and Newspaper Items – those dealing with Covid-19-related Economy and Covid-19-related Deaths. These matches with the inference drawn in section 6.1A.

The overall positive nature of public sentiment traced for all four events, coupled with the general tendency of maintaining a slight positive bias by newsprint-media, corroborated by the last two ratio columns of the above Table 3 as well as the bar charts in section 6.1.B, explains the positive correlation existing by and large in all four cases.

### 6.3. Overall Comparison Chart

The purpose of the following graph is to establish the relative effects of the four events considered over the whole period – taking into account the cumulative counts of positive or negative sentiments together in each case – and viewing them together in the same platform, for both tweets (reflecting public sentiment) and news (reflecting print media instigations).



**Fig. 13: Cumulative Sentiment Counts of Tweet and News Items compared for all Events**

Definitely, the odd-man out here is the Death situation, where news print seem to predominate public sentiment overwhelmingly – which is a danger sign in itself! The significance of these bars are further corroborated by the last ratio columns present in the Table 3 above. The other three cases reflect

healthy situations as public sentiment in each of these scenarios remain optimistic to a large extent, irrespective of news media hypes.

The last sample, dealing with Education is still an oddity on its own, as it depicts sky-high optimism on the part of public opinion regarding the current education policies, supported once again by ratio values in Table 3. This may partially be due to a somewhat lop-sided view taken by the tech-savvy and affluent part of the community who access the twitter media.

# Chapter 7

## Conclusion and Future Scope

Findings suggest that sentiment analysis enables one to understand public sentiments with respect to specific products/services by their Comments, feedback, public opinion, newspaper report etc. and hence can be used for better decision-making approach. In this particular instance, public opinion were collected from a popular social networking service known as Twitter, on four topics relevant to the present pandemic situations. To trace back the possible origin of these opinions another low cost publication media, the daily newspaper, was also accessed. Machine learning tools were utilised to mechanize the collection and assessment procedure as both the volume and nature of data demanded usage of such techniques.

The chosen four topics were pervasive enough to gauge the pulse of the society as a whole in the New Normal situation. Out of these, the first topic, the Public Health Awareness may be the most important one, since it deals with a vitally major weapon to fight the virus till date. A clear positive note in this regard, obtained from the results, is a definite cheering circumstance, even within such a small scale study. The average positive Cosine Similarity index indicates low causal effect between News Media sentiment and Tweet data sentiment for this topic in the one hand, but it also acclaims the indomitable fighting spirit of the Indian Population expressed in the form of a high optimistic trend in Tweeted text in this context.

The next topic 'Economy' is self-evidently a burning question under all trying condition. The role of News Media is very important as it is vital to communicate the exact scenario authentically to the public mind to invoke the right positive trend. As such this demands a high correlation factor – and fortunately this is reflected in the present circumstance through the largest Cosine Similarity index (0.92) for this particular topic. Another advantageous aspect ascertained within the available data is the reasonably higher degree of positivity attained through public sentiment, as is evident from a +ve:-ve ratio of 1.29 factor in case of tweets.

The third topic 'Covid Death' by nature possesses a negative aspect. Naturally, both input tracks convey the least +ve:-ve ratio amongst all four topics, in spite of maintaining a value  $> 1$  on a positive note! In fact, the public sentiment just

avoids a negative majority with a really low ratio value of 1.05. The rapid fluctuations in sentiment values and high correlation between the two input modes both corroborate the very practical dread that uncertainty associated with such a controversial topic induces. A more positive role played by the News Media reports may help to produce a brighter prospect in this case.

The last topic 'Education' was chosen, as it is supposedly a big issue which can adversely affect the very foundation of a civilization if not handled properly. It may be said that the reflection of the public sentiment has not lived upto the expectation of the researchers in the present study. The over optimistic response of the Twitter input displayed by a high ratio value of 1.52, and a relatively lower cosine similarity index with print media news (0.7), indicate a detached and lax attitude amongst the public participants with a great disregard for reality regarding this issue. The section of society probed through the tweets may partially be responsible for this biased result – a more generalized gathering within other social media groups may open up better prospects of circumspection in this regard.

But overall, from the inferences drawn in the preceding Result section it can safely be concluded that the Covid19 pandemic has not sapped the moral of the Indian Society as a whole and the Nation is ready to fight against COVID19 with all its might. That in itself is a positive sign to be rejoiced over!

Future studies can look in to pre and post lockdown tweets and understand whether there was a change in sentiments from the beginning to the end of the lockdown. Some of these can also explore factors that affect mental health during lockdowns and pandemic spreads. Another area for future research could be tackling of fake news that gets circulated through social media, impacting receivers deliberately and falsely.

The approach described in this project work is therefore only reliably usable within the constraints of the corpus that have been collected. In future work, an improved system may be proposed for sentiment prediction based on the lessons learned during the present work. Due to time constraint, the research has been restricted to a sample, but it is hoped that later people may use Twitter sentiment analysis in real time, and be able to predict also from emoji symbols used most frequently in social media currently. Researchers can also think of identifying a more detailed variety of emotions such as happiness, alertness, certainty, and calmness. In this research, Twitter is the only social media taken into account, but various platforms such as, Stock Twits, Yahoo

Finance, Facebook, blogs, discussion forums can also be analysed. More variety in newspaper articles in terms of regional language inputs can also be accommodated to improve flexibility.

## 8. Reference

- [1] Tanu Singhal, “A Review of Coronavirus Disease-2019 (COVID-19)”,2020
- [2] Gopalkrishna Barkur,Vibha,Giridhar B. Kamath. “Sentiment analysis of nationwide lockdown due to COVID 19 outbreak: Evidence from India”,2020
- [3] Christiana Loana Muntean, Gabriela Andreea Morar, Darie Moldovan: Exploring the Meaning behind Twitter Hashtags through Clustering. Business Information Systems Workshop
- [4] Hidenao Abe: Extracting User Behavior-related Words and Phrases using Temporal Patterns of Sequential Patterns Evaluation Indices. Vietnam J Comput Sci, 2017
- [5] Zhao Jianqiang, Gui Xiaolin: Comparison Research on Text Pre-Processing Methods on Twitter Sentiment Analysis. Supported by: NSFC under Grant 1472316(in part), Shaanxi Science and Technology Plan Project under Grants 2016ZDJC-05 and 2013ZS16 -Z01/P01/K01 (in part) and Fundamental Research Funds for Ministry of Education of China under Grant XKJC2014008, February,2017.
- [6] Pang B. and Lee L. Opinion Mining and Sentiment Analysis. Journal Foundation and Trends in Information Retrieval. 2008; 2(1-2): 1 – 135
- [7] A Pappu Rajan and S.P.Victor, “ Web Sentiment Analysis for Scoring Positive or Negative Words using TweeterData”, International Journal of Computer Applications (0975 – 8887) Volume 96– No.6, June 2014
- [8] Shailendra Kumar Singh and Sanchita Paul , “Sentiment Analysis of Social Issues and Sentiment Score Calculation of Negative Prefixes”, International Journal of Applied Engineering Research · June 2015
- [9] Anusha K S , Radhika A D, “A Survey on Analysis of Twitter Opinion Mining Using Sentiment Analysis”, International Research Journal of Engineering and Technology (IRJET)
- [10] <http://www.expertsystem.com/machine-learning-definition/>
- [11] J. Han, M. Kamber, J. Pei, “Data Mining: Concepts and Techniques”, 3<sup>rd</sup> Edition, Morgan Kaufmann Publishers, An imprint of Elsevier, © 2012
- [12] C. C. Aggarwal, C. X.Zhai, “Mining Text Data”, Springer, 2012
- [13]Shiv Kumar Goel, Sanchita Patil, “Twitter Sentiment Analysis of Demonetization on Citizens of INDIA using R”,2017
- [14] <https://www.geeksforgeeks.org/removing-stop-words-nltk-python/>