A project report on

TEXT PROCESSING AND SENTIMENT ANALYSIS OF TWITTER DATA

MASTER'S DEGREE DISSERTATION

A thesis submitted in the partial fulfilment of the requirements for the degree of Masters in Computer Science

Submitted by

DEBLEENA PAL

(ROLL NO. : 15499018012, REG NO. : 181541810007 of 2018-2019)

ARAFAT MONDAL

(ROLL NO. : 15499018017, REG NO. : 181541810002 of 2018-2019)

RISHU SINGH

(ROLL NO. : 15499018004, REG NO. : 181541810015 of 2018-2019)

DEPARTMENT OF MASTERS IN COMPUTER SCIENCE

Under the supervision of

PARAMITA RAY

Dinabandhu Andrews Institute of Technology and Management

Maulana Abul Kalam Azad University of Technology

Date: 08-07-2020

TO WHOM IT MAY CONCERN

This is to certify that the work entitled as **'TEXT PROCESSING AND SENTIMENT ANALYSIS OF TWITTER DATA'** has been satisfactorily completed by:

DEBLEENA PAL

(ROLL NO. : 15499018012, REG NO. : 181541810007 of 2018-2019)

ARAFAT MONDAL

(ROLL NO. : 15499018017, REG NO. : 181541810002 of 2018-2019)

RISHU SINGH

(ROLL NO. : 15499018004, REG NO. : 181541810015 of 2018-2019)

It is a bonafide work carried out under my supervision at DINABANDHU ANDREWS INSTITUTE OF TECHNOLOGY AND MANAGEMENT, Kolkata for partial fulfillment of M.Sc. in Computer Science during the academic year 2018-2019.

It is understood that by this approval the undersigned do not necessarily endure or approve any statement made, opinion expressed or conclusion drawn there in but approve for which it has been submitted.

_____

Project Guide
PARAMITA RAY
Assistant professor
DINABANDHU ANDREWS INSTITUTE OF TECHNOLOGY AND MANAGEMENT, Kolkata

_____

Forward by
Paramita Ray
HOD of Computer science Dept.
DINABANDHU ANDREWS INSTITUTE OF TECHNOLOGY AND MANAGEMENT, Kolkata

_____

Signature of the External Examiner

Date:

DECLARATION OF ORIGINALITY AND COMPLIENCE OF ACADEMIC ETHICS

We hereby declare that this thesis contains original research work done by us, as a part of Master of Computer Science studies. All information in this document has been obtained and presented in accordance with the academic rules and ethical content.

We also declare that, as required by these rules and conduct, we have fully cited and referenced all the materials.

NAME	UNIVERSITY ROLL NO.	REGISTRATION NO.
DEBLEENA PAL	**15499018012**	**181541810007 of 2018-2019**
ARAFAT MONDAL	**15499018017**	**181541810002 of 2018-2019**
RISHU SINGH	**15499018012**	**181541810015 of 2018-2019**

PROJECT TITLE: TEXT PROCESSING AND SENTIMENT ANALYSIS OF TWITTER DATA

_____

SIGNATURE & DATE

ACKNOWLEDGEMENT

We are thankful to Prof. Paramita Ray to accept us as her students and to give us a chance to work under her guidance. She had showed keen interest from the beginning of the project. Her guidance and advices helped us greatly to complete this project. Her feedback in every step really improved us to improve at every step and complete the project within this stipulated period of time.

We are also thankful to the faculty of Dinabandhu Andrews Institute of Technology and Management for providing us a useful and healthy environment to continue our research towards the completion of our project.

We perceive this opportunity as a big milestone in our career development. We will strive to use gained skills and knowledge in the best possible way, and will continue to work on their improvement, in order to attain the best career objectives.

DATE: _____ SIGNATURE: ----------

NAME	UNIVERSITY ROLL NO.	REGISTRATION NO.
DEBLEENA PAL	**15499018012**	**181541810007 of 2018-2019**
ARAFAT MONDAL	**15499018017**	**181541810002 of 2018-2019**
RISHU SINGH	**15499018012**	**181541810015 of 2018-2019**

ABSTRACT

Data mining is essential in today's world to perceive public opinion via increasingly popularized online platform. Companies are inclined towards building algorithms that can understand public sentiment from popular social media platforms like Twitter. Twitter is a rich source of data for opinion mining and sentiment analysis that companies can use to improve their strategy with the public and the stake holders, by getting a clear insight on their emotional tone. However extracting data and performing the analysis still remains a hard task.

With the increase in the importance of computational analysis, many researchers are facing the challenge of using advanced technology that helps in the text analysis. One of the most popular environment for such computational analysis and emerging Data Science is R software. *In this project we have tried to extract the twitter datasets using some of the popular hashtags and have performed sentiment analysis over a period of time based on the datasets. We have selected four recent topics (coronavirus, Bollywood, technology and facebook) based on which we performed the opinion analysis. Our project also includes some of the solutions behind all the negative sentiments which can help in the decision making process.*

CONTENTS

Chapters	Topics	Page no.
	Certificate of Approval	2
	Declaration	3
	Acknowledgement	4
	Abstract	5
	Preface	6
Chapter-1	**Introduction**	9-11
	1.1 Overview of Text Processing	9
	1.2 Definition of Text Processing	9
	1.3 Applications of Text Processing	9
	1.4 Methods and Tools required	10
	1.5 History of Text Processing	10,11
Chapter-2	**Introduction to Sentiment Analysis**	12-15
	2.1 Brief overview of Sentiment Analysis.	12
	2.2 Sentiment Analysis on the microblogging platforms.	12
	2.3 Approaches for performing Sentiment Analysis on the microblogging platforms.	12,13
	2.4 About Twitter	13
	2.5 Literature Survey	14
	2.6 Need for R software in Text Processing.	15
Chapter-3	**Software Requirement Specification (SRS)**	16-19

Chapter-4	**Design Approach**	20,21
Chapter-5	**Development Methodology**	22-33
	5.1 Code for Tweet Extraction and Sentiment Analysis	25-32
Chapter-6	**Results and Discussion**	34-60
Chapter-7	**Conclusion and Future Scope**	61,62
	Appendix	63,64
	Reference	65

PREFACE

This project is divided into six different chapters:

Chapter 1 gives ***a preliminary overview of the text processing***. It provides the historical and the theoretical aspects of the text processing, with some of its applications in various fields. The text can be further processed for determining the emotional concept behind it which has been brought to light in chapter 2.

Chapter 2 gives ***a brief introduction to the sentiment analysis, its impact on various sectors and it brushes on the concept of how it can be used as an extension of text processing***.

Chapter 3 focuses on ***the Software Requirement Analysis (SRS)*** which specifies the general structure of the project and the requirements, both in terms of the hardware and the software. ***Our project mainly focussed on the software analysis***.

The ***flowchart of the design process*** involved in this project has been the topic of the **chapter 4.**

The entire portion of the **chapter 5** covers the ***development methodology*** and the ***code required for carrying out the tweet extraction, calculating the sentiment score and performing the analysis.***

Finally, ***the detailed observation has been described with reference to the visual graphs and the wordcloud*** generated in **chapter 6.**

The list of the libraries used and the references are mentoned at the end of this project.

The most important part of the project is that it extracts data from the hashtags of the most trending topics for four different months and it shows how the opinions of the people varied over a period of those four months, regarding the same topic.

CHAPTER-1

INTRODUCTION

1.1 OVERVIEW OF TEXT PROCESSING

Text processing is the automated way of analysing text for getting structured information. It is used in various sectors of a company, from product teams getting insights from customer reviews to automating various processes in customer service. There are several apps and services which carry out the process of text processing under the hood which can be explained by the following example.

Suppose we want to buy a new laptop, and we search 'top ten laptops trending' on google. Then we visited the online shopping sites and read some reviews, added to cart or made a wish list. Unknowingly, we have already left a trail of text data which is a source of valuable data for the companies. Whether we are aware of the fact or not, we are generating lots of data whenever we are searching things, texting, sending emails…the list is endless!

1.2 DEFINITION OF TEXT PROCESSING

Text processing is the process of analysing and manipulating text information which includes extracting bits of information from the text known as text extraction. We can assign values or tags depending on its content known as text classification. We can even perform calculations depending on the textual information. Text processing is one of the most common task in many ML applications:

1.3 APPLICATIONS OF TEXT PROCESSING

LANGUAGE TRANSLATION: The process of language translation involves translation of a particular sentence from one language to another.

SENTIMENT ANALYSIS: Sentiment analysis is the process of determining, from a text corpus, whether the sentiment towards a particular topic or product is positive, negative or neutral.

SPAM FILTERING: To detect unsolicited and unwanted email/messages.

These applications deal with a huge amount of text to perform classification or translation and it involves a lot of work on the backend. The text is transformed into something which can be actually digested by the algorithm.

1.4 METHODS AND TOOLS REQUIRED

Statistical methods: Statistics and Maths can be considered as the heart of text processing. Several statistical methods like frequency distribution, collocation and concordance can be used. For example,

Word frequency: This statistical method is used to count the most frequently used words or expressions in a text. This insight helps us to identify the problematic areas or the success areas.

Collocation: Collocations help us to find out the words that co-occur together in a text. Bigrams (two adjacent words) and trigrams (three adjacent words) are the most common types of collocations found in text.

Concordance: Concordance helps to decode the ambiguity in human language by analysing how specific words are used in different contexts. For example, the word issue can be applied to different scenarios.

1.5 HISTORY OF TEXT PROCESSING

Word processing or text processing did not develop out of the computer technology. It evolved to fulfil the needs of the writers rather than the mathematicians, gradually merging into one related to computer science technology. The history of word processing includes the gradual automation of the physical aspects of the writing and editing, and the refinement of the technology to make it available to the individuals and also the corporate sectors.

The invention of printing and the movable type in the middle ages emerged to a new world called 'Automation'. The first advancement in the manual writing, as far as the individuals are concerned, was in the form of the 'typewriter'. Henry Mill, an English engineer of the early eighteenth century, is credited with its invention.

The term word processing was invented by the IBM in the 1960s. by 1971, it was recognized by the New York as a 'buzz word'. A 1971 Times article referred to 'the brave new world of word processing'. IBM defined the term in a broad and vague way as the "combination of people, procedures and equipment which transforms ideas into printed communications" and originally used it to include dictating machines and ordinary, manually operated Selectric typewriters. By the early seventies, the term was generally understood to mean semi-automated typewriters affording at least some form of electronic editing and correction, and the ability to produce perfect "originals". Electromechanical paper based equipment such as the Friden Flexowriter was allowed for operations such as the repetitive typing of long letters, and when equipped with an auxiliary reader, could perform an early version of "mail merge". Circa 1970 it began to be feasible to apply electronic computers to office automation tasks. IBM's Mag Tape Selectric Typewriter (MTST) and later Mag-Card Selectric (MCST) were early devices of this kind.

[Source: http://www.computernostalgia.net/articles/HistoryofWordProcessors.html]

With reference to the section 1.3, we can see sentiment analysis is an application of the text processing. In our work, we have extracted twitter data from the hashtags of the trending topics and processed the extracted data. We have incorporated the process of sentiment analysis on the already processed data to perform a detailed analysis on the topics chosen.

So our project focuses on the general concept of the text processing through the tweet extraction and the conversion of the extracted tweets to a feasible format, upon which we can carry out the sentiment analysis.

CHAPTER-2

INTRODUCTION TO SENTIMENT ANALYSIS

2.1 BRIEF OVERVIEW OF SENTIMENT ANALYSIS

Sentiment analysis refers to the broad sector of the field of Natural language Processing, which generally aims at grasping a knowledge about people's opinions, attributes and emotional behaviour regarding a particular subject. The subject may be individuals, an event or trending topics. An immense amount of research work has done in this field, but it has mainly focussed on classifying formal and larger pieces of text data. With the immense growth of the social media and the microblogging platforms, these provide a major resource for the information on a large scale. Research work which have been focussing generally on the articles and the reviews have gradually shifted their domain, from the reviews to the microblogging sites due to their wide popularity.

2.2 SENTIMENT ANALYSIS ON THE MICROBLOGGING PLATFORMS

In the past few years there has been a massive growth in the usage of the microblogging platforms. Spurred by that growth, companies and media are increasingly seeking ways for extracting data from those microblogging platforms. One of the most popular microblogging platforms is Twitter. Companies such as Twitratr (twitrratr.com), tweetfeel (www.tweetfeel.com), and Social Mention (www.socialmention.com) are just a few who offer Twitter sentiment analysis as one of their services. A fair amount of research work has been done on the news articles and other reviews, but there has not been a sufficient amount of research work on these microblogging platforms. The online platform has been a significant medium for the people to express their thoughts and opinions and with the social media platforms there is an abundance of opinions available on a particular topic, product or service. Using sentiment analysis, the polarity of the opinions can be found such as positive, negative or neutral by analysing the text of the opinion. Sentiment analysis has been useful for companies to get their customer's opinions on their products predicting outcomes of elections, and getting opinions from movie reviews. The information gained from sentiment analysis is useful for companies making future decisions.

2.3 APPROACHES FOR PERFORMING THE SENTIMENT ANALYSIS ON THE MICROBLOGGING PLATFORM

The ***bag of words*** method is the traditional approach for the sentiment analysis. The relationship between the collections of words is considered instead of the relationship between individual words. When determining the overall sentiment, the sentiment of each word is determined and combined using a function. Bag of words also ignores word order, which leads to phrases with negation in them to be incorrectly classified. Other techniques

discussed in sentiment analysis include ***Naive Bayes, Maximum Entropy, and Support Vector Machines***. In the Literature Survey section, approaches used for sentiment analysis and text classification are summarised. Features such as automatic part-of-speech tags and resources such as sentiment lexicons have proved useful for sentiment analysis in other domains***. Our project focuses on the fact that whether these techniques can prove to be beneficial for microblogging platforms like Twitter***. The challenge of the microblogging is the wide range of topics it has to cover. People almost tweet about anything and everything. Therefore to mine Twitter data on any particular topic we need methods for quickly identifying the related set of data. ***In this paper we use one such method by using the Twitter hashtags, for example, #news, #technology etc. to determine the sentiments of the people.*** Before going into the depth of the analysis, we would like to throw some light on the microblogging platform that we chose to carry our work on.

2.4 ABOUT TWITTER

Twitter is a microblogging website aired in 2006 with currently having over 550 million users. The user created status messages are termed tweets by the service. The public timeline of twitter service displays tweets all over the world and is an extensive source of real-time information. The original concept behind microblogging was for providing personal status updates but nowadays it has become so popular that it covers almost everything in the world ranging from current political affairs to personal experiences. Movie reviews, travel experiences and current events etc. add to the list. Tweets and microblogs are different from the reviews in general. Reviews are characterised by formal text patterns and are summarised thoughts of authors whereas tweets are restricted to 140 text characters. Tweets offer company an additional review to gather feedback. Sentiment analysis helps customers to decide before making a purchase or planning for a particular movie.

Enterprises find this area useful to research public opinion of their company and products, or to analyse customer satisfaction. Organizations utilize this information to gather feedback about newly released products which supplements in improving further design. Different approaches which include machine learning (ML) techniques, sentiment lexicons, hybrid approaches etc. have been proved useful for sentiment analysis on formal texts. But their effectiveness for extracting sentiment in microblogging data will have to be explored. A careful investigation of tweets reveals that the 140 character length text restricts the vocabulary which imparts the sentiment. The hyperlinks often present in these tweets in turn restrict the vocabulary size. The usage of slangs and the misspelt words are very frequent in the tweets than other language resources is another hurdle that needs to be overcome. On the other way round, the amount of data available in these microblogging websites are diverse than any other platform on the internet. The data available on tweets is incomparable with other platforms. Microblogging language is characterized by expressive punctuations which convey a lot of sentiments. Bold lettered phrases, exclamations, question marks, quoted text etc. leave much scope for sentiment extraction. The proposed work attempts at extracting twitter data for sentiment analysis by aggregating an adapted polarity lexicon which has been learnt from the product reviews and topics under consideration. **[Source: Wikipedia]**

2.5 LITERATURE SURVEY

Sentiment analysis is a growing area of Natural Language Processing with research ranging from document level classification **(Pang and Lee 2008)** to learning the polarity of words and phrases **(e.g., (Hatzivassiloglou and McKeown 1997; Esuli and Sebastiani 2006))**. Given the character limitations on tweets, classifying the sentiment of Twitter messages is most similar to sentence-level sentiment analysis **(e.g., (Yu and Hatzivassiloglou 2003; Kim and Hovy 2004))**; however, the informal and specialized language used in tweets, as well as the very nature of the microblogging domain make Twitter sentiment analysis a very different task. It's an open question how well the features and techniques used on more well-formed data will transfer to the microblogging domain.

Just in the past year there have been a number of papers looking at Twitter sentiment and buzz**.(Jansen et al. 2009 ; Pak and Paroubek 2010; O'Connor et al. 2010; Tumasjan et al. 2010; Bifet and Frank 2010; Barbosa and Feng 2010 ; Davidov, Tsur, and Rappoport 2010)**. Other researchers have begun to explore the use of part-of-speech features but results remain mixed. Features common to microblogging (e.g., emoticons) are also common, but there has been little investigation into the usefulness of existing sentiment resources developed on non-microblogging data. Researchers have also begun to investigate various ways of automatically collecting training data. Several researchers rely on emoticons for defining their training data **(Pak and Paroubek 2010; Bifet and Frank 2010). (Barbosa and Feng 2010)** exploit existing Twitter sentiment sites for collecting training data. **(Davidov, Tsur, and Rappoport 2010)** also use hashtags for creating training data, but they limit their experiments to sentiment/non-sentiment classification. ***Here we use R statistical software for extracting data from Twitter and performing sentiment analysis. We have divided the sentiments into following categories:***

1) ***Positive***
2) ***Anticipation***
3) ***Trust***
4) ***Joy***
5) ***Negative***
6) ***Fear***
7) ***Sadness***
8) ***Surprise***
9) ***Disgust***
10) ***Anger***

This broad classification of the various categories of the opinions of the people regarding the trending topics on the twitter will surely help to bring into focus, the overall feeling regarding that particular topic. Our project not only focuses on the wide range of the classification of the emotions, but it also shows how it differs over a period of time, on that particular topic. For this plotting of the sentiment scores, we have used various R packages (Refer to the Appendix for the list of packages used) which helped in providing proper visuals in the form of wordcoud and the bar plots.

2.6 NEED FOR R SOFTWARE IN TEXT PROCESSING

The importance of computational text analysis in research has increased to a rapid extent nowadays. For this reason, many researchers are facing this challenge of learning how to use the advanced software for the procedure of text analysis. Currently one of the most popular emerging software for the computational analysis and the emerging field of data science is the R statistical software. However for the researchers who does not have a prior knowledge in computer programming language***, R proves to be an easier language for the computational analysis, especially for researchers having a mathematics and statistical background.***

R was specifically designed for statistical analysis, which makes it highly suitable for Data Science applications. Sometimes, programming with R can be a complicated learning experience for the people who does not have a prior programming experience but the tools now available for carrying out text analysis in R makes it a very powerful, cutting-edge text analytics using only a few simple commands. ***The collection of extension software libraries known in R terminology as packages, supplied and maintained by R's extensive user community is one of the keys to R's explosive growth in the sector of data analytics***. Each package extends the functionality of the base R language and core packages, and in addition to functions and data, must include examples and documentation, often in the form of vignettes demonstrating the use of the package. ***The best known package repository, the Comprehensive R Archive Network (CRAN), currently has over 10,000 packages that are published.***

Text analysis has become well established in R. There is a vast collection of packages regarding text processing and text analysis, from low level string operations to advanced text modelling techniques such as Fitting Latent Dirichlet Allocation models, R provides all. ***The main advantage of text processing and the sentiment analysis in R is that it is often possible and relatively very easy to carry out. It is also easy to switch between different packages or to combine them. Recent efforts among the R text analysis developer's community are designed to promote this interoperability to maximise flexibility and the choice among the users***. As a result, learning the basics for text analysis in R provides access to a wide range of advanced text analysis features.

[Source: Hackernoon]

CHAPTER-3

SOFTWARE REQUIREMENT SPECIFICATION (SRS)

Internal Interface requirement:

Project Plan

Introduction

This document lays out the basic plan of the project titled "**Text Processing and Sentiment Analysis of Twitter Data**". The intended readers of the project are generally expected to be the current and the future developers of the field of the sentiment analysis. The sectors for whom the project proves to be beneficial are further mentioned under the intended audience section. The SRS intends to give a brief structure of the project. The pictorial view or the flowchart of the structure is given under the chapter "Design Approach". The detailed process and the code required for the development of the algorithm is described in the chapter "Development Methodology".

Brief overview on the background of the project

The increased exposure of the intended set of customers to the polarised reviews of a certain commodity or a subject have created a significant impact on the companies and the organizations. The collection of the data from these resources have been a major concern in the recent times, as this research has proved as a catalyst for the major uprisings and shifts in the stock market. Due to the vast amount of the data, manual approaches to analysis are no considered feasible. There has been a number of computational approaches which have acted as a substitute for the manual approach and also provided to be the least time consuming and the feasible approach, however there has been a lack of research in this field, mostly because of the lack of the applications in practise and also it has not been incorporated yet in the decision making process. This process of sentiment analysis proved to be a huge potential in the major shift of the events. From the events of mass activism, such as the Egyptian revolution, to the twist in the tales of the stock markets, it has created a huge impact.

[Source: A generic framework for sentiment analysis: Leveraging opinion-bearing data to inform decision making Jacqueline Kazmaier1, Jan H van Vuuren]

Purpose

As we have seen, with the growth in the demand of the opinion mining, there has been an abundant amount of research work done, but it lacks to provide guidance on how to implement these algorithms in practise and how to incorporate them in the decision making

capabilities. The algorithms have also focused analysis, restraining itself to only three sets of polarities and haven't explored much into the sub-categories of the different emotions that people can actually have behind a subject. ***Our project has used wordcloud and different bar-plots as a support to provide a visualisation of the different classifications of the various kinds of emotions regarding a particular topic over a period of time. It has also focused on providing the variations of the sentiments, if any, over a period of four months, which can give a clear insight, of what the people think regarding that particular topic. If the topic had a particular sentiment in one month, it also provides an insight whether the opinions regarding that has changed over a period of time***.

Intended audience

The intended audience can be people related with the sectors related with the topics chosen, the media industry, medical sectors, entertainment industry and people interested in the field of research in the related topic. This can also be beneficial to the sectors of the sales and marketing.

Product scope

***The entire project (Text processing and the Sentiment Analysis) has been carried out using the R software*.**

Project goal

The aim of the project is to extract tweets from the microblogging platform (in our case Twitter).

The extracted tweets are then converted into a data frame which is the feasible form on which we can perform the text processing. After the necessary pre-processing of the data we perform the sentiment analysis and the desired output has been supported with the visuals such as the word cloud and the scores obtained in the sentiment analysis has been plotted in the form of bar plots. The R software contains an indispensable library which has helped us immensely in developing the entire algorithm required for fulfilling the goal of the project. The packages used are listed in the appendix section at the end of the project.

Benefits of the project

The people who will be benefitted from the project are mentioned in the section of the **intended audience.**

A brief overview of the Project

The Product functions are:

- Create a Twitter Application Programming Interface (API).
- Install and load R packages.
- Connect the Twitter account to R.
- Extract the Twitter data using Hashtags.

- Create a data frame using the extracted data.
- Remove the unnecessary components of the extracted tweets i.e., perform data pre-processing. Represent the data extracted in the form of a wordcloud to highlight the major keywords in the tweets.
- Perform sentiment analysis on the extracted data and interpret them in a graphical form

Functionalities

The project covers the most trending topics of the twitter and performs the mining of various opinions, and it covers the analysis over a certain period of time (four months) to determine whether there is any change regarding the particular topic.

- Extract the tweets using the hashtags of the chosen topics
- The tweet is extracted in the form of "lists" in the R programming language which is not feasible for the further text processing
- The tweets are converted to the dataframe.
- The text processing is performed on the extracted tweets converted to dataframe, removing all the unnecessary components which are not required for the sentiment analysis, for example, http links, stop words, stemming words, punctuation marks, numbers etc.
- After the required processing, the wordcloud is created which highlights the most commonly used words in the tweets.
- The sentiment analysis is performed and score is calculated based on each sentiment
- A visual representation of the analysis is provided using a bar plot. Scores are plotted on the Y-axis while the respective sentiments are plotted on the X-axis. .

Non Functional Requirement:

Performance Requirements:

If there are performance requirements for the product under various circumstances, state them here and explain their rationale, to help the developers understand the intent and make suitable design choices. Specify the timing relationships for real time systems. Make such requirements as specific as possible. You may need to state performance requirements for individual functional requirements or features.

Safety Requirements:

Specify those requirements that are concerned with possible loss, damage, or harm that could result from the use of the product. Define any safeguards or actions that must be taken, as well as actions that must be prevented. Refer to any external policies or regulations that state safety issues that affect the product's design or use. Define any safety certifications that must be satisfied.

Security Requirements:

Specify any requirements regarding security or privacy issues surrounding use of the product or protection of the data used or created by the product. Define any user identity authentication requirements. Refer to any external policies or regulations containing

security issues that affect the product. Define any security or privacy certifications that must be satisfied.

Software Quality Attributes:
Specify any additional quality characteristics for the product that will be important to either the customers or the developers. Some to consider are: adaptability, availability, correctness, flexibility, interoperability, maintainability, portability, reliability, reusability, robustness, testability, and usability. Write these to be specific, quantitative, and verifiable when possible. At the least, clarify the relative preferences for various attributes, such as ease of use over ease of learning.

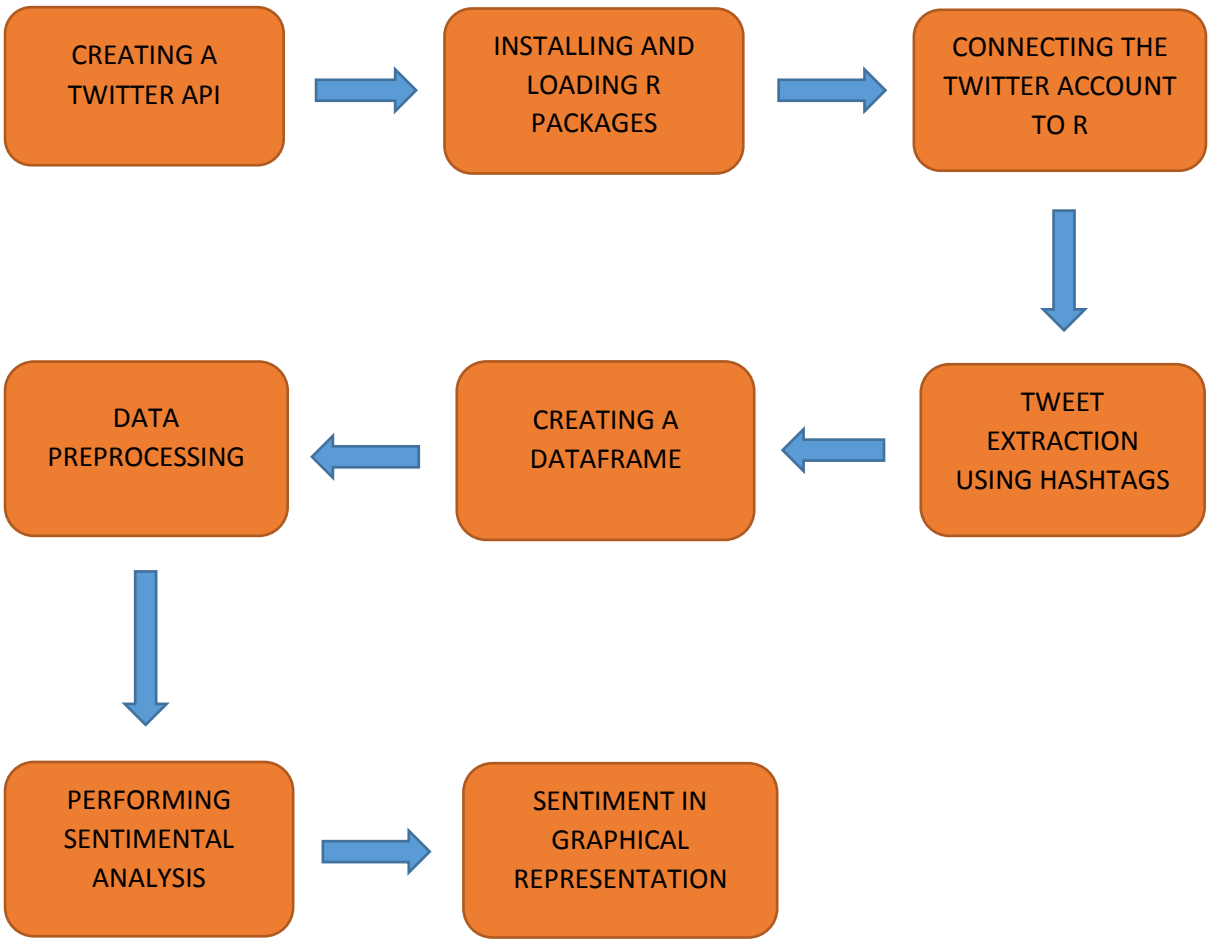
Other Requirements:

- Linux operating systems/Windows
- R statistical software
- R studio
- Sublime text 3
- Modern web browser (Google chrome, Mozilla Firefox)
- Twitter API, Google API

CHAPTER-4

DESIGN APPROACH

The general flowchart of the structure of the project is given as:



Algorithm:

- Create a twitter application programming interface (API) which helps to link the twitter to the R software.
- R provides a varied set of packages which helps to execute various set of instructions. We install and load the required R packages. **[Refer to the appendix section at the end of the project]**
- Connect the twitter account to R using the API key, secret key, access token secret, access token from the twitter.
- Now we write the code in R for the tweet extraction using the hashtags.
- The tweet extracted is generated in the form of "list" as an output and hence it is converted to the data frame which is the feasible format for carrying out the text processing and the sentiment analysis.
- The sentiment analysis is supported with the help of visuals such as the wordcloud and the bar plot which can be generated by using the extensive libraries in R.

The detailed description of the step by step process of the algorithm is described in the next chapter.

CHAPTER-5

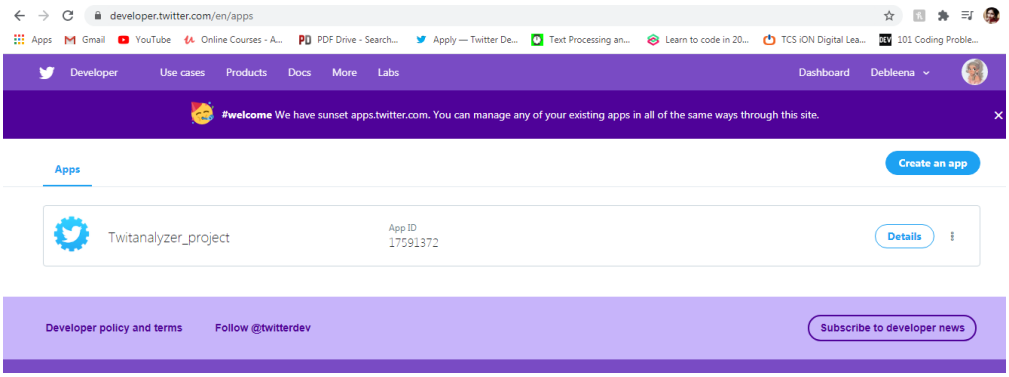
DEVELOPMENT METHODOLOGY

Text processing and the sentiment analysis constitutes the major topics of our project. For carrying out both the processes, we need to extract data from the twitter. We have to connect the twitter to the R software to carry out the procedure.

Creating a twitter API

To connect the twitter account with the R software, we need to create a twitter API. For creating a twitter API we need to follow the required steps:

1. We need to go to the following link https://developer.twitter.com/en/apps and click on the "create an app" option to create a new application. For our project we have already created an app called "Twitanalyzer_project" which is shown in the following screenshot below:



2. On clicking the above option we will get a window which will require to fill up the following details:
   App details, app name, application description, website URL, and we need to select few options.
3. After all the details are verified, we will be granted the API key, secret key, access token secret and the access token which is the required gateway to connect our twitter account to R software.

Loading the required R packages

R provides an extensive library of packages. For performing the text processing and the sentiment analysis, we need to install and load certain R packages. For the complete list of the packages used in the code, refer to the appendix section at the end of the project.

The list of the packages used in our code are:

```
install.packages("twitteR")
install.packages("ROAuth")
install.packages("NLP")
install.packages("syuzhet")
install.packages("tm")
install.packages("SnowballC")
install.packages("stringi")
install.packages("topicmodels")
install.packages("RColorBrewer")
install.packages("wordcloud")
install.packages("ggplot2")



library("twitteR")
library("ROAuth")
library("NLP")
library("syuzhet")
library("tm")
library("SnowballC")
library("stringi")
library("topicmodels")
library("RColorBrewer")
library("wordcloud")
library("ggplot2")
```

- The twitter account is connected to R software by connecting with the keys generated after creating the twitter API.

```
consumer_key <- 'jWUrQ8zKbsTEyD5F5qbUbQFdt'
consumer_secret <- 'B8fbQFW8yCHqlwys4FfPIEVLRVu8qJoiMKdnKCRc3mRClXLJbx'
access_token <- '1238733730719485954-UGCjewMYRpNcAziJotflZpfR0Gprq5'
access_secret <- 'zTFLOdVa4zeEFEJ37z3o3Vc2HuOuhzVNqrhXLDzd8gQRj'
setup_twitter_oauth(consumer_key,consumer_secret,access_token,access_secret)
```

- The tweets are extracted using the hashtags of the topics. Special features such as the usage of emoticons and hashtags carry an analytical value. Hashtags are labels used to conduct search and categorization of the data and are included in the text, prepended by the use of “#” sign. Here we develop a code for the tweet extraction using some of the popular hashtags used in the recent times. For the full algorithm refer to the code section. The extracted tweets are generated in the form of “lists” which does not support text processing or the sentiment analysis. Therefore it needs to be converted to “data frames” to support the process.

  In our project we choose to select four popular hashtags which are “#coronavirus”,”#bollywood”,”#technology” and “#facebook”.

  The data frames generated in the process is given below

7	Stocks and shares tumble, #freight rates climb. I>	FALSE	0	NA	2020-03-24 13:52:27	TRUE	NA
8	RT @DrDenaGrayson: <U+0001F6A8>BREAKING: A man DI>	FALSE	0	NA	2020-03-24 13:52:27	FALSE	NA
9	RT @SkyNews: BREAKING: Singapore has re-introduce>	FALSE	0	NA	2020-03-24 13:52:27	FALSE	NA
10	RT @larrykim: #Olympics2020 new Olympic logo for >	FALSE	0	NA	2020-03-24 13:52:27	FALSE	NA
11	.@katvnews: Arkansas #coronavirus cases up to 206>	FALSE	0	NA	2020-03-24 13:52:27	TRUE	NA
12	RT @StefSimanowitz: 1/3. On 13 March, @Channel4Ne>	FALSE	0	NA	2020-03-24 13:52:27	FALSE	NA
13	RT @MumbaiPolice: For once (and only once), the s>	FALSE	0	NA	2020-03-24 13:52:27	FALSE	NA
14	Respected Chief Secretary, \nBVR Subrahmaniyam Ji>	FALSE	0	NA	2020-03-24 13:52:26	TRUE	NA
15	RT @andreasharsono: PM Shinzo Abe of Japan <U+000>	FALSE	0	NA	2020-03-24 13:52:26	FALSE	NA
16	RT @AmyKremer: After Pelosi, Schumer and the Sena>	FALSE	0	NA	2020-03-24 13:52:26	FALSE	NA
17	RT @TVMP_Pride: 'No evidence' that your pet can g>	FALSE	0	NA	2020-03-24 13:52:26	FALSE	NA
18	RT @RajatSharmaLive: Vice-President Venkaiah Naid>	FALSE	0	NA	2020-03-24 13:52:26	FALSE	NA
19	RT @TrendsSuriyaKL: Exclusive :#CoronaVirus Aware>	FALSE	0	NA	2020-03-24 13:52:26	FALSE	NA
20	RT @Being_Punjabii: #DiseaseFree_With_TrueWorship>	FALSE	0	NA	2020-03-24 13:52:26	FALSE	NA

- Performing the natural language processing on the textual data from Twitter data brings a new set of challenges because of the informal state of the data. Tweets can often contain misspelt words, and the introduction of restrictive word limit of up to 140 characters often encourages the usage of abbreviations and slang. Unconventional linguistic means are also used, such as capitalisation and elongation of words to lay emphasis on certain piece of text. The texts are full of special characters and the unnecessary data which can create a hindrance to the proper text analysis. Therefore it becomes extremely important to pre-process the data before carrying out the process of text analysis. Pre-processing text information also includes removal of stemming words. An example of stemming words are "computer", "computational" and "computation" to the root "compute". The process of pre-processing involves converting all the words into lowercase, removal of links to the web-pages (http elements), and deleting punctuations as well as stop words.

A brief introduction to stopwords

Stop words are the most common terms we hear when we are working with text mining applications. They are also called "stop word list" or even ""stop list". Stop words are the set of commonly used words in any language, other than English. The reason stop words are needed to be removed because then the most desired and important keywords can be focused upon easily. Stop words can mean different things to different applications. To some applications, however, removal of the stop words can prove to be detrimental. For example, the removal of stop words which include the adjectives which indicates the positive sentiments such as "good", "nice" or the negative sentiments such as "not" can throw the algorithms off their tracks.

- We have completed with the data pre-processing part and we finally jump to the part of the analysis. Now we perform the sentiment analysis of the processed data we obtain after the tweet extraction. Sentiment analysis or opinion mining as it is also called, is a research area that deals with extracting the opinions or emotions conveyed in the text, and it is commonly handled as natural language processing (NLP) task. The sentiment information is found across various domains such as product reviews, blogs comments, movie reviews and many other sources of textual information**. [Refer to chapter 2 for the overview of the Sentiment Analysis]**

- The code calculates the score of each sentiment. It assigns a “positive” or “negative” sentiment to each word in its lexicon. The overall score of the expression is calculated by adding up the scores of each expression, which helps in the further sentiment analysis. If the overall score of a piece of text adds up to be a positive number, then the output of the sentimental analysis is observed to be a positive one. If the overall score turns out to be negative, then the text is turned out to be of a negative sentiment. If the overall score is equal to zero, then the text turns out to be of a neutral sentiment.

CODE FOR THE TWEET EXTRACTION AND THE SENTIMENT ANALYSIS

Code for #coronavirus

```
install.packages("twitteR")
install.packages("ROAuth")
install.packages("NLP")
install.packages("syuzhet")
install.packages("tm")
install.packages("SnowballC")
install.packages("stringi")
install.packages("topicmodels")
install.packages("RColorBrewer")
install.packages("wordcloud")
install.packages("ggplot2")
library("twitteR")
library("ROAuth")
library("NLP")
library("syuzhet")
library("tm")
library("SnowballC")
library("stringi")
library("topicmodels")
library("RColorBrewer")
library("wordcloud")
library("ggplot2")

consumer_key <- 'jWUrQ8zKbsTEyD5F5qbUbQFdt'
consumer_secret <- 'B8fbQFW8yCHqlwys4FfPIEVLRVu8qJoiMKdnKCRc3mRClXLJbx'
access_token <- '1238733730719485954-UGCjewMYRpNcAziJotflZpfR0Gprq5'
access_secret <- 'zTFLOdVa4zeEFEJ37z3o3Vc2HuOuhzVNqrhXLDzd8gQRj'
setup_twitter_oauth(consumer_key,consumer_secret,access_token,access_secret)

tweets_c <- searchTwitter("#coronavirus",n=1000,lang="en")
corona_tweets <- twListToDF(tweets_c)
View(corona_tweets)
corona_text <- corona_tweets$text
```

```
corona_text <- tolower(corona_text)
corona_text <- gsub("@\\w+","",corona_text)
corona_text <- gsub("http\\w+","",corona_text)
corona_text <- gsub("[ |\t]{2,}","",corona_text)
corona_text <- gsub("^ ","",corona_text)
corona_text <- gsub(" $","",corona_text)
docs <- Corpus(VectorSource(corona_text))
inspect(docs)
toSpace <- content_transformer(function (x , pattern ) gsub(pattern, " ", x))
docs <- tm_map(docs, toSpace, "/")
docs <- tm_map(docs, toSpace, "@")
docs <- tm_map(docs, toSpace, "\\|")
docs <- tm_map(docs, content_transformer(tolower))
docs <- tm_map(docs, removeNumbers)
docs <- tm_map(docs, removeWords, stopwords("english"))
docs <- tm_map(docs, removeWords, c("blabla1", "blabla2"))
docs <- tm_map(docs, removePunctuation)
docs <- tm_map(docs, stripWhitespace)
docs <- tm_map(docs, stemDocument)
dtm <- TermDocumentMatrix(docs)
m <- as.matrix(dtm)
v <- sort(rowSums(m),decreasing=TRUE)
d <- data.frame(word = names(v),freq=v)
head(d, 10)
set.seed(1234)
wordcloud(words = d$word, freq = d$freq, min.freq = 1,
        max.words=200, random.order=FALSE, rot.per=0.35,
        colors=brewer.pal(8, "Dark2"))

mysentiment_corona<-get_nrc_sentiment((corona_text))
Sentimentscores_corona<-data.frame(colSums(mysentiment_corona[,]))
names(Sentimentscores_corona)<-"Score"

Sentimentscores_corona<-cbind("sentiment"=rownames(Sentimentscores_corona),Sentimentscores_corona)
rownames(Sentimentscores_corona)<-NULL

ggplot(data=Sentimentscores_corona,aes(x=sentiment,y=Score))+geom_bar(aes(fill=sentiment),stat =
"identity")+
  theme(legend.position="none")+
  xlab("Sentiments")+ylab("scores")+ggtitle("Sentiments of people behind the tweets on the outbreak of the
pandemic COVID-19")
```

Code for #bollywood

```
install.packages("twitteR")
install.packages("ROAuth")
install.packages("NLP")
install.packages("syuzhet")
install.packages("tm")
install.packages("SnowballC")
install.packages("stringi")
install.packages("topicmodels")
install.packages("RColorBrewer")
install.packages("wordcloud")
install.packages("ggplot2")
library("twitteR")
library("ROAuth")
library("NLP")
library("syuzhet")
library("tm")
library("SnowballC")
library("stringi")
library("topicmodels")
library("RColorBrewer")
library("wordcloud")
library("ggplot2")
consumer_key <- 'jWUrQ8zKbsTEyD5F5qbUbQFdt'
consumer_secret <- 'B8fbQFW8yCHqlwys4FfPIEVLRVu8qJoiMKdnKCRc3mRClXLJbx'
access_token <- '1238733730719485954-UGCjewMYRpNcAziJotflZpfR0Gprq5'
access_secret <- 'zTFLOdVa4zeEFEJ37z3o3Vc2HuOuhzVNqrhXLDzd8gQRj'
setup_twitter_oauth(consumer_key,consumer_secret,access_token,access_secret)
tweets_b <- searchTwitter("#bollywood",n=1000,lang="en")
bolly_tweets <- twListToDF(tweets_b)
bolly_text <- bolly_tweets$text
bolly_text <- tolower(bolly_text)
bolly_text <- gsub("@\\w+","",bolly_text)
bolly_text <- gsub("[[:punct:]]","",bolly_text)
bolly_text <- gsub("http\\w+","",bolly_text)
bolly_text <- gsub("[ |\t]{2,}","",bolly_text)
bolly_text <- gsub("^ ","",bolly_text)
bolly_text <- gsub(" $","",bolly_text)
docs <- Corpus(VectorSource(bolly_text))
inspect(docs)
toSpace <- content_transformer(function (x , pattern ) gsub(pattern, " ", x))
docs <- tm_map(docs, toSpace, "/")
docs <- tm_map(docs, toSpace, "@")
docs <- tm_map(docs, toSpace, "\\|")
```

```
docs <- tm_map(docs, content_transformer(tolower))
docs <- tm_map(docs, removeNumbers)
docs <- tm_map(docs, removeWords, stopwords("english"))
docs <- tm_map(docs, removeWords, c("blabla1", "blabla2"))
docs <- tm_map(docs, removePunctuation)
docs <- tm_map(docs, stripWhitespace)
docs <- tm_map(docs, stemDocument)
dtm <- TermDocumentMatrix(docs)
m <- as.matrix(dtm)
v <- sort(rowSums(m),decreasing=TRUE)
d <- data.frame(word = names(v),freq=v)
head(d, 10)
set.seed(1234)
wordcloud(words = d$word, freq = d$freq, min.freq = 1,
        max.words=200, random.order=FALSE, rot.per=0.35,
  colors=brewer.pal(8, "Dark2"))

mysentiment_bolly<-get_nrc_sentiment((bolly_text))
Sentimentscores_bolly<-data.frame(colSums(mysentiment_bolly[,]))
names(Sentimentscores_bolly)<-"Score"
Sentimentscores_bolly<-cbind("sentiment"=rownames(Sentimentscores_bolly),Sentimentscores_bolly)
rownames(Sentimentscores_bolly)<-NULL
ggplot(data=Sentimentscores_bolly,aes(x=sentiment,y=Score))+geom_bar(aes(fill=sentiment),stat =
"identity")+
  theme(legend.position="none")+
  xlab("Sentiments")+ylab("scores")+ggtitle("Sentiments of people behind the tweets on BOLLYWOOD film
industry")
```

Code for #technology

```
install.packages("twitteR")
install.packages("ROAuth")
install.packages("NLP")
install.packages("syuzhet")
install.packages("tm")
install.packages("SnowballC")
install.packages("stringi")
install.packages("topicmodels")
install.packages("RColorBrewer")
install.packages("wordcloud")
install.packages("ggplot2")

library("twitteR")
library("ROAuth")
library("NLP")
library("syuzhet")
library("tm")
library("SnowballC")
library("stringi")
library("topicmodels")
library("RColorBrewer")
library("wordcloud")
library("ggplot2")

consumer_key <- 'jWUrQ8zKbsTEyD5F5qbUbQFdt'
consumer_secret <- 'B8fbQFW8yCHqlwys4FfPIEVLRVu8qJoiMKdnKCRc3mRClXLJbx'
access_token <- '1238733730719485954-UGCjewMYRpNcAziJotflZpfR0Gprq5'
access_secret <- 'zTFLOdVa4zeEFEJ37z3o3Vc2HuOuhzVNqrhXLDzd8gQRj'
setup_twitter_oauth(consumer_key,consumer_secret,access_token,access_secret)

tweets_t <- searchTwitter("#ai",n=1000,lang="en")
tech_tweets <- twListToDF(tweets_t)
tech_text <- tech_tweets$text
tech_text <- tolower(tech_text)
tech_text <- gsub("@\\w+","",tech_text)
tech_text <- gsub("[[:punct:]]","",tech_text)
tech_text <- gsub("http\\w+","",tech_text)
tech_text <- gsub("[ |\t]{2,}","",tech_text)
tech_text <- gsub("^ ","",tech_text)
tech_text <- gsub(" $","",tech_text)
docs <- Corpus(VectorSource(tech_text))
inspect(docs)
toSpace <- content_transformer(function (x , pattern ) gsub(pattern, " ", x))
```

```
docs <- tm_map(docs, toSpace, "/")
docs <- tm_map(docs, toSpace, "@")
docs <- tm_map(docs, toSpace, "\\|")
docs <- tm_map(docs, content_transformer(tolower))
docs <- tm_map(docs, removeNumbers)
docs <- tm_map(docs, removeWords, stopwords("english"))
docs <- tm_map(docs, removeWords, c("blabla1", "blabla2"))
docs <- tm_map(docs, removePunctuation)
docs <- tm_map(docs, stripWhitespace)
docs <- tm_map(docs, stemDocument)
dtm <- TermDocumentMatrix(docs)
m <- as.matrix(dtm)
v <- sort(rowSums(m),decreasing=TRUE)
d <- data.frame(word = names(v),freq=v)
head(d, 10)
set.seed(1234)
wordcloud(words = d$word, freq = d$freq, min.freq = 1,
        max.words=200, random.order=FALSE, rot.per=0.35,
        colors=brewer.pal(8, "Dark2"))

mysentiment_tech<-get_nrc_sentiment((tech_text))
Sentimentscores_tech<-data.frame(colSums(mysentiment_tech[,]))
names(Sentimentscores_tech)<-"Score"
Sentimentscores_tech<-cbind("sentiment"=rownames(Sentimentscores_tech),Sentimentscores_tech)
rownames(Sentimentscores_tech)<-NULL
ggplot(data=Sentimentscores_tech,aes(x=sentiment,y=Score))+geom_bar(aes(fill=sentiment),stat =
"identity")+
theme(legend.position="none")+
xlab("Sentiments")+ylab("scores")+ggtitle("Sentiments of people behind the tweets on the overall
technology")
```

Code for #facebook

```
install.packages("twitteR")
install.packages("ROAuth")
install.packages("NLP")
install.packages("syuzhet")
install.packages("tm")
install.packages("SnowballC")
install.packages("stringi")
install.packages("topicmodels")
install.packages("RColorBrewer")
install.packages("wordcloud")
install.packages("ggplot2")

library("twitteR")
library("ROAuth")
library("NLP")
library("syuzhet")
library("tm")
library("SnowballC")
library("stringi")
library("topicmodels")
library("RColorBrewer")
library("wordcloud")
library("ggplot2")

consumer_key <- 'jWUrQ8zKbsTEyD5F5qbUbQFdt'
consumer_secret <- 'B8fbQFW8yCHqlwys4FfPIEVLRVu8qJoiMKdnKCRc3mRClXLJbx'
access_token <- '1238733730719485954-UGCjewMYRpNcAziJotflZpfR0Gprq5'
access_secret <- 'zTFLOdVa4zeEFEJ37z3o3Vc2HuOuhzVNqrhXLDzd8gQRj'
setup_twitter_oauth(consumer_key,consumer_secret,access_token,access_secret)
tweets_f <- searchTwitter("#socialmedia",n=1000,lang="en")
facebook_tweets <- twListToDF(tweets_f)
View(facebook_tweets)
facebook_text <- facebook_tweets$text
facebook_text <- tolower(facebook_text)
facebook_text <- gsub("@\\w+","",facebook_text)
facebook_text <- gsub("[[:punct:]]","",facebook_text)
facebook_text <- gsub("http\\w+","",facebook_text)
facebook_text <- gsub("[ |\t]{2,}","",facebook_text)
facebook_text <- gsub("^ ","",facebook_text)
facebook_text <- gsub(" $","",facebook_text)
docs <- Corpus(VectorSource(facebook_text))
inspect(docs)
toSpace <- content_transformer(function (x , pattern ) gsub(pattern, " ", x))
```

```
docs <- tm_map(docs, toSpace, "/")
docs <- tm_map(docs, toSpace, "@")
docs <- tm_map(docs, toSpace, "\\|")
docs <- tm_map(docs, content_transformer(tolower))
docs <- tm_map(docs, removeNumbers)
docs <- tm_map(docs, removeWords, stopwords("english"))
docs <- tm_map(docs, removeWords, c("blabla1", "blabla2"))
docs <- tm_map(docs, removePunctuation)
docs <- tm_map(docs, stripWhitespace)
docs <- tm_map(docs, stemDocument)
dtm <- TermDocumentMatrix(docs)
m <- as.matrix(dtm)
v <- sort(rowSums(m),decreasing=TRUE)
d <- data.frame(word = names(v),freq=v)
head(d, 10)

set.seed(1234)
wordcloud(words = d$word, freq = d$freq, min.freq = 1,
        max.words=200, random.order=FALSE, rot.per=0.35,
        colors=brewer.pal(8, "Dark2"))

mysentiment_facebook<-get_nrc_sentiment((facebook_text))
Sentimentscores_facebook<-data.frame(colSums(mysentiment_facebook[,]))
names(Sentimentscores_facebook)<-"Score"
Sentimentscores_facebook<-
cbind("sentiment"=rownames(Sentimentscores_facebook),Sentimentscores_facebook)
rownames(Sentimentscores_facebook)<-NULL
ggplot(data=Sentimentscores_facebook,aes(x=sentiment,y=Score))+geom_bar(aes(fill=sentiment),stat =
"identity")+
  theme(legend.position="none")+
  xlab("Sentiments")+ylab("scores")+ggtitle("Sentiments of people behind the tweets on the social
networking site FACEBOOK")
```

- Data visualizations (such as graphs, charts, infographics etc.) provide the text analysts a valuable and easier way to communicate important information and observations at a glance. There are various ways of using a stunning visualization process to highlight important textual points such as creating a wordcloud, and using graphical visualization. These can help to make dull data sizzle and provide crucial information. Word clouds are also known as text clouds or tag clouds. The words which appear more frequently appear bolder and bigger in the wordcloud. *In our project we provided the data visualization using the wordcloud as well as the graphical representation. The various sentiments regarding the particular topics are represented in a visualization form using graphs*. **[Refer to the chapter 6 for the graphical representation and the detailed analysis of the visualization]**

CHAPTER-6

RESULTS AND DISCUSSION

PLOT FOR #CORONAVIRUS

Coronavirus disease (COVID-19) is an infectious disease caused by the newly discovered coronavirus. It is an ongoing global pandemic of the coronavirus disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The outbreak was first identified in Wuhan, China, in December 2019. The World Health Organisation declared the outbreak a Public Health Emergency of International Concern on 30th January, 2020 and a pandemic on 11th march, 2020. More than 11.1 million cases have been recorded so far in more than 188 countries and the union territories, resulting in 528,000 deaths, more than 6.03 million people have recovered.

The main source of the spreading is between the people having close contact, mostly due to the spreading of the droplets via coughing, sneezing and talking. The droplets usually fall to the ground or onto the surfaces rather than travelling through air or long distances. However, research as of June 2020 has shown that speech-generated droplets may remain airborne for tens of minutes. Less commonly, people may become infected by touching a contaminated surface and then touching their face. It is most contagious during the first three days after the onset of symptoms, although spread is possible before symptoms appear, and from people who do not show symptoms. Common symptoms include cough, fever, fatigue and shortness of breath. Complications may include pneumonia and acute respiratory distress syndrome. The time from exposure to onset of symptoms is typically around five days but may range from two to fourteen days. There is no known vaccine or specific antiviral treatment.

Recommended preventive measures include hand washing, covering one's mouth while coughing, and maintaining distance from the other people, wearing a face mask in the public settings and monitoring self-isolation for the people who suspect they are infected. **[Source: Wikipedia]**

We have performed the sentiment analysis of the COVID-19 situation for a day, for the four months of March, April, May and June to predict how the situations are during a day of these months***. From the plot as we can see that the maximum score is pointing to the "positive" sentiment during all the four months [figure-1,figure-2,figure-3,figure-4]***, which indicates that in spite of all the negativity around, most people are still trying to stay positive amidst the tough situations. Most people are trying to remain positive by their own means. Many people have tried to find their inner ways of happiness even through these tough times, by discovering new hobbies and new ways of spending the quarantine days. The people have been forced to stay in lockdown and they are utilising this time to reconnect with their friends and families, with whom they were unable to connect due to the lack of time. Most of the times, the people remain engulfed in some form of digital

equipment, scrolling through the various forms of social media. People generally remain active on social media without being socially active. But this lockdown has forced people to be more interactive with their family life and spend more quality time with their family. The lockdown has forced people to start a new hobby which they have been trying to adopt from a long time ago. People are spending time to learn something new. While the world has slowed down from its everyday rat race, it has encouraged people to look at the life in a different way. Most of the sectors have started their remote working schedule, schools and universities have started to shift their classes online. This had led to the creation of various apps which helps in carrying out with the process of remote working schedule and also taking remote classes.

This lockdown has led to a downfall of many business sectors, particularly the travel and hospitality industries, but it has also proved to be a golden opportunity for some industries such as digital entertainment sectors. People living in the quarantine are generally spending most of their time watching Netflix, amazon prime video and many other online digital streaming platforms. Even youtube has also become one of the most common sources of digital business and marketing. Many people are slowly adopting business in a digital way through the form of social media marketing and digital marketing, as the importance of social media platforms are increasing day by day. Business has also increased in the telecom sectors, since there is a rapid usage in the internet. Teenagers are spending their time playing online games which has also led to an increase in the streaming of the online games. Looking at the countries like Spain and Italy, it's being predicted that after lockdown and maintaining social distancing among the people can help in reducing the spread of the deadly virus.

The second maximum plot is the sentiment of "negative" which is varying throughout the period of four months. It can be seen that the plot is maximum during the month of March and April [figure-1 and figure-2] and comparatively lesser during the months of May and June [figure-3 and figure-4]*, *with a slight increase during the month of June. As we are all aware of the deadly virus, it's bound to have negative feelings, during this tough time. As the number of infections are increasing exponentially on a daily basis, it is becoming difficult for the people to maintain their internal peace of mind. Lockdown across the world is hampering the economy as the stock markets are crashing down and the unemployment is increasing mostly in the developed countries. The patients who almost got cured of the disease are being infected again. People are forced to remain in lockdown and the toxicity around them is increasing daily. These results in increase in domestic violence, anxiety, fear and depression. To cope with the deteriorating mental stress, people are getting more addicted to alcohol, smoking and drugs, which can actually create an adverse impact on the immune system. Fake social media news and WhatsApp forwards are creating more stress among the people. As the number of infections and the deaths are increasing on a daily basis across the world, the plots were maximum during the months of March and April. This states that this was the time of the month where the statistics of number of the daily number of infections and the death cases were increasing across the world. Complete lockdown was imposed in many countries to prevent the wide spread of the deadly virus. As we approach towards the months of the May and June, we can see that the situations are

on the verge of improvement, though the fear of the rising infections and the cases still remains. But all the countries have increased their number of the testing and the governments have taken all kinds of steps to curb the spread of infectious diseases. Moreover, the number of cured patients have overtaken the statistics of daily infected patients which forms a source of relief. And if we look at the overall statistics of the death rate, it is very minimal as compared to the number of infections.

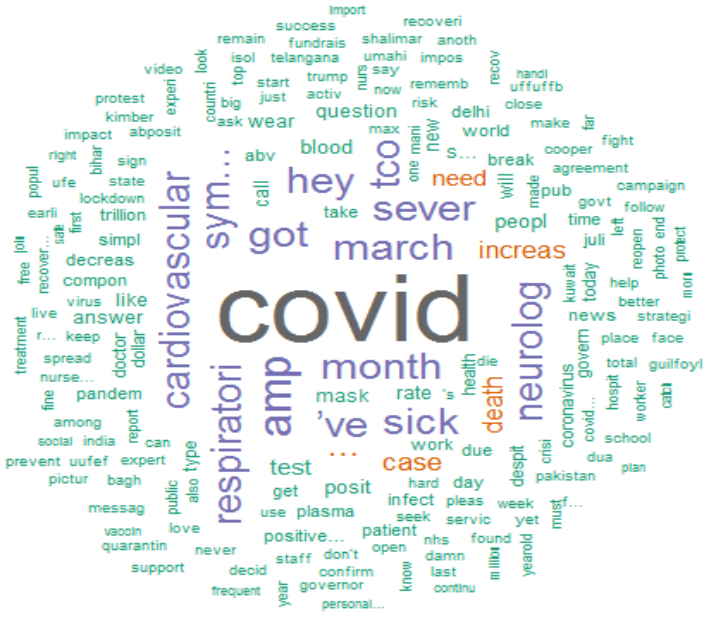
The next comes the feeling of "trust" as ***implied in the plots of all the months [figure-1, figure-2, figure-3, figure-4***] which implies that all the people around the world have immense trust on the doctors, nurses and all the social workers in the community and even people of some sectors in the society who are continuously serving their society and doesn't have the privilege to work from home. The feeling of "fear" also comes into this chapter as these people who are our real superheroes of the society also have a family of their own and do have close people who cares for them. Deep within them, as the number of infections arise and also the rise in the number of deaths, there lies a fear in them***. The feeling of "sadness" [varying throughout the period of four months, denoted by figure-1, figure-2, figure-3, figure-4] is quite obvious as the people are always shocked by the bad news*** surrounding them in this tough times and it is really becoming difficult for the people to remain optimistic. ***The plot of "sadness" decreases from the month of March to the month of June and there has been a constant maximum score of "positive" sentiment throughout the period of the four months***. ***The feeling of "anger" and "disgust" [denoted by the varying score in the plots in all the four months] is coming from the fact that in some countries, this adverse situation is becoming a weapon of the politics, due to which the common citizens have to suffer a lot***. This is causing the people to lose faith in their government and creating a spark for the internal rivalry and hatred among themselves.

The unexpected tragic news is creating a massive shock every day and hence following the news every day always comes with a "shocking surprise" which has more or less maintained a constant score which implies that there was no element of surprising facts regarding this situation for the chosen period.

AREAS THAT NEED FOCUS BASED ON THE SENTIMENT ANALYSIS

- COVID-19 has become a deadly pandemic which has no cure till now except the body's inbuilt immune system. As the number of cases are rising across the world, the negativity surrounding us is increasing and so is the "negative" sentiment. From the month of March to May [Figure-1 to Figure-3], we can see a decrease in the score of the negative sentiment, with a slight exception in the month of June.
- The pandemic is a very new situation for us and our daily social life has been affected badly. There has been a huge disrupt in the world economy. Major business sectors such as airlines, travel, hospitality sectors have faced a huge loss. Many people working in these sectors are experiencing lay-offs. To solve this problem, government needs to provide financial aids to them to support their family. There should be new jobs created for them, so that they can support their family. This pandemic is not a new blow for the travel and the tourism industry. According to the statistics shown by the financial express, It has suffered a huge loss during the 9/11 attack, the SARS epidemic in the areas of China and East Asia during the year 2003**.** **[Source: https://www.financialexpress.com/industry/travel-tourism-sector-has-emerged-from-many-crises-in-past-heres-what-makes-coronavirus-different/1922357/]**. This pandemic is calamitous to a high scale, uncharted before and it has been a huge blow to the travel industry creating a 25% decline in the global travel alone. India is supposed to suffer a loss of five lakh crore and job losses of about four to five crores. The figures depend on the condition of the pandemic and till how long the lockdown persists.
  [Source: https://www.thehindubusinessline.com/companies/impact-of-the-coronavirus-pandemic-might-lead-to-a-bleak-fy21-for-the-hospitality-and-tourism-industry-care-ratings/article31445306.ece]
- The key takeaway from this crisis should be that these sectors and as well as the government should stay prepared for a virus aided pandemic breakout like this. The hygiene measures should be taken in plenty to avoid certain last moment situations and they should try to innovate certain alternative ways of their business, so that both the employers and the employees do not have to suffer in such kind of situations. **To quote, Farhat Jamal**:
  "In the aftermath of COVID-19, we must accept that epidemics and virus breakouts may return to haunt us again in the future. Preparedness should be our key takeaway from this experience," he said. Hotels will not only need to ramp up technology but also ensure availability of basic protective equipment such as masks, infra thermometers, gloves and a set or two of PPEs. The government and the industry also needs to prescribe minimum hygiene and sanitation guidelines and compliance standards."

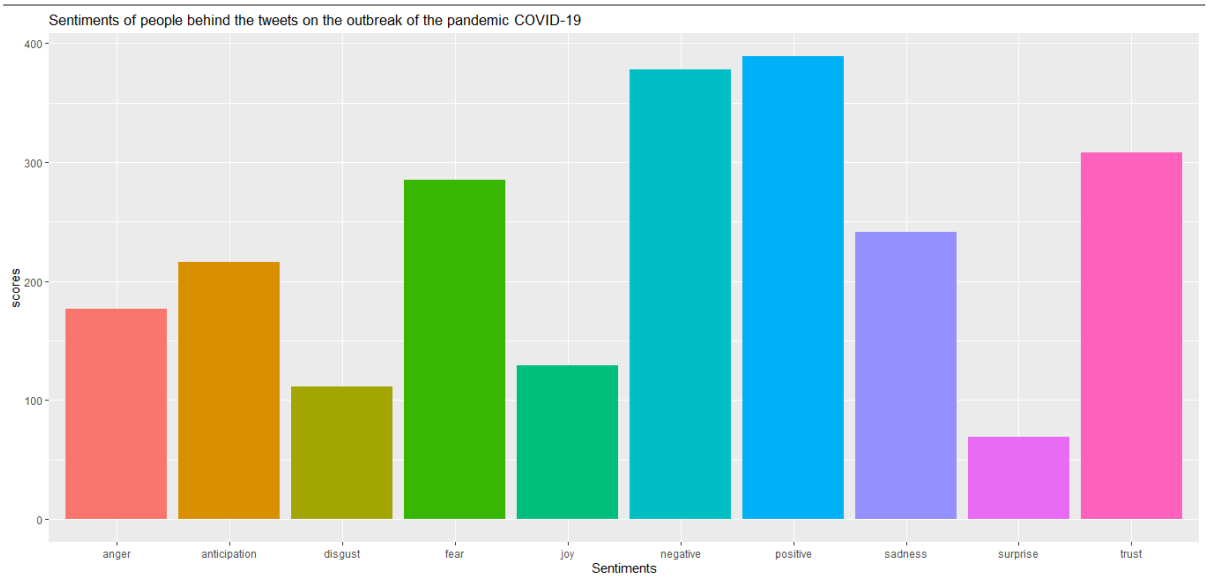
The wordcloud for the COVID-19 disease is given as:



The visual analysis of the opinions regarding the current pandemic of covid-19 for the period of four months is given below:

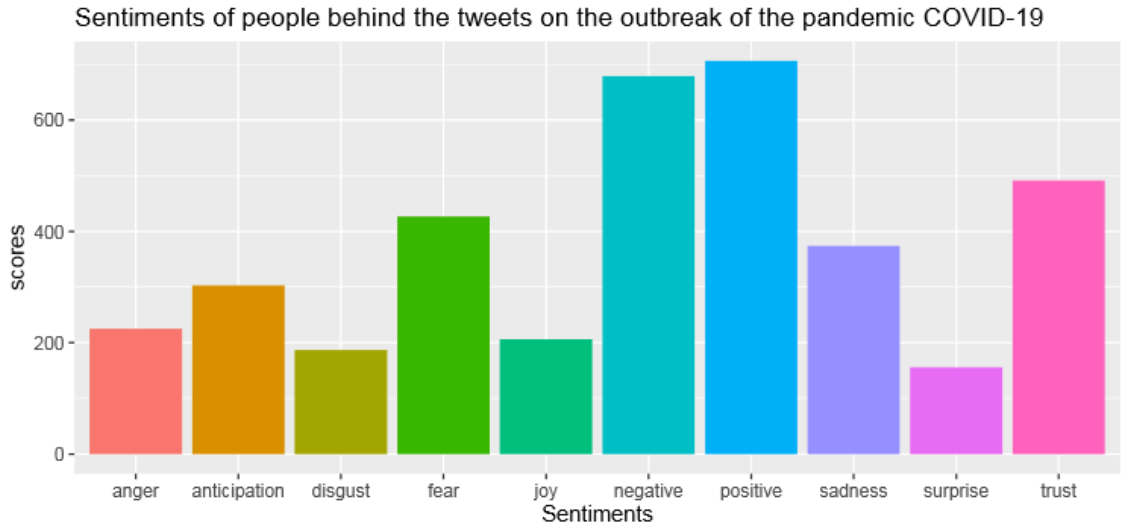
Analysis of the #coronavirus for the month of March

Figure-1



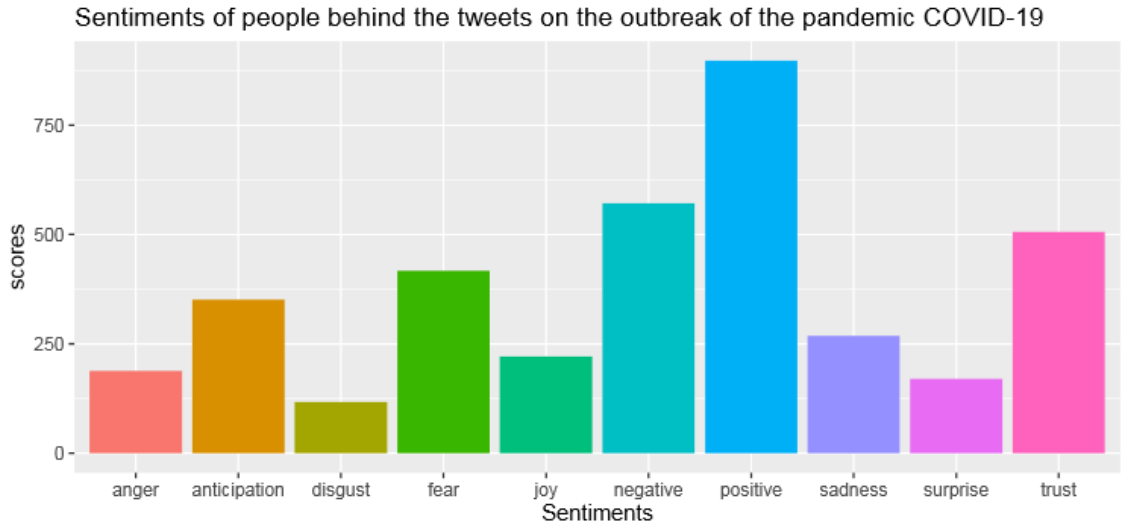
Analysis of the #coronavirus for the month of April

Figure-2



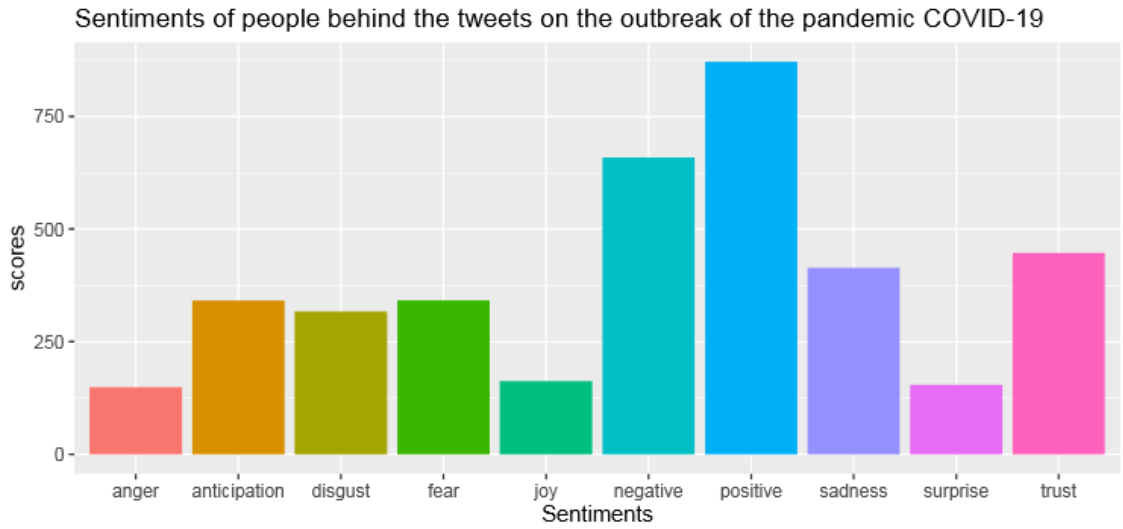
Analysis of the #coronavirus for the month of May

Figure-3



Analysis of the #coronavirus for the month of June

Figure-4



PLOT FOR #BOLLYWOOD

Bollywood (Indian Cinema) was also known as the Bombay Cinema formerly. It is the Indian Hindi language based film industry in Mumbai. The term 'Bollywood' can be claimed as a portmanteau of 'Bombay' and 'Hollywood'. This industry is related to the cinema of the Southern India and the other film industries making the Indian cinema the biggest producer of the world's feature films. Bollywood makes films and movies which depicts our country's culture and heritage. Many foreign countries portray India as they watch in the Indian movies.

Indian cinema has an annual output of 1,986 feature films in 2017. Bollywood is its largest film producer, with 364 Hindi films produced in 2017. Bollywood represents 43 percent of Indian net box-office revenue; Tamil and Telugu cinema represent 36 percent, and the remaining regional cinema constituted 21 percent in 2014. Bollywood is one of the largest centres of film production in the world. In 2001 ticket sales, Indian cinema (including Bollywood) reportedly sold an estimated 3.6 billion tickets worldwide, compared to Hollywood's 2.6 billion tickets sold. Bollywood films tend to use vernacular Hindustani, mutually intelligible by people who self-identify as speaking either Hindi or Urdu, and modern Bollywood movies increasingly incorporate elements of Hinglish.

[Source: Wikipedia]

India is fast becoming a superpower, says Shashi Tharoor, not just through trade and politics, but through the "soft" power, its ability to share its culture to the world through food, music, technology, Bollywood.

In the words of Dr. Shashi Tharoor,

"…Bollywood is now taking a certain aspect of Indian-ness and Indian culture around the globe, not just in the Indian diaspora in the U.S. and the U.K., but to the screens of Arabs and Africans, of Senegalese and Syrians. I've met a young man in New York whose illiterate mother in a village in Senegal takes a bus once a month to the capital city of Dakar, just to watch a Bollywood movie. She can't understand the dialogue. She's illiterate, so she can't read the French subtitles. But these movies are made to be understood despite such handicaps, and she has a great time in the song and the dance and the action. She goes away with stars in her eyes about India, as a result. And this is happening more and more…."

[Source: Why nations should pursue "soft" power | Dr. Shashi Tharoor | TedTalks]

In our project, we have plotted the sentiments regarding the famous Bollywood industry in India for a period of four months, March, April, May and June***. From the bar plots we can say that the people have a general positive sentiment regarding the industry [figure-1, figure-2, figure-3], which shows an exception in the month of June [figure-4]***. ***There has also been a change in the sentiments of "trust", "fear" and "sadness" over a period of time, which can clearly show that there has been a lot of ups and downs in the industry for the past few months [denoted by all the plots].***

Movies are basically assumed to be fictional in most cases, but it also depicts some plots based on the real life issues, circumstances and even incidents. In most movies, the plot, the scenes and the dialogue of the story is scripted and directed in such a way that it creates an everlasting impact on the viewer. It gets them to relate their life with the life of the characters, the incidents and to learn from the character's mistakes. There are a number of iconic films in the Bollywood which teaches us many things. The films based on the life stories of famous people are the most inspirational while the films of mythology and action heroes filled with special effects are a source of entertainment for the small children. ***These comes under the "positive", "trust" and "joy" sentiments [denoted by plots of all the four months]. These sentiments have been constant throughout the period.*** Bollywood movies have always been a source of escape among the Indians in spite of having the internal ups and downs which has been the recent talk of the period.

If we have a look at the scores of the "negative" sentiments, it has undergone a significant rise during the months of May and June. During June, it has the maximum score, surpassing the "positive" sentiment [figure-3, figure-4].

Bollywood is an industry where many of the youngsters aspire to get a chance to work, because of the money, fame and the glamourous lifestyle. It is a field of cut throat competition and nearly one out of ten can strive to that point of success. Others do get chance in Bollywood but can only work for a few films and cannot bear the pressure of the competition. Many of them gets into addiction, partying and many in disciplined habits which worsens their career. Others can get into the mental depression of not getting the desired success in the field. It really becomes more difficult for the outsiders to attain the same amount of stardom since the Bollywood is already filled with famous celebrities. So, the younger generation of the celebrities are given more preference than the outsiders. Not being a part of the star background, can really be a tough situation for the outsiders to create a place for them in the industry. These biasness in the industry is more prominent nowadays making the field more toxic as a part of the career option for the outsiders. Not to mention, the standard of the Bollywood films are really going down. People are preferring to watch more short films and documentaries rather than spending money and watching long hours of movies which have no valuable content in them which can actually create an impact on the audience. Bollywood industry have lost their originality in scripting the films and are generally covering up by making the sequel of already made films or remaking the films which have been already made once. Even the songs of the films are actually the modern version of the old songs filled with electronic effects and some of them are even auto tuned. These are some of the negative impacts Bollywood have on the people's minds. The "negative","sadness","fear","disgust" and "anger" plots reveal these kind of opinions regarding the film industry and also the impact of the movies. This has always been a serious issue in the sector, which has been on the rise since the past few months as the plots score a maximum in the month of June. ***The tragic incident of the suicide of the talented and the charming actor Sushant Singh Rajput has unleashed a dark side of the Bollywood industry and "nepotism" has become the buzzword for the past few days since the month of June and thus the discussions revolving around this topic are mainly about the people discussing the negative sides of the Bollywood and condemning some of the***

dark sides which resulted in the maximum negative score in the month of June. This incident has sent shockwaves across the country and the netizens have been pointing to a certain sector of the Bollywood and there has been a huge gossip throughout the whole country, huge protests have started for banning the Bollywood industry and banning nepotism.

"Bollywood's nepotism didn't start with Karan Johar. But it must end with Sushant Singh Rajput. "

Nepotism-led 'othering' has reached a toxic level in Bollywood. Regardless of the extent of its contribution to Sushant's death, the Hindi film industry must do some soul-searching.

[Source: The Print, 5th July, 2020]

Also, not to mention***, the recent demise of the famous Bollywood stars Rishi Kapoor, Irrfan Khan during the month of April has been the topic of discussion and has accounted for a huge loss in the industry. All the losses and the recent shocking news have led to an increase in the "negative" sentiments for the period of four months in the industry.***

AREAS THAT NEED FOCUS BASED ON THE SENTIMENT ANALYSIS

From the overall plot over the period of four months, we can see that there is a massive increase in the "negative" sentiments in the Bollywood. It is advisable to focus on the negative aspects of a subject and try to bring some healthy changes into it. So here we are diving deep into the topic and analyse how can we try to approach to a solution.

- The gradual increase of the score of the negative sentiment should be of major concern for the industry. This reflects the unhealthy competition and the environment for the people who are striving hard to make a career in the industry and it is also reflecting on the artists who already have a flourishing career. The industry should have a well-balanced environment and the outsiders who are willing to pursue a career in the industry should be given chance to portray their acting skills. Many actors and actresses who have been the outsiders have admitted to the media that they have been the victim of unnecessary partiality and the nepotism. **To quote few articles:**

  "Sharmila tagore's son Saif claims he is a nepotism victim"
  [Source: newsbytesapp.com]

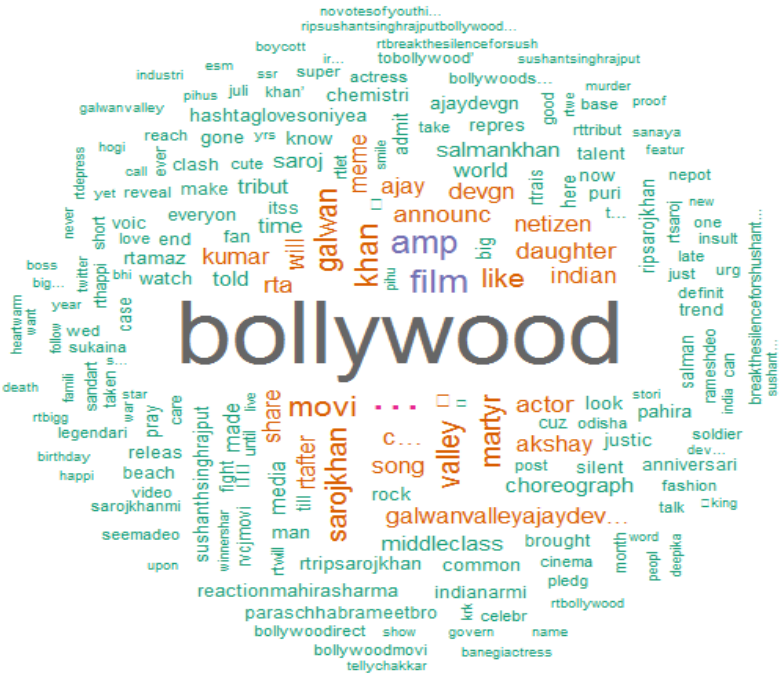
  "Sushant's brother-in-law introduces nepometer to fight nepotism in Bollywood"
  [Source: newsbytesapp.com]

  "13 Bollywood actors who have openly spoke about nepotism in the industry"
  [Source: metrosaga.com]

- The "negative" fact about the industry not only sticks with the tragic incidents like nepotism and partiality, but it also brings light to the facts that Bollywood movies have lost their originality. The entire plot of the movie is either a remake of some movie belonging to a different culture or language or the plot lacks the captivating storyline. Even the songs have been a remake since recent times and it has been the source of complain for many. People are preferring short films or TV series streaming on various OTT platforms which have more original content than the long hour Bollywood movies.
  [Source: https://toistudent.timesofindia.indiatimes.com/news/top-news/has-bollywood-run-out-of-original-ideas/7205.html]
- Days have long gone when people used to depend on the Bollywood movies as the only source of entertainment. Nowadays there are various OTT platforms through which the entire world is connected and people can watch TV commercials and TV series belonging to a different part of the world, whose content are far more original than the Bollywood movies which still remains in the era of the stereotypical romance of virtual reality. To keep up the pace, it really has to step up the game as Bollywood is one of the many industries which represents India and the Indian

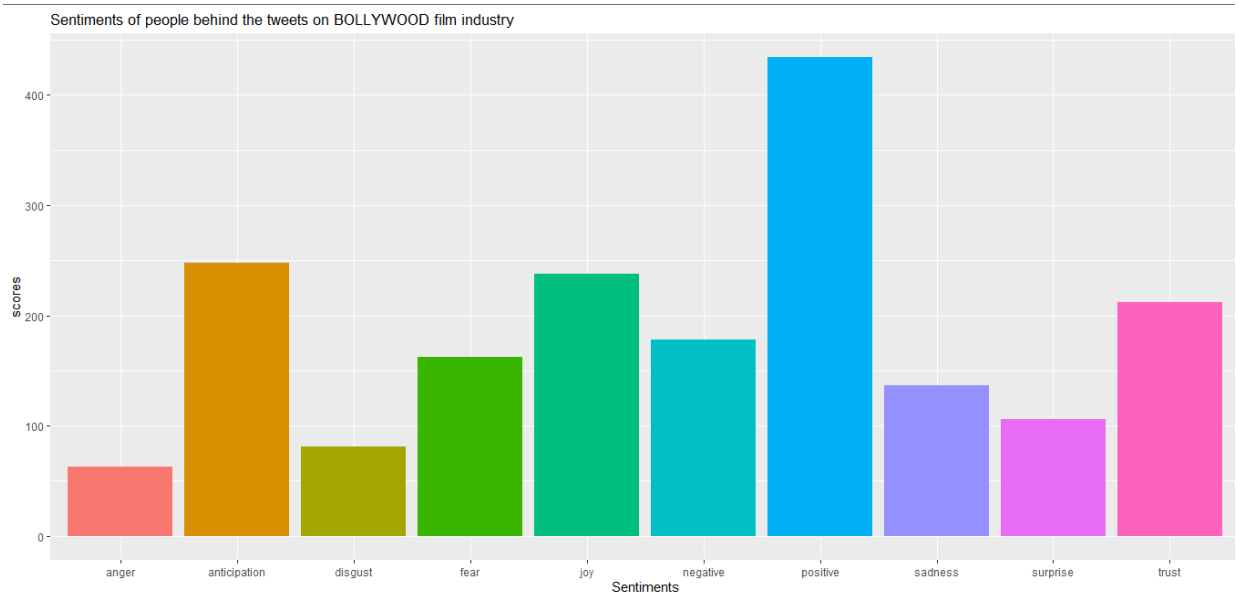
culture across the whole world. The more negative impact the industry has on itself, the more negatively, it will represent India across the whole world.

The wordcloud regarding the #bollywood is given below:



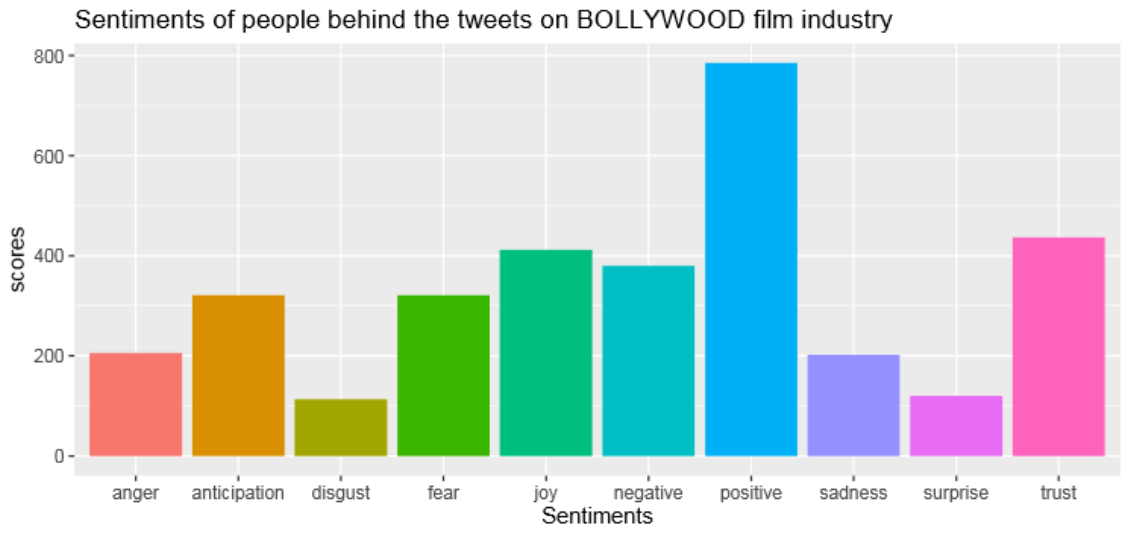
Analysis of the #bollywood for the month of March

Figure-1



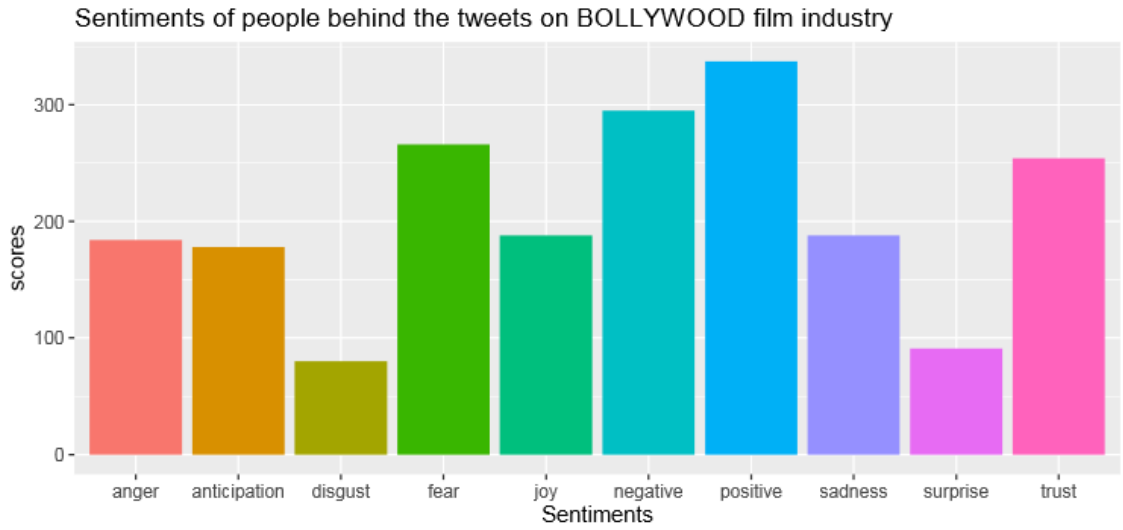
Analysis of the #bollywood for the month of April

Figure-2



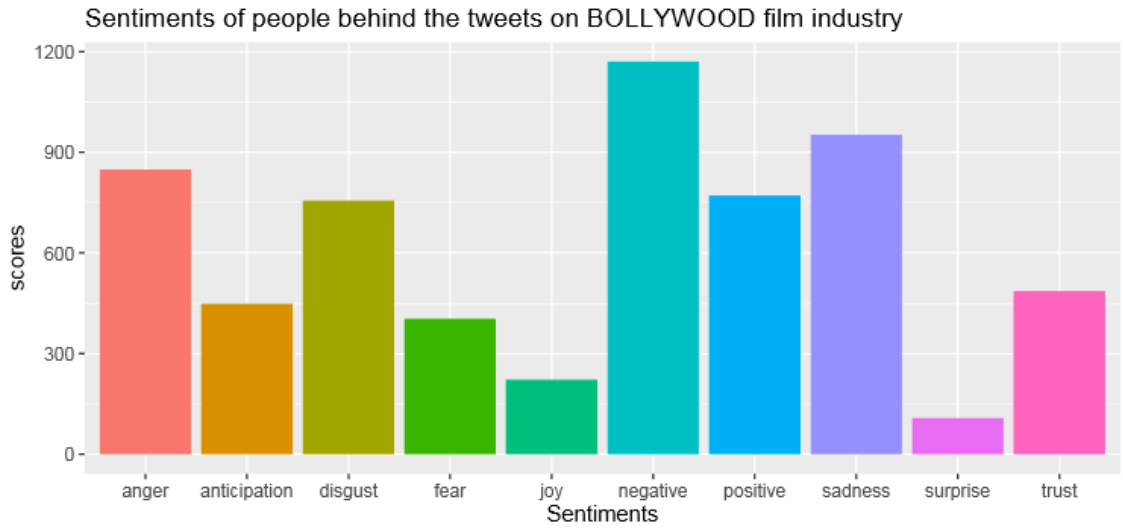
Analysis of the #bollywood for the month of May

Figure-3



Analysis of the #bollywood for the month of June

Figure-4



PLOT FOR #TECHNOLOGY

If we analyse the visual results regarding the sentiments of the overall technology, during the period of four months, then we can see the “positive” plot has the maximum score. The introduction of the technology in our daily lives has improved our way of living. Our lives have been much more comfortable as the technology was introduced in all spheres of our lives slowly. As the technology is advancing now, it is easing our lives. In the modern world, people are making the maximum of the technology as it is evolving significantly with the time. Today life seems impossible with the absence of the digital gadgets and the technologies such as smartphones, TVs and laptops, etc. The continuous evolution of technology has been of the utmost importance to the humans such as the developments made in the field of medical science, has enabled us to treat many health conditions like cancer and other chronic diseases.

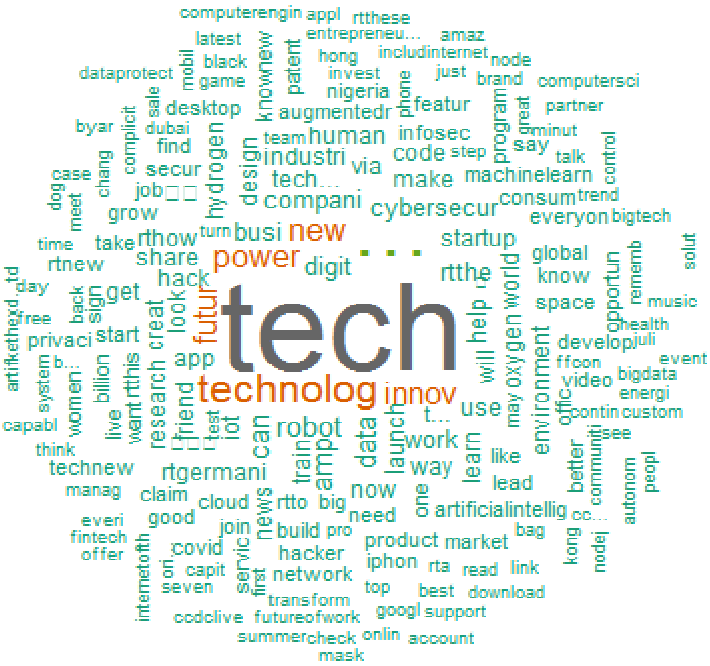
The invention of the internet, computers and the mobile phones have turned out to be a boon for the mankind and has made the process of communication faster, simpler and effective. The inventions of the latest technology and machinery have increased the overall production rate manifolds. As the technology is advancing, new inventions and discoveries are speeding up.

The advancements in the technology have created a more secured environment for the humans. Banking and money management sectors has become more manageable and secured. The invention of webcams, CCTV cameras and surveillance cameras have added an extra advantage in making a more secured environment.

The advancement of the technology has made it easier for the knowledge to be easily accessible. People are well acquainted with the knowledge all across the world. The 21st century has been an era of Science and the Technology. The invention of the World Wide Web has make the world an easy place to live in with the easy access of information across the world. Advancement in the technology has also made things cost effective and less time consuming.

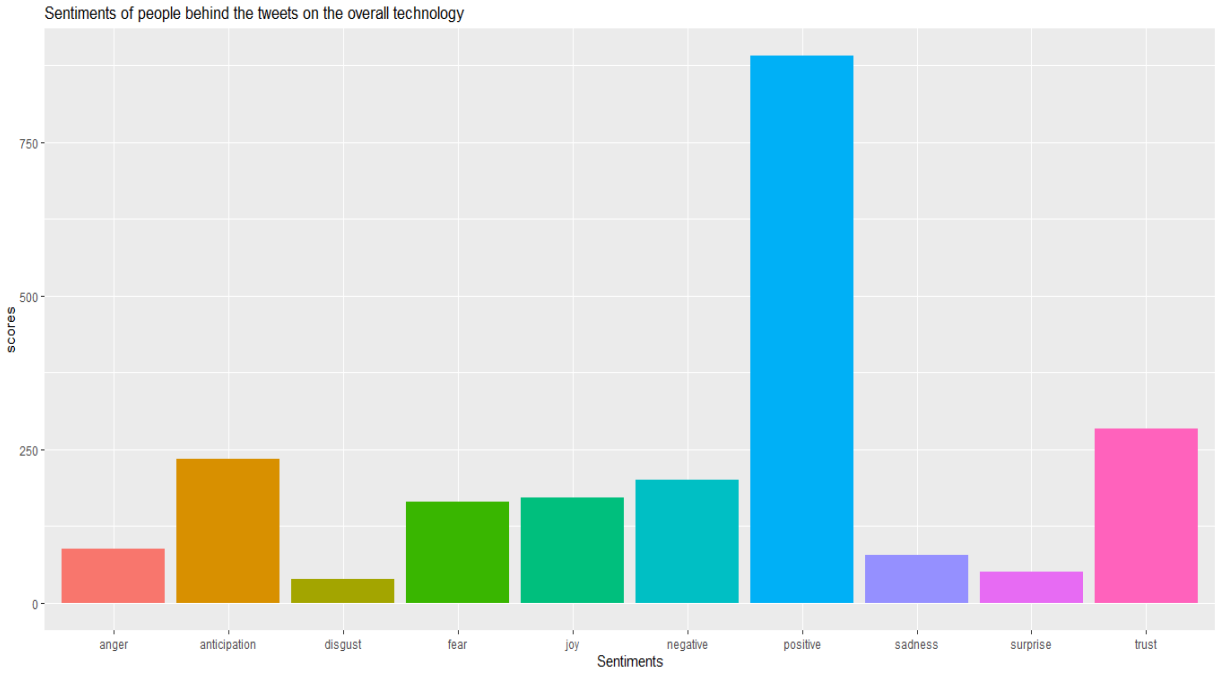
Technology has truly resulted in making our lives much easier and has also led to modernization in many fields. Be the field of medicine, farming or electronics; technology has resulted in a global revolution and advancements. Hence if we analyse the opinions of the people regarding the technology, the score with respect to the “positive” sentiments is the maximum. The “negative” aspects of the technology is much lower as compared to the “positive” aspects of the technology. As it is widely said, there are two sides of a coin. Similarly technology has both its positive and negative aspects. With the advancement in the technology, people are prone to laziness and they have lost their social nature. They are more widely involved in technology and not making the good use of technology can lead to addiction and create adverse effects.

The wordcloud for #technology:



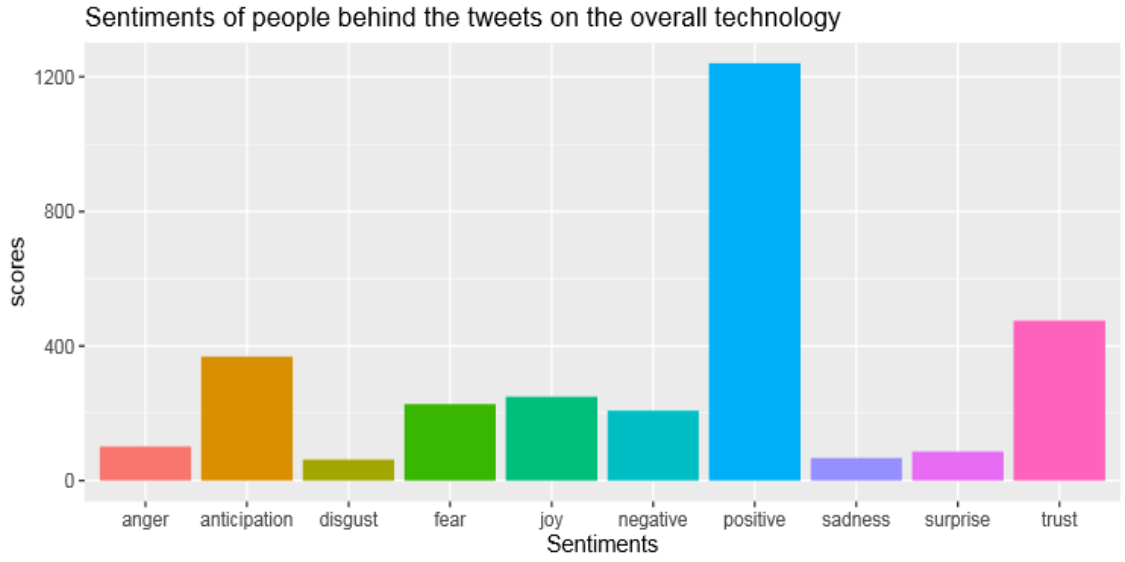
Analysis of the #technology for the month of March

Figure-1



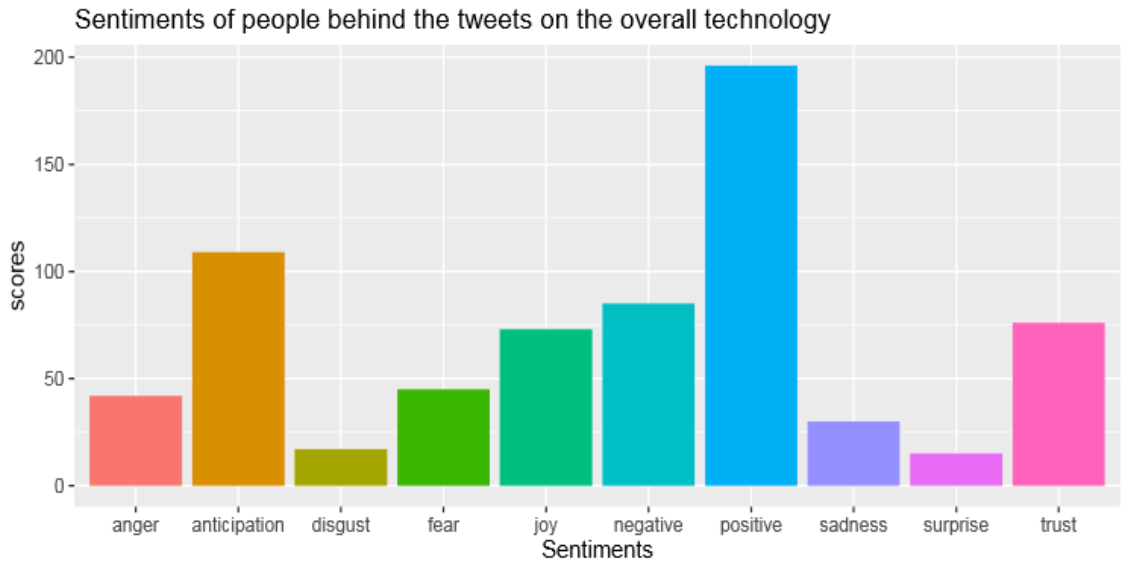
Analysis of the #technology for the month of April

Figure-2



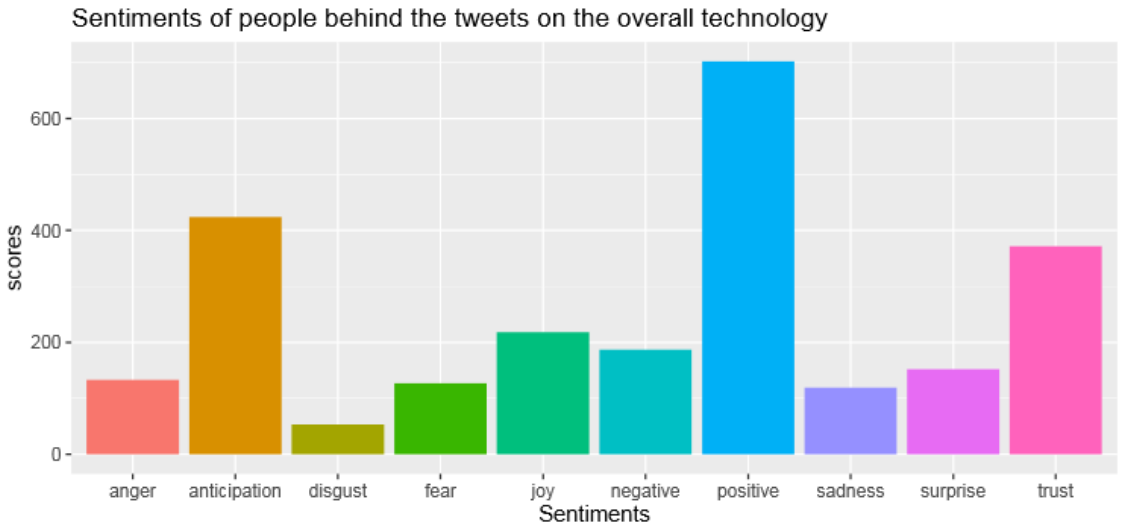
Analysis of the #technology for the month of May

Figure-3



Analysis of the #technology for the month of June

Figure-4



AREAS THAT NEED FOCUS BASED ON THE SENTIMENT ANALYSIS

From the plots of the sentiments behind the hashtag technology, we can clearly see that the positive sentiments of the people obtain a maximum score in all the four months [figure-1 to figure-4]. We need to focus on the positive aspects of the technology and try to implement them to eradicate the negative aspects of any sectors. In the earlier sections, for the plots of the coronavirus and the Bollywood industry, we have seen an equal score of the "negative" sentiments with respect to the "positive" sentiments. In some cases there has been an increase in the negative sentiments too. From the above analysis of the "positive" sentiments regarding the technology, we have a few opinion regarding how it can be used to solve the problems of the other topics.

- Technology is already ruling every domain of our lives. We have seen a huge development in the IT and the software development sectors all across the world. In the current situation of the coronavirus breakout, we have failed to curb the virus and the cases are growing widely in the country. This is because we have not updated the technology in the medical sectors, the backbone of our society. This crisis points out that the government needs to be prepared for such kind of sudden outbreaks which can be more devastating than any war, if the protective measures are not being taken properly.
- The various industrial sectors should learn from this crisis and upgrade their technology or try to introduce certain new technology measures so that they can be prepared for any micro-organism breakout. Many IT companies have already upgraded their workforce and technology to such an extent such that they can continue their "work from home" schedule even during the upcoming years***. The headline, "TCS 'work from home' policy: Only one-fourth of workers to come to office; CEO explains Vision 25×25" [Source: Financial Express]***
- The "Work from home" culture is the new approach that will be adopted by many sectors/industries to maintain the habit of social distancing in the near future even if the crisis is over. The world is bound to change and there will be enhancement in the personal hygiene and the hygiene will be the most important criteria in the workplace. To maintain this level of hygiene while continuing with the smooth operation of the industry, the technology plays an important role in maintaining this balance. New softwares are being invented to provide this balance.
- Technology has also tried to maintain the pace of education for the school and the college going students. Their education has been shifted from the actual classes to virtual classes thereby helping them to continue with their studies even during the lockdown being imposed. The freshers who are in their final year degree courses, applying for jobs, can currently apply for virtual trainings and can opt for digital internships to boost their profile as a graduate. All this is possible because of the advantages in the technology.

- Such approaches should be maintained in the near future and also new innovative approaches should be brought into the plate, to avoid any unnecessary costs and hamper the environment.

PLOT FOR #FACEBOOK

Facebook (styled as **facebook**) is an American online social media and social networking service based in Menlo Park, California and a flagship service of the namesake company Facebook, Inc. It was founded by Mark Zuckerberg, along with fellow Harvard College students and roommates Eduardo Saverin, Andrew McCollum, Dustin Moskovitz and Chris Hughes.

The founders initially limited Facebook membership to Harvard students. Membership was expanded to Columbia, Stanford, and Yale before being expanded to the rest of the Ivy League, MIT, and higher education institutions in the Boston area, then various other universities, and lastly high school students. Since 2006, anyone who claims to be at least 13 years old has been allowed to become a registered user of Facebook, though this may vary depending on local laws. The name comes from the face book directories often given to American university students.

Facebook can be accessed from devices with Internet connectivity, such as personal computers, tablets and smartphones. After registering, users can create a profile revealing information about themselves. They can post text, photos and multimedia which is shared with any other users that have agreed to be their "friend", or, with a different privacy setting, with any reader. Users can also use various embedded apps, join common-interest groups, buy and sell items or services on Marketplace, and receive notifications of their Facebook friends' activities and activities of Facebook pages they follow. Facebook claimed that it had more than 2.3 billion monthly active users as of December 2018.

[Source: Wikipedia]

Facebook is non-arguably one of the most popular social networking sites on the internet. In this visual analysis of the people's opinions regarding Facebook, we can see that there are both varying scores of "positive" and "negative" sentiments throughout the timespan of the four months. According to the "positive" sentiments, there are many advantages of the Facebook. Some of them including, that we can connect with a large number of people from all across the world. It's a major source of entertainment for many people across the globe. Since Facebook is a global social networking site, location doesn't prove to be a barrier.

With the translation feature and many other features in Facebook, it has made a more widely popular social networking site all across the globe. We can connect with different people speaking different languages. Moreover it's easy to find likeminded people on Facebook by looking at the interests, hobbies, likes and dislikes. We can easily connect with people through their walls, private messages and the video chat.

Facebook also allows the establishments of the partnerships between various projects. Moreover using the Facebook page, it can help to increase the brand value of the business.

The major advantage of using the Facebook is that it is a real time social networking site and hence it is the best platform which helps us to remain updated about the latest news and updates. Bloggers and internet marketers can subscribe to popular blog fan pages and keep themselves aware of the latest job updates.

There are several disadvantages of using Facebook are its privacy issues. Even if Facebook has made changes to its privacy issues many outsiders can actually stalk a user's profile to actually gain more information about them and leak their private information. Facebook is full of fake profiles, created mostly by the stalkers and the marketers to gain more friends and to use it for their marketing purposes. Facebook is a vast platform where people share vast topics and is mostly time-consuming and can make a person addictive to Facebook.

All these drawbacks makes up the all the "negative" opinions about Facebook.

AREAS THAT NEED FOCUS BASED ON THE SENTIMENT ANALYSIS

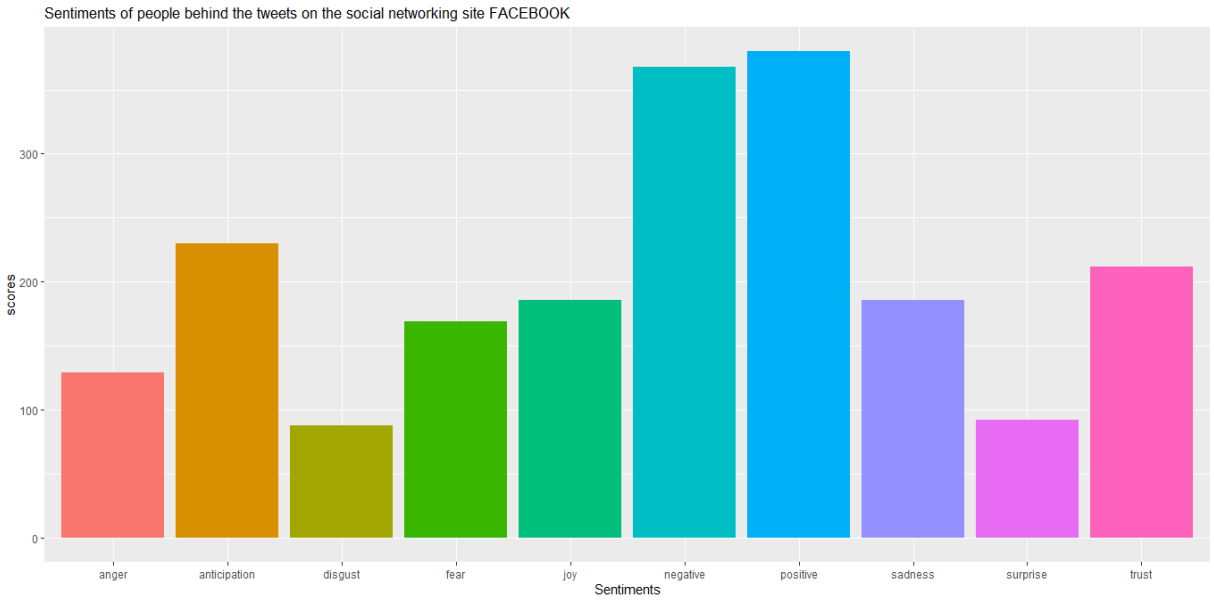
From the observations of the plots of the facebook, we can see that there is not much negativity regarding the sentiments behind this popular social media platform.

- Due to this crisis, people have to maintain social distancing norms and hence they are forced to be in a condition of lockdown to curb the curve of the COVID-19. Social media platforms such as facebook has been the important source of medium to maintain contact with each other.
- Facebook has also served as a medium for the news across the world, regarding every situation of the pandemic. The negative aspect of the facebook is that, people are also striving to spread fake rumours regarding any particular event is happening across the world, creating more panic among the people which should not be done.
- Business sectors have already made their way to collaborate with the digital platforms as these platforms have seen an immense growth in the past few years. in the coming years, this will flourish immensely and the businesses in the social media platforms will be opted as a career opportunity by many students in the near future.
- In the amidst of the coronavirus crisis, facebook has been the platform of many virtual events being organised by sevral organisations. Many event organisers are promoting their events on this platform, and those who are interested, they can register themselves and can access permission to attend the event. This had already started earlier, but this will be of a great importance in the near future which can help to maintain social distancing and can be a preventive measure for the smooth operation of the business and the events while also being prepared for any such kind of pandemic breakout.
- These improvements in the technology and more innovative ideas required will drive more people in the sectors of the technology to learn new things about the technology so that they can pave their way for employment in the new technology sectors. Many startups will require young professional minds to maintain their technological field in the near future.

The wordcloud for the #facebook is:

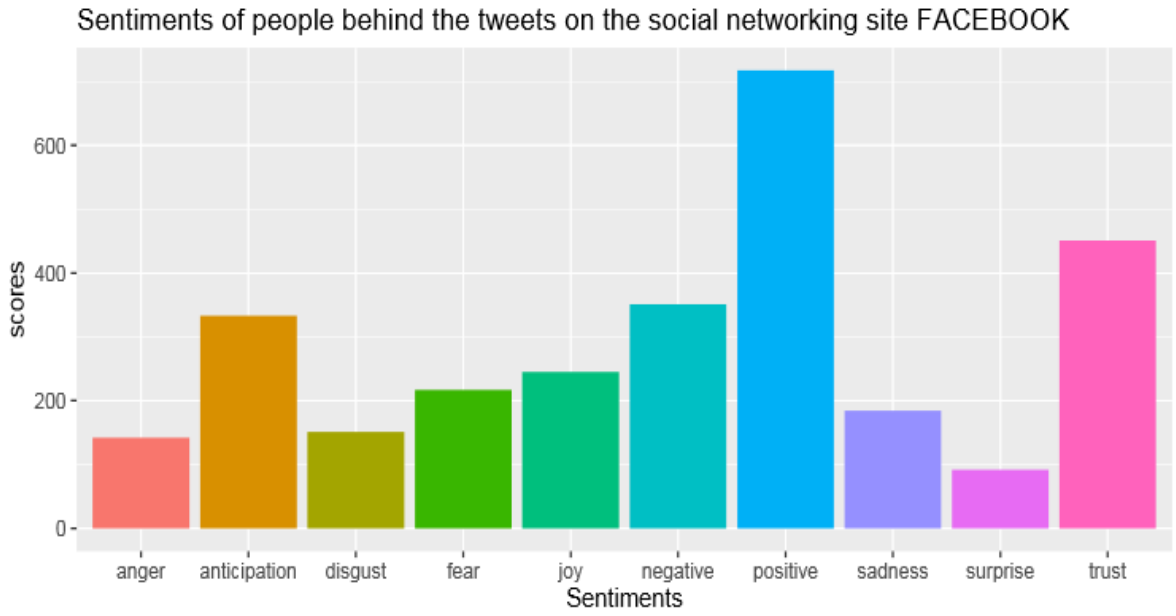
Analysis of the #facebook for the month of March

Figure-1



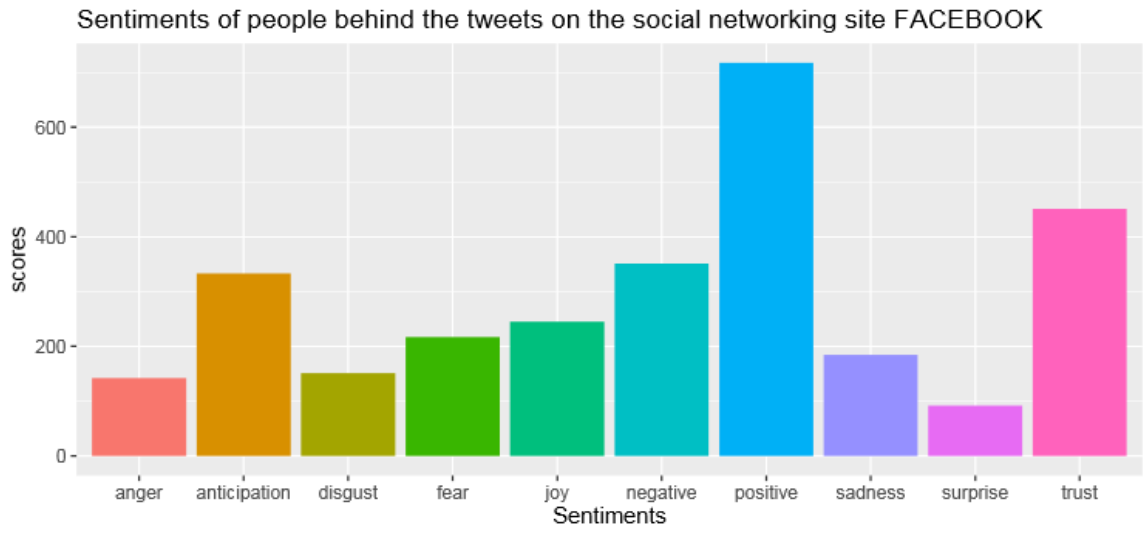
Analysis of the #facebook for the month of April

Figure-2



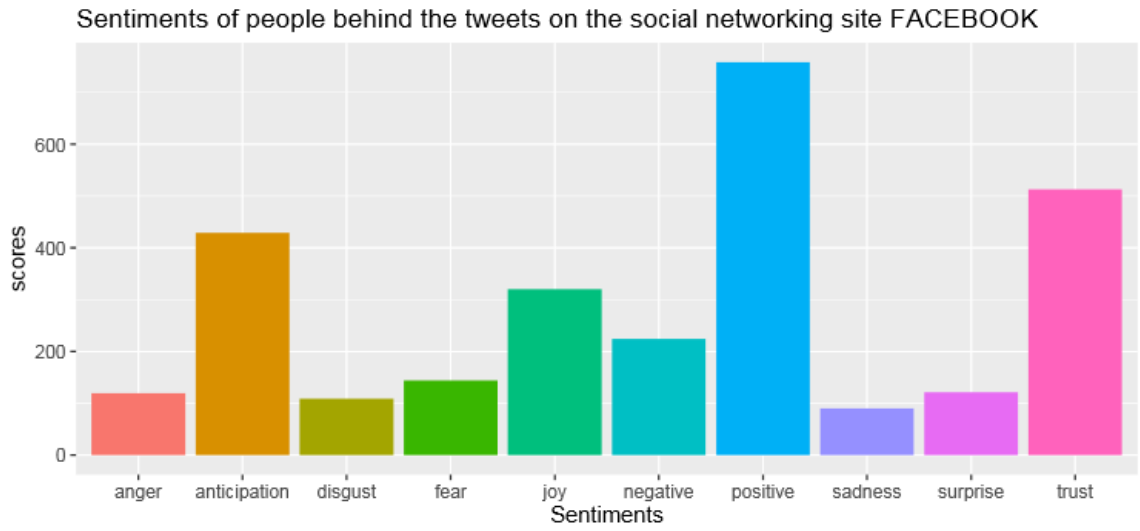
Analysis of the #facebook for the month of May

Figure-3



Analysis of the #facebook for the month of June

Figure-4



CHAPTER-7

CONCLUSION AND FUTURE SCOPE

Sentiment analysis of twitter data in the domain of microblogging is a relatively new research topic so there is a lot of room for research in this particular domain. Decent amount of work has already been done on sentiment analysis of movie reviews, document/web blogs, and articles and general phrase level sentiment analysis. This differs from twitter due to the word limit of 140 characters which forces the twitter user to express his opinions in a very compressed form of text. The best results reached in sentiment classification use supervised learning techniques such as Naive Bayes and Support Vector Machines, but the manual labelling required for the supervised approach is very expensive. Some work has been done on unsupervised and semi-supervised approaches, and there is a lot of room of improvement. Various researchers testing new features and classification techniques often just compare their results to base-line performance. There is a need of proper and formal comparisons between these results arrived through different features and classification techniques in order to select the best features and most efficient classification techniques for particular applications

The process of the sentiment analysis is still on the verge of advancement, especially in the domain of microblogging. We propose a couple of ideas which are worth exploring in the future and can help in the future improved performance.

In this we have worked with the simplest unigram models. Improvements can be brought in this models by adding some extra information like the closeness of the word with respect to the negation word. The nearer the word is to the negation word, the more impact it has on the overall polarity of the sentence. If the word is farther away from the negation word, it has a much less impact on the polarity of the sentence.

Apart from this, we are currently focussing on the unigrams. In a similar way, we need to lay our focus on the effect of the bigrams and trigrams as well. When unigrams are used with the bigrams and trigrams, it results in the improvement of its performance. However for the bigrams and trigrams to be an effective feature, we need to have a labelled data set of more than 9000 tweets.

In recent works of sentiment analysis we are currently working on the Parts of Speech separately from the unigram models. If we work on the unigram models along with the Parts of Speech, it will result in a more enhanced form of the analysis.

Another feature is worth of mentioning, is whether a relative position of a particular word in the tweet has any effect on the classifier. Although Pang et al. explored a similar feature and reported negative results, their results were based on reviews which are very different from tweets and they worked on an extremely simple model.

In this project we have mainly focussed on the general sentiment analysis. There is a lot of potential work available in this context with partial knowledge of the context. We have chosen

four hashtags which is currently trending today and we have extracted the twitter data relative to those context and performed a general analysis of the sentiments of the twitter users.

Appendix

The list of libraries used:

Name of the library	Usage
twitteR	Provides access to the twitter API in R.
ROAuth	Provides an interface to the **OAuth** 1.0 specification allowing users to authenticate via **OAuth** to the server of their choice
NLP	For the purpose of natural language processing.
syuzhet	An **R package** for the extraction of sentiment and sentiment-based plot arcs from text.
tm	Provides a comprehensive text mining framework for R.
SnowballC	Provides exactly the same API as Rstem, but uses a slightly different design of the C libstemmer library from the Snowball project. It also supports two more languages.
stringi	Provides R language wrappers to the International Components for Unicode (ICU) library and allows for: conversion of text encodings, string searching and collation in any locale, Unicode normalization of text, handling texts with mixed reading direction (e.g., left to right and right to left), and text boundary analysis (for tokenizing on different aggregation levels or to identify suitable line wrapping locations).
topicmodels	Provides an interface to the C code for Latent Dirichlet Allocation (LDA) models and Correlated Topics Models (CTM) by David M. Blei and co-authors and the C++ code for fitting LDA models using Gibbs sampling by Xuan-Hieu Phan and co-authors.
RColorBrewer	**RColorBrewer** is an **R package** that contains a ready-to-**use** color palettes for creating beautiful graphics.
wordcloud	Provides a visualisation similar to the famous wordle ones: it horizontally

	and vertically distributes features in a pleasing visualisation with the font size scaled by frequency.
ggplot2	**ggplot2** is a system for declaratively creating graphics, based on The Grammar of Graphics. You provide the data, tell **ggplot2** how to map variables to aesthetics, what graphical primitives to **use**, and it takes care of the details.

REFERENCE:

[1] Efthymios Kouloumpis and Johanna Moore,IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 3, July 2012

[2] S. Batra and D. Rao, ”Entity Based Sentiment Analysis on Twitter”, Stanford University,2010

[3] Saif M.Mohammad and Xiaodan zhu ,Sentiment Analysis on of social media texts:,2014

[4]Ekaterina kochmar,University of Cambridge,at the Cambridge coding Academy Data Science.2016

[5] Manju Venugopalan and Deepa Gupta ,Exploring Sentiment Analysis on Twitter Data, IEEE 2015

[6] Brett Duncan and Yanqing Zhang, Neural Networks for Sentiment Analysis on Twitter.2017

Additional links and websites:

1) https://hackernoon.com/text-processing-and-sentiment-analysis-of-twitter-data-22ff5e51e14c
2) https://lionbridge.ai/articles/the-essential-guide-to-sentiment-analysis/?ref=hackernoon.com
3) https://www.kaggle.com/rtatman/tutorial-sentiment-analysis-in-r
4) https://uvastatlab.github.io/2019/05/03/an-introduction-to-analyzing-twitter-data-with-r/
5) https://www.datacamp.com/community/tutorials/sentiment-analysis-R
6) https://github.com/Twitter-Sentiment-Analysis/R/blob/master/Final%20Report%20on%20Twitter%20Sentiment%20Analysis.pdf
7) https://stackoverflow.com/questions/37364908/how-to-extract-tweets-between-a-certain-time-period-in-r
8) http://www.sthda.com/english/wiki/text-mining-and-word-cloud-fundamentals-in-r-5-simple-steps-you-should-know
9) https://en.wikipedia.org/wiki/COVID-19_pandemic
10) https://theprint.in/opinion/bollywood-nepotism-karan-johar-sushant-singh-rajput/445850/
11) https://en.wikipedia.org/wiki/Facebook
12) https://www.english-video.net/v/en/689