Social Media Sentiment Analysis

Major project report submitted in partial fulfillment of the
requirement for the degree of Bachelor of Technology
in

Computer Science and Engineering

By
AYUSH BARI (181379)

UNDER THE SUPERVISION OF
Dr. HARI SINGH
Assistant Professor (SG)

Department of Computer Science & Engineering and Information Technology

Jaypee University of Information Technology, Waknaghat, 173234, Himachal Pradesh, INDIA

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Dr. Hari Singh, Assistant Professor (SG), CSE & IT Department**, Jaypee University of Information Technology. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:
Dr. Hari Singh
Assistant Professor (SG)
CSE & IT Department
Jaypee University of Information Technology

Submitted by:
Ayush Bari
(181379)
Final Year Student, B. Tech
Jaypee University of Information Technology

CERTIFICATE

This is to certify that the work which is being presented in the project report titled **“SOCIAL MEDIA SENTIMENT ANALYSIS** " in partial fulfillment of the requirements for the award of the degree of BTech in Computer Science And Engineering and submitted to the Department of Computer Science And Engineering, Jaypee University of Information Technology, Waknaghat is an authentic record of work carried out by “**Ayush Bari (181379)”** during the period from August 2021 to December 2021 under the supervision of **Dr. Hari Singh**, **Assistant Professor (SG), CSE & IT Department**, Jaypee University of Information Technology, Waknaghat.

Ayush Bari
(181379)

The above statement made is correct to the best of my knowledge.

Dr. Hari Singh
Associate Professor (SG)
CSE & IT Department
Jaypee University of Information Technology, Waknaghat, India
Dated: 09/12/2021

ACKNOWLEDGMENT

First, I express our heartiest thanks and gratefulness to Almighty God for His divine blessing makes it possible to complete the project successfully.

We are grateful and wish my profound indebtedness to **Supervisor Dr. Hari Singh**, **Associate Prof**, Department of CSE Jaypee University of Information Technology, Waknaghat. Deep Knowledge & keen interest of my supervisor in the field of "**Sentiment Analysis**" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

We would like to express my heartiest gratitude to **Dr. Hari Singh,** Department of CSE, for his kind help to finish my project.

We would also generously welcome each one of those individuals who have helped us straightforwardly or in a roundabout way in making this project a win. In this unique situation, I might want to thank the various staff individuals, both educating and non-instructing, which have developed their convenient help and facilitated my undertaking.

Finally, we must acknowledge with due respect the constant support and patients of my parents.

Ayush Bari
(181379)

ABSTRACT

Sentiment Analysis is a component of machine learning which processes natural language, analyzes text and then extract some emotions out of it. It is used to learn subjective Information and the state of a person.

Social Media has become a popular place for people to express their opinion about a brand, talk about it, and give feedback. It’s helps in understanding people's sentiment over any topic and incident. Analyzing sentiments help in understanding how people are thinking emotionally and classifying it as negative, positive or neutral.

Such data is available in big quantities which will be difficult to evaluate manually, examine and identify. So instead of doing this time-consuming exercise, we are going to use technical aspects to solve this problem. This dataset used here is a collection of many text and internet blogs. Many different machine learning classifiers are used here, so that a person's sentiment can be identified. All these classifiers are applied and then the best classifier with the best result will be chosen in order to predict people's emotions.

By this analysis the professionals can evaluate more of people's emotions accurately and it will help them identify early symptoms of distress.

TABLE OF CONTENT

Chapter-1 INTRODUCTION

1.1 Introduction

As the world is advancing and the internet era has begun, many youngsters have developed a habit of starting their new day with a good morning text. The intention of these technologies was to make human life easier and friendly and they are useful in every aspect of our life. Today the machines have become so advanced that these machines can even predict the future with the help of artificial intelligence based on current data.

As social media websites keep evolving and slowly become the source of all kinds of information, people started posting their opinion on various topics, discussions, issues, complaints and expressing negative, positive, or neutral emotions in response to the product they use or the condition they go through. Many brands and companies even conduct polls on these sites and blogs to understand the general people sentiment and demand of their various offerings. This is requirement for some technology that can identify and summarize overall people's sentiment.

The advancement of various machine learning and artificial intelligence algorithms have changed how we understood the world earlier. In the process of this people started neglecting connections with the other as they did before. In this fast-paced world, everyone wants to contribute to the advancement of society and they try to become the best in this world. This hunger of success between people is one of the many reason that build stress and pressure on people ultimately leading towards the path of tension, depression, anxiety and sometimes even suicide. The sad thing is those people suffering are themselves less aware of this [1]. The mental illness, depression and suicide are slowly becoming a global phenomenon. People have become less aware of other’s emotions as they become isolated in their social media bubble. In the era of COVID-19 such people have increased even more. The recent studies show there is an increase of 67% in suicidal behaviors in lockdowns era [2]. The sharp increment of cases at this rate brings the need to correctly identify a person’s emotions so that right assistance can be provided at the right time. In older times, we needed to visit someone’s social media manually and go through their activities which was a very time consuming and ineffective method. These artificial intelligence and machine learning models can be used to process data at very large scale. This will make the process more effective with less time consuming. Sentiment analysis is gaining popularity for product evaluation, customer feedback, and understanding public demand.

1.2 Problem Statement and Objective

Problem Statement -

Social media is the hub of public opinion where millions of opinions are shared every few minutes. As more and more people are expressing their thoughts on social media which can be both neutral and polarizing. The problem is to plot public sentiments on the relative topic and monitor it.

Objective -

Objective of the project is to extract tweets from twitter and extracts sentiments out of them. We will be plotting a graph based on those sentiments to check their sentiments and plot a graph based on subjectivity and polarity of their tweets.

CHAPTER 2 - LITERATURE SURVEY

Literature review is an important aspect of this project since it helps in establishing familiarity with the topic. With the help of various literature review we can understand the current research in the respective field. It make things more clear and helps in greater focus to the research problem and understand the findings.

Machine Learning

Machine learning is a study of computer-based algorithms which is used to improve software functioning through experience and using data in abundance. Machine learning is a component of Artificial Intelligence. To build machine learning models we use machine learning algorithms on sample data or training data. The more the software will learn about available data, the more accurate results it will predict.

Machine learning algorithms are used in all kinds of variety applications like recommender systems which recommend people the next thing they should interact on their devices or helping people by filtering out redundant email often called spam. Subjects like speech recognition and computer vision have become very convenient and accessible by use of machine learning. Many of today's big tech companies like Facebook, Google and Apple are using machine learning as a central part of their functioning.

There are four basic approaches in machine learning

- Supervised Learning – Here, we apply algorithms on labeled training data and software learns a function by example of input output pairs. In this, every pair consists of an input object and a desired value. It helps organizations solve plenty of real-world problems at scale. Some models used in supervised learning are naïve bayes, support vector machine (SVM), linear regression, logistic regression, random forest and many more.

- Unsupervised Learning – Here, we use Machine learning algorithms to identify patterns in data sets having data points which are neither classified nor labeled. This is used where we want to discover hidden patterns, similarities and differences in information available to us. Principal component analysis, k-means clustering, neural networks are commonly used approaches for this type of learning.

- Reinforcement Learning – It is used to teach a machine to complete a multi – step process for when the rules are clearly defined. This algorithm is capable of deciding on its own what steps should be taken along the way.

Sentiment Analysis

Sentiment analysis is the process of judging whether the opinion expressed is positive, negative or neutral. A sentiment analysis system for text analysis uses natural language processing (NLP) with machine learning techniques to assign scores to each entity's topics, categories, themes for every sentence and phrase shared by the user.

It helps organizations and data scientists gauge, express opinion, and conduct nuanced market research. It helps in monitoring complex things like bank reputation, likes, dislikes about some certain product and identifying people's emotions. People going through mental illness, depression, suicidal thoughts can be approached early on which can help in reducing already very high suicide rates.

There are many types of sentiment analysis which are useful for different purposes like

- <u>Fine – Grained Sentiment Analysis</u>

  It is used when an organization wants to understand what kind of feedback they are getting for their products. It helps in judging the customer reaction towards the products of related accessories and helps brands customize its portfolio. The fine-grained analysis has the ability to provide more precise results to a system that prioritizes addressing customer complaints.

  The data usually ranges from very positive to very negative. It uses words like anger, sadness, worries that are associated with negative sentiment, and for positive sentiment it has words like happiness, love, enthusiasm.

- <u>Emotion Detection Sentiment Analysis</u>

  It is an interesting way to extract the emotions out of a piece of text. Predefined lexicon which are the words that are either positive and negative are used to determine the sentiment. It helps the company understand the needs of customers.

  On the downside front, emotion detection sentiment analysis system encounters confusion when it has to process varied lexicons from the customer. It usually happens when a word is used for both positive and negative perception of the product or service.

  Example – Virat Kohli is killing it is a positive emotion comment but it can create confusion to the algorithm.

- Aspect based

  Aspect based sentiment analysis is usually used for a particular aspect of service or product. For example, a laptop vendor might use it for one aspect like processing speeds or graphics performance. They can understand customer feelings about an attribute by using this. By using specific details, the composed opinion can be derived to provide insightful feedback to the company. This makes it very easy to detect customer complaints and resolve them.

- Intent Analysis

  Many times, it is required to focus on the right customer. Focusing on the right customer can help save resources for a company, so that they can utilize their limited resources better. It helps companies understand whether the customer actually intends to use that service or not. It is highly important for marketing and advertising purposes.

Sentiment Library – These are basically very large collections of words like wonderful, horrible, bad, dirty and phrases like amazing show, horrible story, good, performance, wonderful game. These things are pre scored by humans and given a sentiment weight. A basic sentiment analysis algorithm uses such a library to understand sentiments of the phrases it encounters.

This type of manual sentiment scoring can be a tricky process as it involves decision on how positive or negative a word/phrase is going to get rated. These libraries require consistent maintenance by tweaking scores based on trend and adding new phrases.

This can be used to execute sentiment analysis without training machine learning models which are known and rules-based sentiment analysis.

Sentiment analysis is getting used everywhere

Sentiment Analysis can be performed on both Static Data and Streaming Data. Static data refers to a fixed data set. This type of data set is collected beforehand and is processed later to get required results. This is quite different from Streaming data. Streaming Data is data that is continuously generated in real time. This type of data needs to be processed constantly using suitable stream techniques. While static data is predefined and is not changing, streaming data continually changes after its recorded and hence provides better accuracy results.

Related Works –

This topic sentiment analysis and emotion analysis has gone through a lot of research in recent times even if most of the research is done to classify positive and negative emotions on textual data.

Zhao Jianqiangi and Gui Xiaolini (2016) determined various Text pre-processing methods affect the performance of sentiment classification. In their research they compared six different pre-processing methods that show different sentiment polarity classification results in Twitter. Experimental results indicated that the performance of classifiers is not affected by the removal of URLs, the removal of stop words and numbers although they are helpful in reducing noise. But on the other hand, replacing negations and expanding acronyms. The study concludes that selection of appropriate pre-processing methods for different classifiers produce different results.

Mondher Bouazizi and Tomoaki Otsuki (2016) in order to extract patterns that determined the level of sarcasm of tweets used Parts of Speech tags. Upon performing different natural language processing (NLP) tasks with Apache OpenNLP tool they got good results by applying a selected approach. The result could be better when the training set is bigger. To cover larger data set of sarcastic tweets Bridianne O'Dea, Philip J. Batterham, Mark E. Larsen, Cecile Paris, Tjeerd W. Boonstra, and Helen Christensen (2015) presented the “We Feel system” to get a real-time emotional sentiment analysis tool for Twitter. In their collective efforts, they were able to analyze 2.73×109 tweets over a 12-week period. They detected significant events while analyzing the weekly variations in emotional expression. They also observed some indices of anxiety, depression and suicidal thoughts, which indicates some social media tool that can measure mental health could also be made.

Ping Feng Pai and chia Hsin Liu (2018) [5] developed a framework consisting of time series forecasting models and multivariate regression technique to predict monthly total vehicle sales.

The researchers got more accurate forecasting results after seasonalizing procedures with hybrid multivariate data.

El Aloui [4] performed sentiment analysis on the data obtained by 2016 tweets of ongoing US elections. They presumed that every class has its own version of us elections. S. Kaur [6] had used sentiment analysis with the N-gram algorithm which he used to feature KNN and extraction for classification. That model was able to achieve an accuracy of over 86%.

M. Wankhede [8] predicted food reviews by applying multiple machine learning models. In all those models, he achieved the highest accurate result using logistic regression. One failure of the model is it didn’t perform equally well for predicting negative reviews.

N. Zainuddin [7] selected chi-square for her feature extraction. She used various term-weighting schemes and SVM for classification. On the Taboada corpus dataset she achieved 73.21% maximum accuracy.

CHAPTER 3 - SYSTEM DEVELOPMENT

3.1 Background of the project

Natural Language Processing helps developing machines that understand text data or voice data and respond on their own using text or voice data. It is pretty much similar to how humans interact.

Natural Language Processing is a branch of Artificial Intelligence AI developed to give computers the ability to understand text and words in similar ways to humans.

NLP helps in computer software to take text from one language and convert it to another. It makes them respond to various commands humans spoke with the help of statistics, machine learning, and deep learning neural models. In and all these technologies enable software to understand human text to higher meaning and work with human sentiment and intent.

We interact with NLP in voice assistants, GPS systems, they help us summarize large volumes of text. All the text to dictation programs works with NLP, customer service chatbots and many other consumer convenience services. NLP is very useful to enterprises as it helps in streamlining operations, simplifies critical mission business processes and increases productivity in office and for employees.

Human language is filled with so many nuances and ambiguity that can make it difficult to manage software and determine what customers want from respective text or voice data. Human speech contains many things like sarcasm, idioms, metaphors, grammar, homophones and many more like variations in human speech. It's also full of irregularities.

As indicated by the business assesses as it were 21% of the accessible information is present in the organized structure information has been created as we talk as we treat as we communicate something specific on WhatsApp, Instagram hello informing and different stages larger part of this information exists in the text based structure which is profoundly unstructured in nature now to produce critical and noteworthy bits of knowledge from this text information it is essential to get to know the strategies and the rule of regular language handling so we should comprehend what precisely is NLP now regular language handling that is NLP alludes to the man-made reasoning technique speaking with a keen framework utilizing normal language now it is a section of software engineering and counterfeit knowledge which manages the human language by using NLP and parts one can coordinate the enormous pieces of text information play out various robotized assignments and take care of a wide scope of issues for example, mechanization outline machine interpretation named substance acknowledgment relationship extraction wistful investigation discourse acknowledgment and subject divisions

Many fields where NLP is extensively used are as follows

1.Speech Recognition – It is also called speech to text and is a reliable method to convert voice data to text data. Applications that require speech voice command conversion require speech recognition. Its major challenges are to understand what people talk about, the shortcuts and slangs and often incorrect grammar. Understanding different accents is also a problem. Watchword search and it is additionally a field where NLP is intensely utilized presently extricating data from sites or on the other hand a specific archive additionally requires the information on NLP. Now one of the coolest applications of NLP is the bolt bother it matching which is fundamentally a proposal of promotions dependent on your search what it does is examines the text of the information which we are now utilizing or then again looked and coordinate it with the text

2. Part of Speech tagging - It is the problem of understanding the part of speech of any specific word on some human written piece of text based on where and in what context use and context.

3. Named entity recognition – Many phrases are identical and many phrases are also identical which are used as useful entities. NEM helps identify Which word is the location and which is the name etc.

4. Coreference resolution - Coreference resolution helps in when there is a need to identify if they are both or more same words to the same word and same entity in a given piece of text. One of the most common practices is to identify the person or object to which a certain noun or pronoun refers but it can also be involved in identifying an idiom or idiom or a sarcasm in the text.

5. Sentiment Analysis – It helps organizations and data scientists gauge, express opinion, and conduct nuanced market research. It helps in monitoring complex things like bank reputation, likes, dislikes about some certain product and identifying people’s emotions.

6. Natural language generation – Natural Language generation is basically defined as the opposite of speech recognition and also the opposite of speech to text. We use this to complete the task of putting structured and proper information into human language.

Python programming Language –

Here in this project, I am using python programming language which is extensively used for machine learning and artificial Intelligence fields. Python comes with all kinds of libraries like pandas, NumPy, SciPy and hence working with them becomes quite easy. Python is also used in those fields where big data and statistics play a dominant role and need to be worked with. So here, python will be used as the main programming language.

Python is a general-purpose high-level language and has been a very popular language for a long time. It is getting heavily used in domains of web development, machine learning applications as it brings all the easy-to-use advanced features to the programmers. Python also allows object oriented and procedural paradigms and code can be written that way. In python people have to generally type less indentation which makes it more readable to everyone. It is now used all across the IT sector and big tech companies, in fact all the big tech companies like google, Microsoft, apple, all are using python to develop their cutting-edge technology and bring advanced customization to their customers. It also supports advanced word processing (late binding), which binds the way the words change during the process. Python has reference computation inbuilt integrated and owing to the less redundant nature of the language, programmers have to write less lines of code as compared to other high-level languages like C, C++ and java.

Python applications contain less lines than programs written in other languages. That is why we choose Python for artificial intelligence, artificial consciousness, and dealing with massive volumes of data. Python is an article-oriented programming language. Classes, objects, polymorphism, exemplification, legacy, and reflection are all concepts in Python.

Many python libraries are used in this project like –

- NumPy
- pandas
- Matplotlib
- seaborn
- io
- NLTK
- Scikit-learn

A little explanation on the libraries used

- NumPy – NumPy is a python library used for working with arrays. It is an open-source project where NumPy stands for numerical python. Originally python has a list to work as arrays but they are quite slow in actual programming which creates a problem and this problem is solved by NumPy as it stores arrays in a continuous place.

- Pandas – Pandas is a python library which makes it easier to operate with data. It is powerful, faster in speeds and quite flexible. Pandas is also an open-source data manipulation and analysis tool.

- Matplotlib – A lot of plotting is required in machine learning projects and this is where matplotlib helps. It is a plotting library developed to work with python programming language. It provides an object-oriented API and it is used for generating plots and graphs.

- Seaborn - It is also based on matplotlib. It is a high-level interface to draw attractive and statistically informative graphics. It supports high level abstraction for making multi-plot grids.

- io – This allows us to operate on file input and file output operations. It is advantageous to use because the function and classes here allow us to extend the functionality and are able to write Unicode data.

- Scikit-learn – scikit-learn is a robust and useful library used in machine learning on the python platform. It contains efficient tools and various statistical models like regression, clustering which are used in machine learning.

- Textblob – It helps in accomplishing many tasks of NLP through its API which includes sentiment analysis, classification, translation, part of speech tagging and more.

Natural Language Toolkit (NLTK)

Natural Language Toolkit is used to implement NLP in python programming language. It contains any libraries for task in NLP and additional libraries for many different small tasks including segmentation of words, parsing of sentences, lemmatization, stemming which is a method to clean words down to their roots and tokenization that is used so that we can transform sentences, phrases, paragraphs and lines into small tokens that helps software to better execute machine learning algorithms and produce better results. The NLTK library is the regular language tool stash so NLTK is mainly used for building Python projects that can work with human language information as its males simple to interact with to 50 corpora and lexical assets. This also helps in drawing conclusions from the human written text.

Pickle

Pickle is a python module which we use for serializing and deserializing any object in python structure. We can pickle any object in python to save it on disk. The main things pickle achieves is that it serializes the given object before it writes the object in a file. So it get used as a tool to convert any object python has like list, dict, etc. into a character stream. The resultant character stream actually contains all the necessary information needed to reconstruct it back into another python script.

Twitter API

Twitter API is helpful as it provides developers access to Twitter and its most of the given functionality. We can use this API and read and write various information present in Twitter website like tweets, trends, users.

Twitter API provides access to lot of HTTP endpoints like

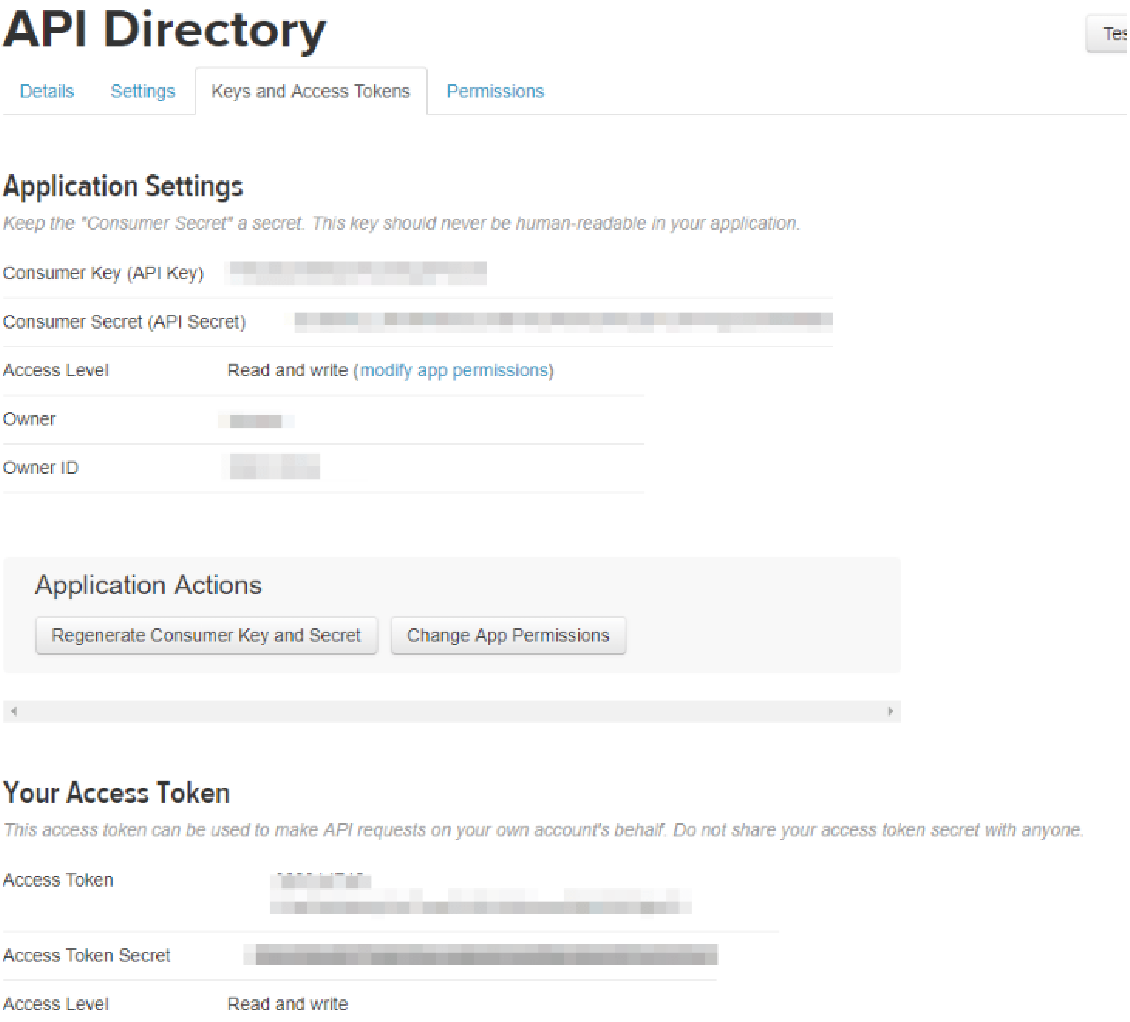
- Tweets
- Retweets
- Trends
- Media
- Likes

To access Twitter API, we first need to authorize using varies keys and token given to us.

Tweepy

It is an open source Python package that provides a easy and reliable way to access Twitter API through Python. Tweepy consists of range of different classes and methods which helps in accessing Twitter's models and API endpoints. It takes care of various things like

- Streams
- OAuth authentication
- Rate Limits
- HTTP request



API Directory

Application Settings

Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.

Consumer Key (API Key)	[redacted]
Consumer Secret (API Secret)	[redacted]
Access Level	Read and write (modify app permissions)
Owner	[redacted]
Owner ID	[redacted]

Application Actions

Your Access Token

This access token can be used to make API requests on your own account's behalf. Do not share your access token secret with anyone.

Access Token	[redacted]
Access Token Secret	[redacted]
Access Level	Read and write

Image – Authentication details of Twitter API

Machine Learning Models –

There are various Machine Learning models to choose from

- Logistic Regression
- Bernoulli Naïve Bayes
- Random Forest Regression
- Multinomial Naïve Bayes classifier

To determine the emotion behind given textual data.

Logistic Regression

Logistic regression is a machine learning classification algorithm that is used to observe a discrete set of classes. For example, spam filters of emails, checking and validating online transactions, medical uses etc. Logistic Regression uses its logistic sigmoid to form its output so that it can return a probability value.

It is a predictive analysis algorithm. Logistic regression is based on the concept of probability.

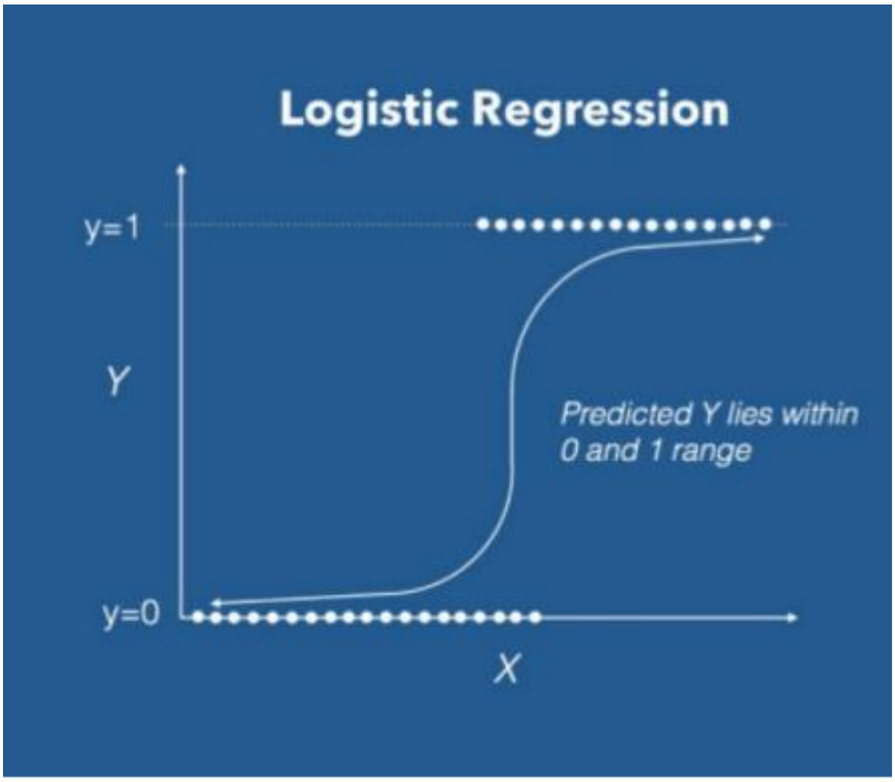


Fig - Logistic Regression range

In this figure, the range of Logistic Regression is shown which always lies between 0 to 1.

Logistic Regression is a great regression analysis technique that explains the relationship between a dependent nominal variable, binary variable, interval variable and ratio-level variable. It is a supervised algorithm that uses a logical function that covers log odds to probability.

$$log[p/(1-p)]$$

This is the function used when we have to convert log odds to respective probability.

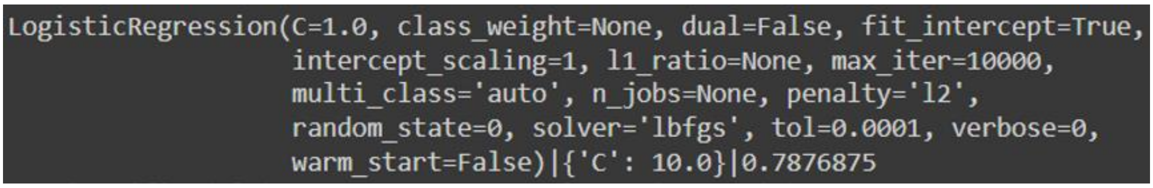
Here, p is the probability of an event taking place, [1-p] is the probability of an event not taking place. Analogous models that don’t have logistic function but use sigmoid function can also be used the curve and sigmoid function is represented by

$$Accuracy = \frac{1}{1 + exponent^{-value}}$$

This is a sigmoid function. In order to predict values to probabilities, sigmoid functions get used. This function maps values between 0 to 1.

<u>Benefits of Logistic Regression -</u>

- Logistic regression works well for linearly separable data sets.
- Logistic regression discusses the coefficient size indicator in both positive and negative scope.
- Logistic Regression is less prone to overfitting
- Logistic regression is very easy to train, easier to interpret and implement.



```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, l1_ratio=None, max_iter=10000,
                   multi_class='auto', n_jobs=None, penalty='l2',
                   random_state=0, solver='lbfgs', tol=0.0001, verbose=0,
                   warm_start=False)|{'C': 10.0}|0.7876875
```

Fig - Applying logistic regression

In this figure, logistic regression is shown, where we set the inverse of regularization at 1.0 while max iteration was set up to 10000. All the rest of the parameters are set to be default.

Naïve Bayes Classifier-

Naïve Bayes classification is a machine learning model which runs on the method of probability. Naïve bayes uses probability for completing classification task
This classifier is based on Bayes Theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes Theorem

Here P(A|B) = Probability of A when B is true
P(B|A) = Probability of B when A is true

P(A) = P(B) = Independent probability

Bayes theorem is a probability theorem using which we can find the probability of A happening, if the B has already occurred. Here, Since B has already happened, B is evidence while A is just a hypothesis we are going to extract. Bayes theorem works on the assumption that predictors are independent which means presence of one feature does not impact other, that's why it is called as naïve

This probabilistic approach makes the naïve bayes classifier the fastest among other classifiers.

Advantage of using Naïve Bayes Classification –

- It is simple and easy to implement
- It works even without much training data.
- It can handle both discrete and continuous data
- Its fast enough to be used for real life calculations
- It does not get affected by irrelevant features
- It is a highly scalable classifier

Application of Naïve Bayes algorithm –
It has successful applications like

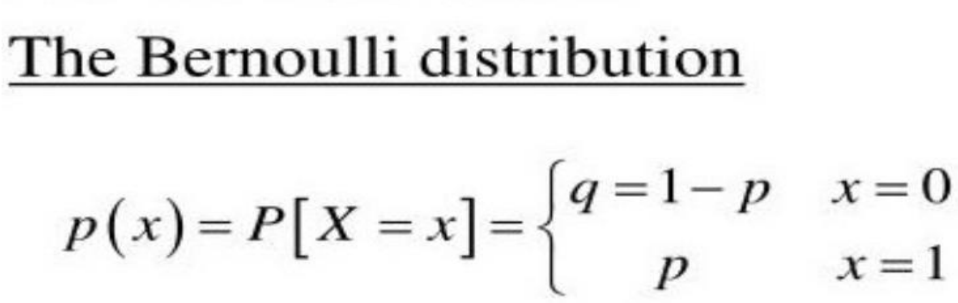
- Spam filtering
- Test Classification
- Sentiment Analysis
- Recommender System

i) Bernoulli Naïve Bayes –

Bernoulli Naïve Bayes is a variation of Naïve Bayes. Its works on discrete data according to Bernoulli distribution. The main thing about Bernoulli Naïve is that it only works in binary values, it only accepts values containing true or false or yes or no or success or failure, 0 or1 and so on. So, this classifier mainly works when the main feature values are binary.

Bernoulli distribution –

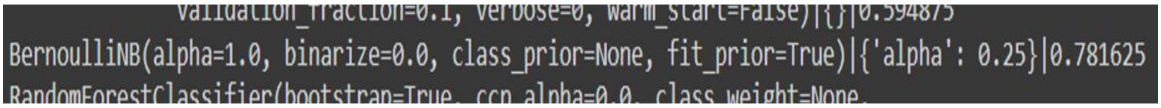
The random Bernoulli distribution is like this



The Bernoulli distribution

$$p(x)=P[X=x]=\begin{cases} q=1-p & x=0 \\ p & x=1 \end{cases}$$

Here ‘p’ is the probability of success and
‘q’ is probability of failure, q = 1-p



```
              validation_fraction=0.1, verbose=0, warm_start=False)|{}|0.594875
BernoulliNB(alpha=1.0, binarize=0.0, class_prior=None, fit_prior=True)|{'alpha': 0.25}|0.781625
RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
```

Fig - Applying Bernoulli Naïve Bayes

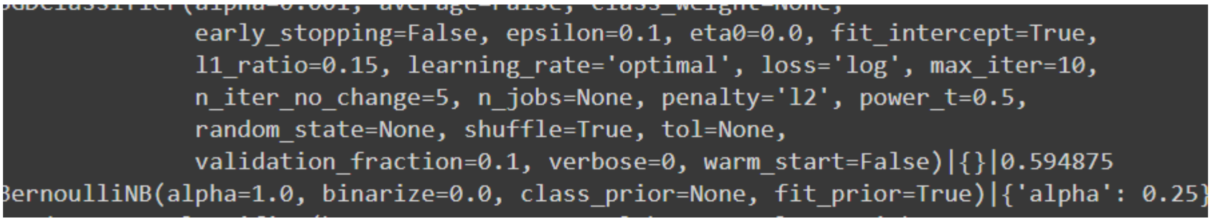
In this figure, Bernoulli Naïve Bayes is shown, where the highest accuracy came when we set additive parameters to 0.25 while setting all other parameters as default.

ii) Multinomial Naïve Bayes

Multinomial Naïve bayes is another machine learning algorithm which is the variation of Naïve Bayes. Multinomial Naïve Bayes is mostly used in Natural Language Processing (NLP). This algorithm is derived from Naïve Bayes and it uses a probabilistic approach to machine learning classification. Multinomial Naïve Bayes classifier comes with features containing the frequency of specific word in document, it is calculated probability of every tag given in a respected sample and then whichever tag has the highest probability is the answer.

Since it is based on Naïve Bayes, the presence or absence of some feature doesn’t not impact presence or absence of some other feature. Multinomial Naïve bayes normally require integer feature counts but it can also work with fractional counts for example TF-IDF.

.



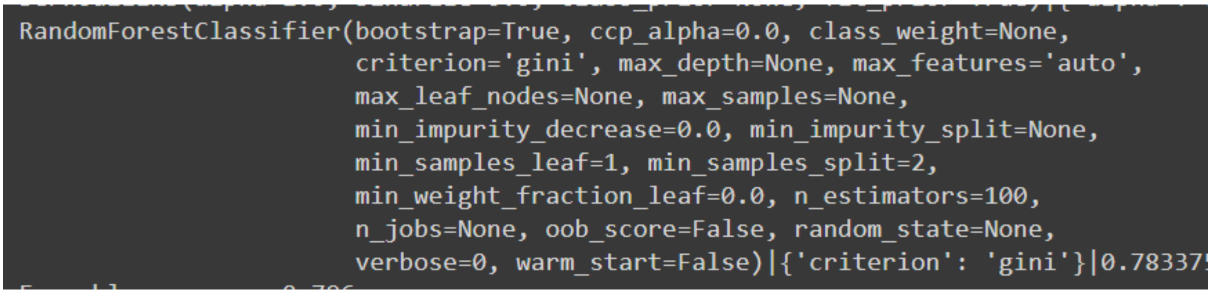
```
SGDClassifier(alpha=0.001, average=False, class_weight=None,
              early_stopping=False, epsilon=0.1, eta0=0.0, fit_intercept=True,
              l1_ratio=0.15, learning_rate='optimal', loss='log', max_iter=10,
              n_iter_no_change=5, n_jobs=None, penalty='l2', power_t=0.5,
              random_state=None, shuffle=True, tol=None,
              validation_fraction=0.1, verbose=0, warm_start=False)|{}|0.594875
BernoulliNB(alpha=1.0, binarize=0.0, class_prior=None, fit_prior=True)|{'alpha': 0.25}
```

Fig - Multinomial Naïve Bayes Classification

In this figure, Multinomial Naïve Bayes is shown, where we got the best result when additive smoothing parameter was 0.0 and all other parameters were set to default.

Random Forest Classifier

Random forest classifiers consist of randomly selected decision trees, these trees are processes individually that can operate as an ensemble. Random forest model is made of multiple decision trees. Decision trees seek to find the best split to split the data. These are mainly trained through classification and regression trees. In this classification model, each separate tree provides a separate class prediction and after all whichever class gets the most votes become model prediction. There was a little correlation between predictions made by these individual trees to each other. It uses the majority vote for class and the common for regression to generate selection timber from diverse samples. Since it incorporates various characteristic choices in its constructing techniques, they are able to accommodate a large number of functions at excessive prediction accuracy.



```
RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
                       criterion='gini', max_depth=None, max_features='auto',
                       max_leaf_nodes=None, max_samples=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, n_estimators=100,
                       n_jobs=None, oob_score=False, random_state=None,
                       verbose=0, warm_start=False)|{'criterion': 'gini'}|0.78337
```

Fig - Random Forest classification

In this figure, Random Forest classifier is shown, where we set the measure of quality of the split to ‘Gini’ and all other parameters to default.

Benefits of using Random Forest Classification –

- Reduces risk of overfitting
- Provides flexibility
- Easy to determine feature importance
- Effective tool for estimating missing values

3.2 Algorithm –

This contains the description of all the methods involved in this project, these are the methods we are using to extract appropriate emotions from the people's opinions we used.

The detailed description of all steps involved in given below

1. **Fetching data** – Firstly, we will authenticate our Twitter API credentials to fetch tweets. After authentication will be provide the person whose tweets we are going to analyze and fetch their tweets. We will store those tweets after fetching for further usage.

2. **Data Pre-processing Phase** – This is the process performed over raw data according to algorithm needs. It involves cleaning of data, integration of data reduction of data and finally transformation of data. In my project, initially the data is a bunch of extracts taken from twitter API, these will be stored in another text file. But before storing them we will clean it and remove all the unnecessary items from it. We are using stop-words here that are present in the Natural Language toolkit so that we can remove all the common words like the, me, you, a, an etc.

   Stop-words are the commonly used words for example as, the, a, an, in, etc. We do not want to waste time over these common redundant words which contribute nothing to our sentiment analysis and hence we remove them through the NLTK library. This helps to eliminate redundant clutter and focus on drawing out correct emotions. Here, we are using lemmatization which can reduce words to their root status. Lemmatization uses a vocabulary and processes the data in order to remove any redundant suffix or prefix and reduce the word to its core, that is lemma. For example, it will reduce words like ran, dun, running to run which is its lemma. Another alternative to this is stemming which reduces the words to stems but lemmatization is better as it stays closer to the real meaning inflicted in text.

   We will clean tweets and remove many component of tweets and doesn't assist in our sentiment analysis.
   These things include
   - Emoticons
   - Retweets
   - Hyperlinks
   - Advertisement
   - Mentions
   - Tags

3. **Term-Weighting Scheme** – The Term weighting schemes are very useful in the essence that they are useful to collect essential information from the unstructured data. It processes all the words and assigns them a certain numerical value. This value is used to calculate the weight of that particular word and distinguish it from the others. This method predicts the importance of different words by calculating the predictable emotions embedded within them. It calculates the frequency of the word used and then uses that to calculate its relative importance. Here TF_IDF has been used.

TF_IDF –

TF (Term frequency) IDF (inverse document frequency) is used for scoring the relative importance of words. It is a statistical method to determine relevance and importance of any given word in the given text and what relevance is the information that it provides about the context. The number is inversely proportional if the same word is common in other documents as well although directly proportional to the number of times a word is present in a phrase, paragraph or document.

This has been shown with the help of below equations

$$idf(t,d,D) = tf(t,d).idf(t,D) \qquad (1)$$

$$(t,d) = log(1 + freq(t,d)) \qquad (2)$$

$$idf(t,d) = log(\frac{N}{count(d\epsilon D : t\epsilon d)}) \qquad (3)$$

In the above written equations, equation 1 denotes the calculations in formal mathematical terms. Term frequency simply means how many times a word has appeared in a given document. This can be calculated simply by the raw count. Then we do its inverse. Inverse Document Frequency shows how less or how rare a word is in the entire document set. This is calculated as the logarithm of total documents divided by total number of documents where the word exists.

3.3 System Requirements

System Hardware requirements for the project:

- (Intel / AMD architecture) x86 64-bit CPU
- Four GB RAM
- Five GB loose disk space

System Software Requirements

- Operating system: Windows 10
- Jupiter with Python3
- Spyder
- Scikit-learn library
- NLTK library

CHAPTER 04: PROJECT'S IMPLEMENTATION

Real analysis will be done on twitter. We will take live fresh tweets out of twitter in real time and will apply our model, naïve bayes in our case to differentiate between positive and negative tweets This live extraction part is the main component of our data set. People express different opinions and based on their writing style their emotions can be extracted. For example, when a person suffers, he/she goes through a lot of emotional distress, they often don't interact with others, their eating habits and their sleeping habits and many more. One particular change is how they are expressing their opinions.

An emotionally distressed person often uses word like frustrated, sad, etc. frequently while a person with suicidal thinking uses kill, death, hopelessness types words in their writing. Their writing styles consist more of extremist words with heavy usage specific words like full, must, absolute, never, etc. They often look at words as black and white and nothing in between which is visible in their writing.

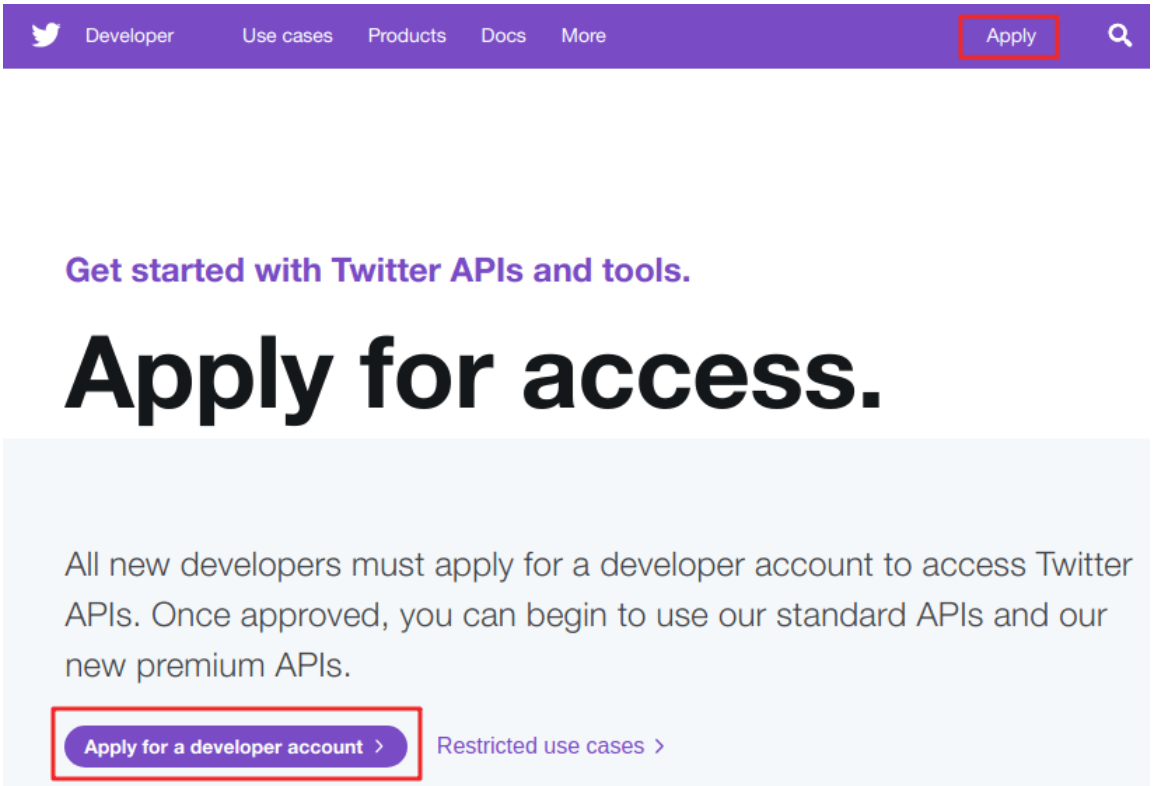
For the dataset, I have used Twitter site which contains opinions on all types of topics ad events and from there we will be live extracting them through our code.

4.1. Authenticate Twitter API

First we have to authenticate our access to Twitter API through tweepy.OAuthhandler

- consumer_key
- consumer_secret
- access_token
- access_secret

These keys are available once we make our twitter developer account. These keys are private to each user.



Get started with Twitter APIs and tools.

Apply for access.

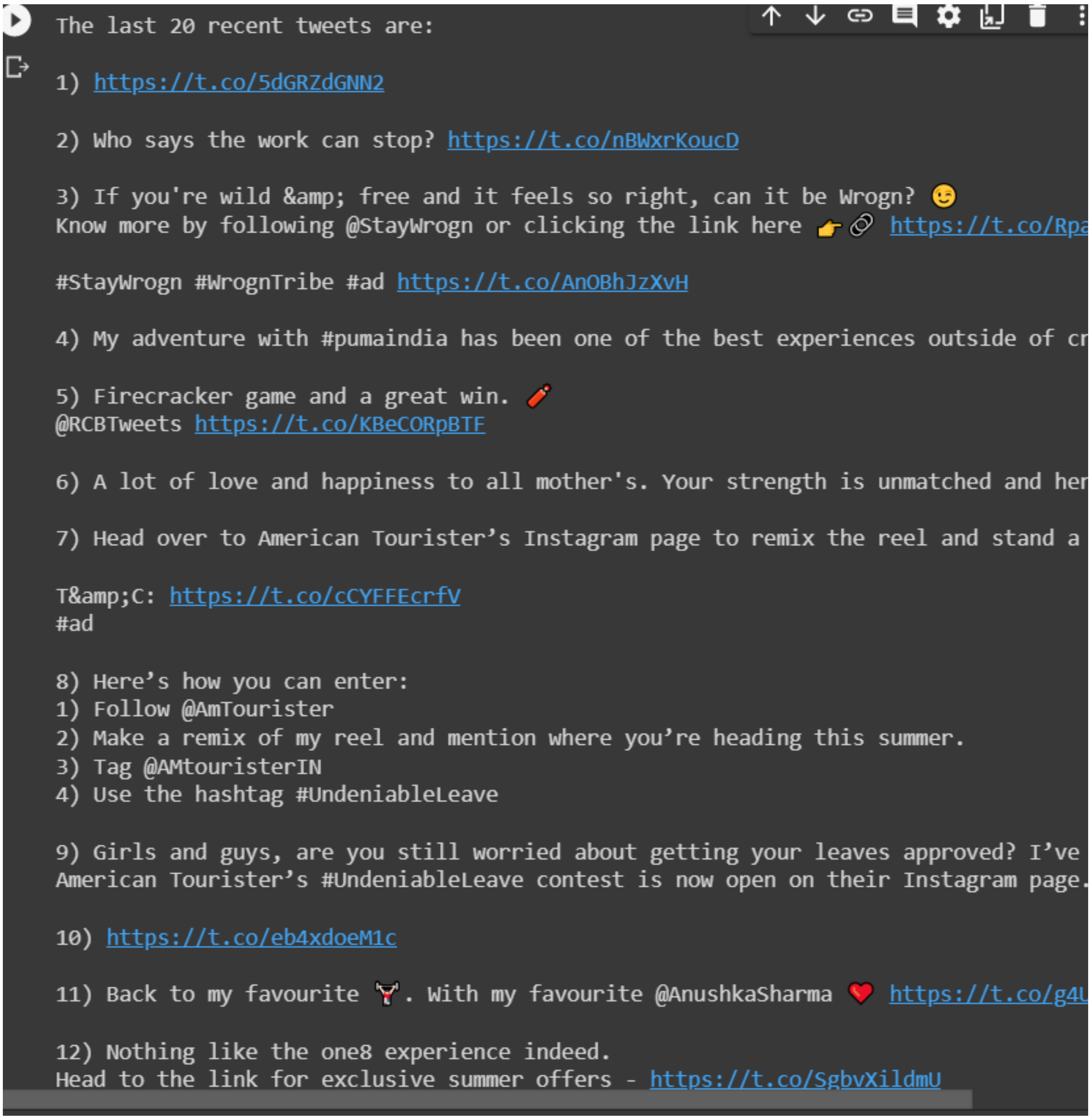
All new developers must apply for a developer account to access Twitter APIs. Once approved, you can begin to use our standard APIs and our new premium APIs.

Image – Signup page for Twitter Developer Account

4.2. Extracting Tweets

Then, we will enter handle id of twitter user we want to extract the tweets. Here, I am extracting 100 tweets from twitter users.

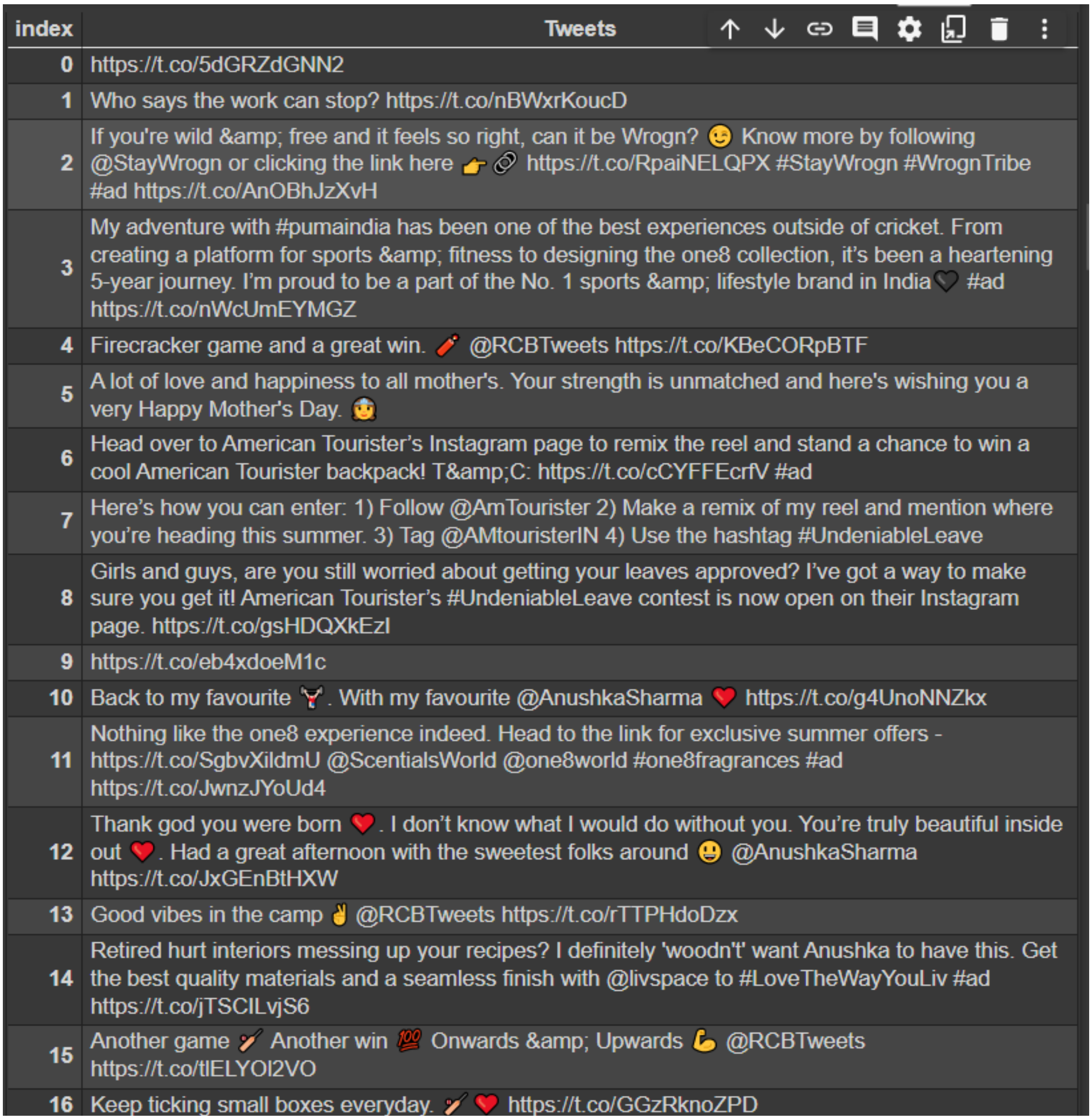
For example these are the last 5 tweets of Virat Kohli (tweeter handle : @imVkohli)



```
The last 20 recent tweets are:

1) https://t.co/5dGRZdGNN2

2) Who says the work can stop? https://t.co/nBWxrKoucD

3) If you're wild &amp; free and it feels so right, can it be Wrogn? 😉
Know more by following @StayWrogn or clicking the link here 👉🔗 https://t.co/Rpa

#StayWrogn #WrognTribe #ad https://t.co/AnOBhJzXvH

4) My adventure with #pumaindia has been one of the best experiences outside of cr

5) Firecracker game and a great win. 🧨
@RCBTweets https://t.co/KBeCORpBTF

6) A lot of love and happiness to all mother's. Your strength is unmatched and her

7) Head over to American Tourister’s Instagram page to remix the reel and stand a

T&amp;C: https://t.co/cCYFFEcrfV
#ad

8) Here’s how you can enter:
1) Follow @AmTourister
2) Make a remix of my reel and mention where you’re heading this summer.
3) Tag @AMtouristerIN
4) Use the hashtag #UndeniableLeave

9) Girls and guys, are you still worried about getting your leaves approved? I’ve
American Tourister’s #UndeniableLeave contest is now open on their Instagram page.

10) https://t.co/eb4xdoeM1c

11) Back to my favourite 🏋. With my favourite @AnushkaSharma ❤️ https://t.co/g4U

12) Nothing like the one8 experience indeed.
Head to the link for exclusive summer offers - https://t.co/SgbvXildmU
```

Image – Last 20 Tweets of @imVKohli

We will store these tweets in a dataFrame which is available in pandas library. DataFrames are 2D data structure just like 2D array with rows and columns. These are used mainly in data science, machine learning and many more other data requiring fields.



index	Tweets
0	https://t.co/5dGRZdGNN2
1	Who says the work can stop? https://t.co/nBWxrKoucD
2	If you're wild &amp; free and it feels so right, can it be Wrogn? 😉 Know more by following @StayWrogn or clicking the link here 👉🔗 https://t.co/RpaiNELQPX #StayWrogn #WrognTribe #ad https://t.co/AnOBhJzXvH
3	My adventure with #pumaindia has been one of the best experiences outside of cricket. From creating a platform for sports &amp; fitness to designing the one8 collection, it's been a heartening 5-year journey. I'm proud to be a part of the No. 1 sports &amp; lifestyle brand in India🖤 #ad https://t.co/nWcUmEYMGZ
4	Firecracker game and a great win. 🧨 @RCBTweets https://t.co/KBeCORpBTF
5	A lot of love and happiness to all mother's. Your strength is unmatched and here's wishing you a very Happy Mother's Day. 👸
6	Head over to American Tourister's Instagram page to remix the reel and stand a chance to win a cool American Tourister backpack! T&amp;C: https://t.co/cCYFFEcrfV #ad
7	Here's how you can enter: 1) Follow @AmTourister 2) Make a remix of my reel and mention where you're heading this summer. 3) Tag @AMtouristerIN 4) Use the hashtag #UndeniableLeave
8	Girls and guys, are you still worried about getting your leaves approved? I've got a way to make sure you get it! American Tourister's #UndeniableLeave contest is now open on their Instagram page. https://t.co/gsHDQXkEzI
9	https://t.co/eb4xdoeM1c
10	Back to my favourite 🏋. With my favourite @AnushkaSharma ❤️ https://t.co/g4UnoNNZkx
11	Nothing like the one8 experience indeed. Head to the link for exclusive summer offers - https://t.co/SgbvXildmU @ScentialsWorld @one8world #one8fragrances #ad https://t.co/JwnzJYoUd4
12	Thank god you were born ❤️. I don't know what I would do without you. You're truly beautiful inside out ❤️. Had a great afternoon with the sweetest folks around 😃 @AnushkaSharma https://t.co/JxGEnBtHXW
13	Good vibes in the camp ✌ @RCBTweets https://t.co/rTTPHdoDzx
14	Retired hurt interiors messing up your recipes? I definitely 'woodn't' want Anushka to have this. Get the best quality materials and a seamless finish with @livspace to #LoveTheWayYouLiv #ad https://t.co/jTSCILvjS6
15	Another game 🏏 Another win 💯 Onwards &amp; Upwards 💪 @RCBTweets https://t.co/tlELYOI2VO
16	Keep ticking small boxes everyday. 🏏 ❤️ https://t.co/GGzRknoZPD

Image – Tweets stored in data frame

4.3. Cleaning the text

A tweet consists of many things which are not necessary for our analysis.
These things include

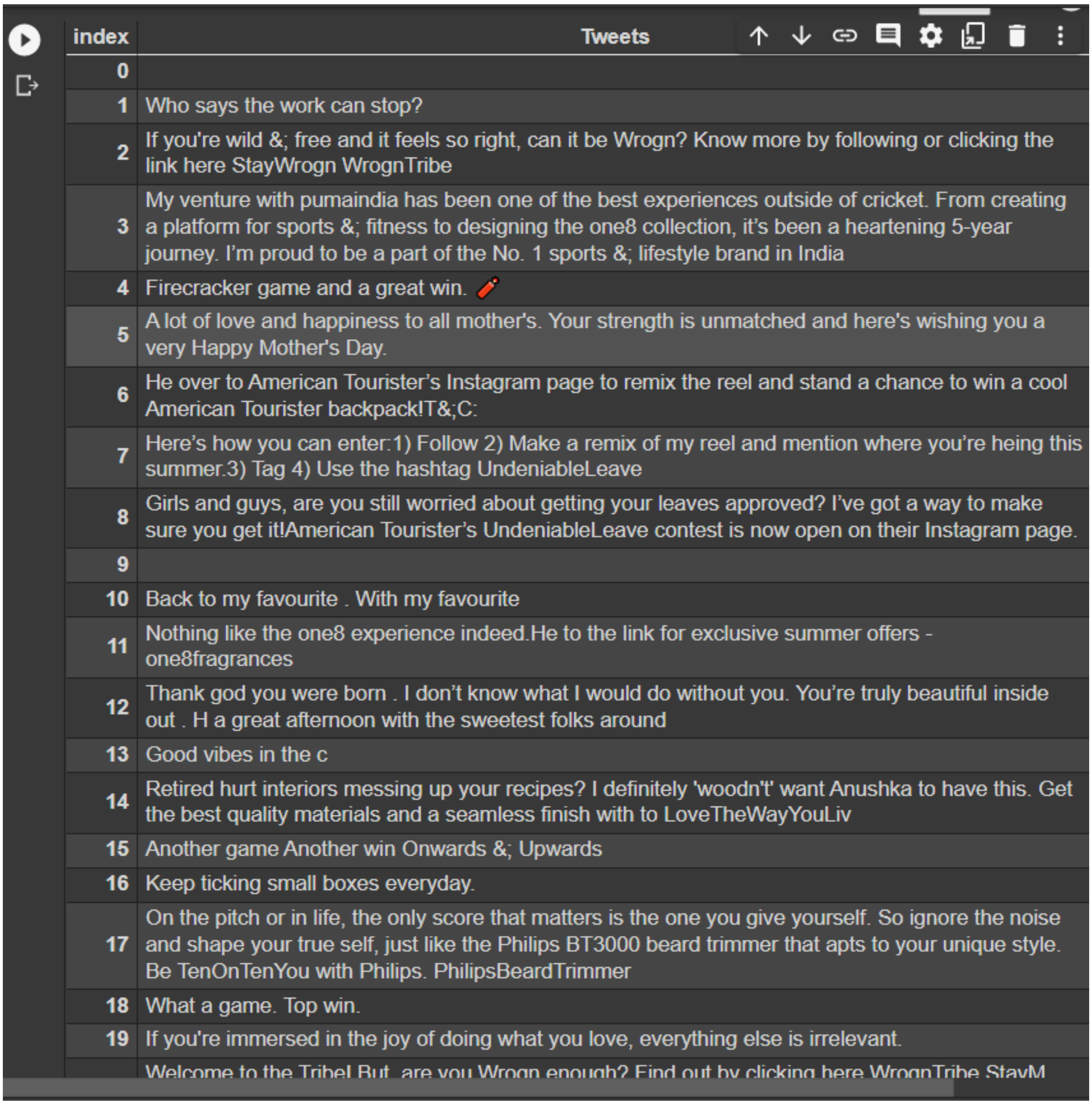
- Emoji – Nowadays, people use emojis everywhere, so its important to remove them before moving ahead

- Mentions (@) – Twitter is a interactive website and people mentions others while commenting but these mentions don’t contribute towards our analysis, so we will remove them.

- Tags (#) – Tags are related to the topic user is commenting. WE will remove it.

- Hyperlinks (https\\) – So many tweets contain hyperlinks about external sites which doesn’t assist in our analysis.

- Retweets – Retweets means reposting someone’s tweets. We will remove it since its someone else tweets and not of user itself.



index	Tweets
0	
1	Who says the work can stop?
2	If you're wild &; free and it feels so right, can it be Wrogn? Know more by following or clicking the link here StayWrogn WrognTribe
3	My venture with pumaindia has been one of the best experiences outside of cricket. From creating a platform for sports &; fitness to designing the one8 collection, it's been a heartening 5-year journey. I'm proud to be a part of the No. 1 sports &; lifestyle brand in India
4	Firecracker game and a great win. 🧨
5	A lot of love and happiness to all mother's. Your strength is unmatched and here's wishing you a very Happy Mother's Day.
6	He over to American Tourister's Instagram page to remix the reel and stand a chance to win a cool American Tourister backpack!T&;C:
7	Here's how you can enter:1) Follow 2) Make a remix of my reel and mention where you're heing this summer.3) Tag 4) Use the hashtag UndeniableLeave
8	Girls and guys, are you still worried about getting your leaves approved? I've got a way to make sure you get it!American Tourister's UndeniableLeave contest is now open on their Instagram page.
9	
10	Back to my favourite . With my favourite
11	Nothing like the one8 experience indeed.He to the link for exclusive summer offers - one8fragrances
12	Thank god you were born . I don't know what I would do without you. You're truly beautiful inside out . H a great afternoon with the sweetest folks around
13	Good vibes in the c
14	Retired hurt interiors messing up your recipes? I definitely 'woodn't' want Anushka to have this. Get the best quality materials and a seamless finish with to LoveTheWayYouLiv
15	Another game Another win Onwards &; Upwards
16	Keep ticking small boxes everyday.
17	On the pitch or in life, the only score that matters is the one you give yourself. So ignore the noise and shape your true self, just like the Philips BT3000 beard trimmer that apts to your unique style. Be TenOnTenYou with Philips. PhilipsBeardTrimmer
18	What a game. Top win.
19	If you're immersed in the joy of doing what you love, everything else is irrelevant.
	Welcome to the Tribe! But, are you Wrogn enough? Find out by clicking here WrognTribe StayM

Image – Tweets after pre-processing

4.4. Get Subjectivity and Polarity

Subjectivity – Subjectivity refers to the personal judgement, thinking and opinion.
The output of subjectivity lies between [0,1]

Polarity – The main thing in sentiment analysis is to analyze a tweet and extract positive or negative sentiment out of it. These sentiments are denoted in values : positive values for positive emotions and negative values for negative emotions.

This polarity score defines if the overall tweet is positive or negative.
Its score ranges from -1(very negative) to +1 (very positive).

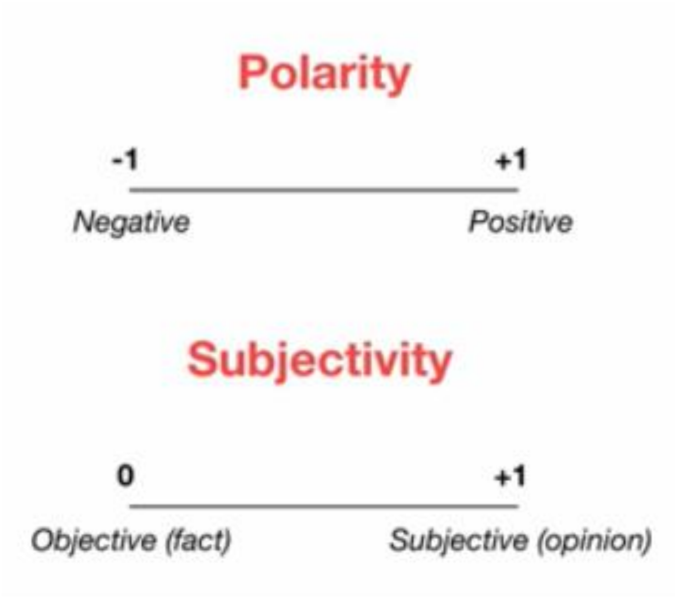
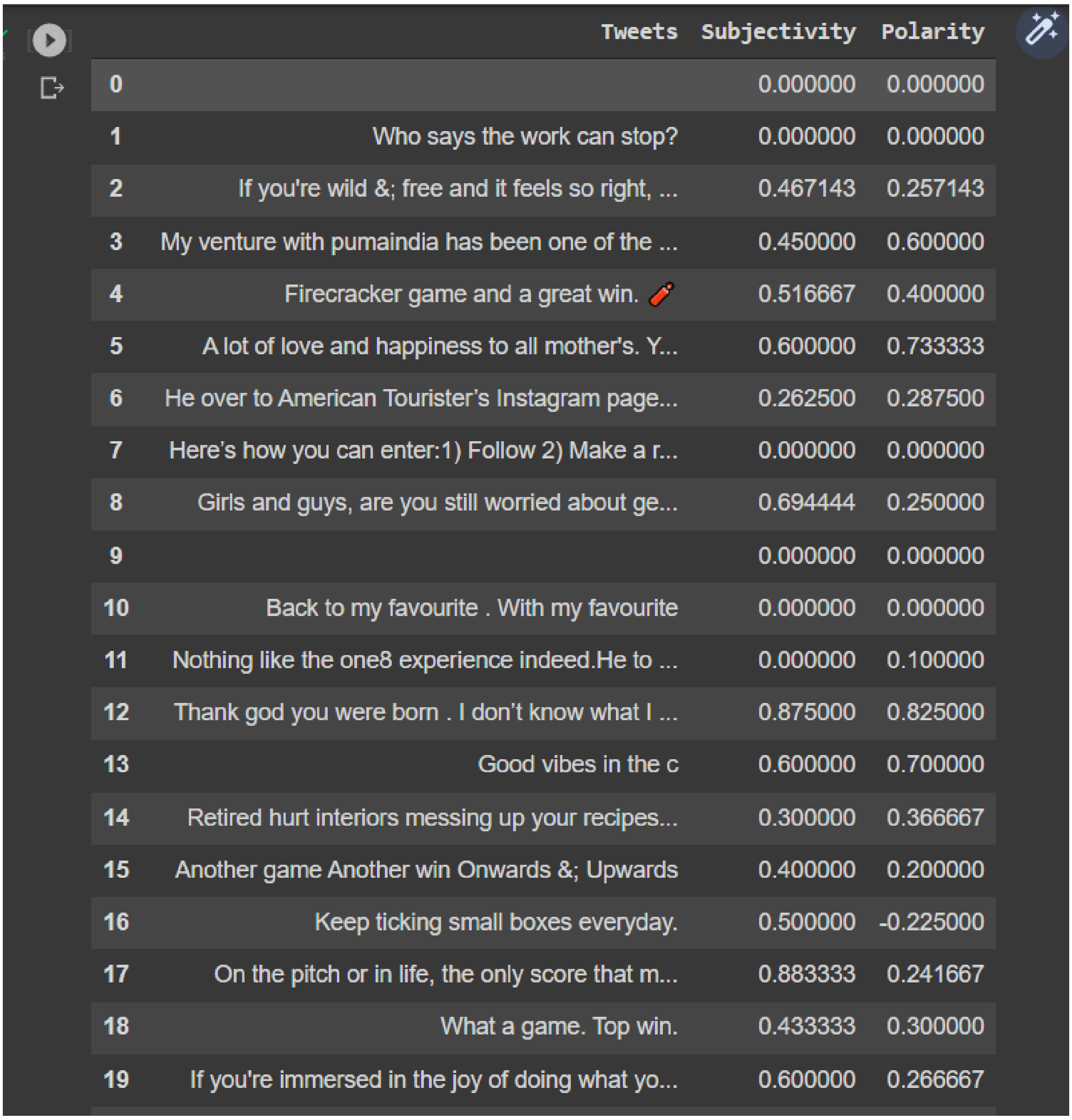


Image – Polarity and Subjectivity



	Tweets	Subjectivity	Polarity
0		0.000000	0.000000
1	Who says the work can stop?	0.000000	0.000000
2	If you're wild &; free and it feels so right, ...	0.467143	0.257143
3	My venture with pumaindia has been one of the ...	0.450000	0.600000
4	Firecracker game and a great win. 🧨	0.516667	0.400000
5	A lot of love and happiness to all mother's. Y...	0.600000	0.733333
6	He over to American Tourister’s Instagram page...	0.262500	0.287500
7	Here’s how you can enter:1) Follow 2) Make a r...	0.000000	0.000000
8	Girls and guys, are you still worried about ge...	0.694444	0.250000
9		0.000000	0.000000
10	Back to my favourite . With my favourite	0.000000	0.000000
11	Nothing like the one8 experience indeed.He to ...	0.000000	0.100000
12	Thank god you were born . I don’t know what I ...	0.875000	0.825000
13	Good vibes in the c	0.600000	0.700000
14	Retired hurt interiors messing up your recipes...	0.300000	0.366667
15	Another game Another win Onwards &; Upwards	0.400000	0.200000
16	Keep ticking small boxes everyday.	0.500000	-0.225000
17	On the pitch or in life, the only score that m...	0.883333	0.241667
18	What a game. Top win.	0.433333	0.300000
19	If you're immersed in the joy of doing what yo...	0.600000	0.266667

Image – Polarity and Subjectivity on @imVKohli tweets

Plotting the word cloud –

I plotted the word cloud for the tweets assembled here.

The most common words are shown in bigger font.

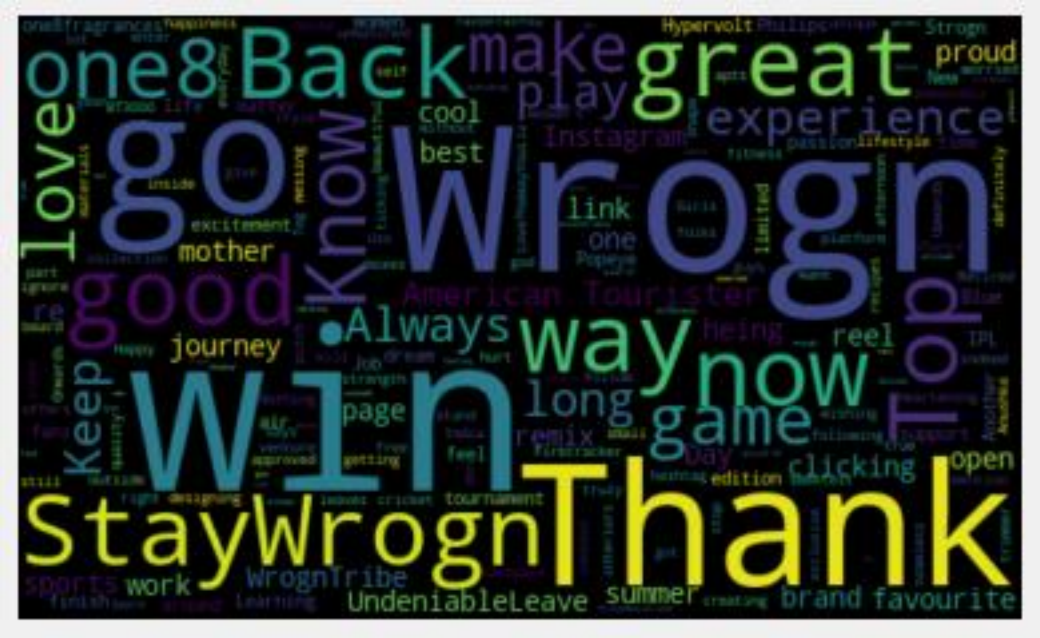
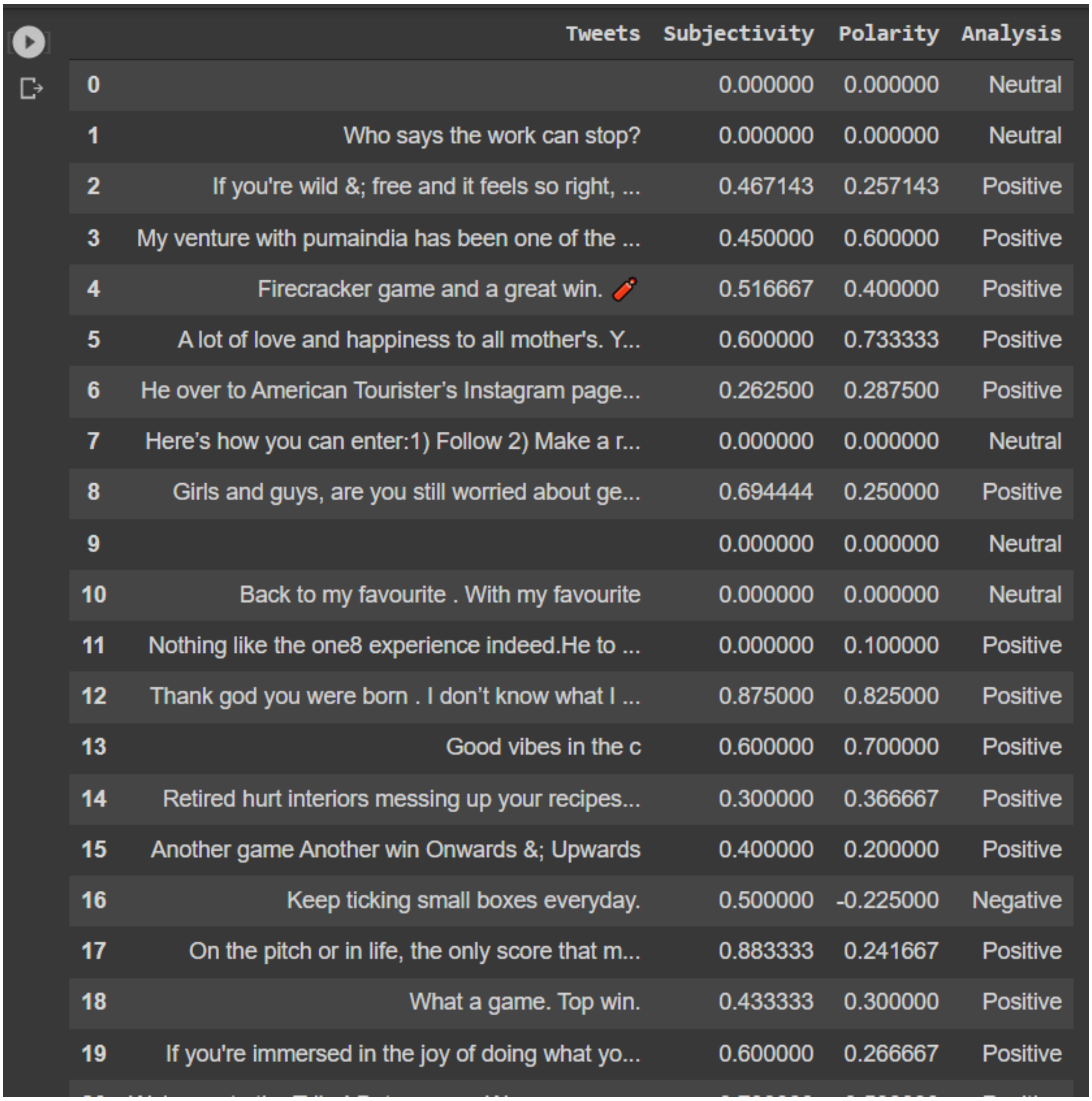


Image – Word cloud of @imVKohli tweets

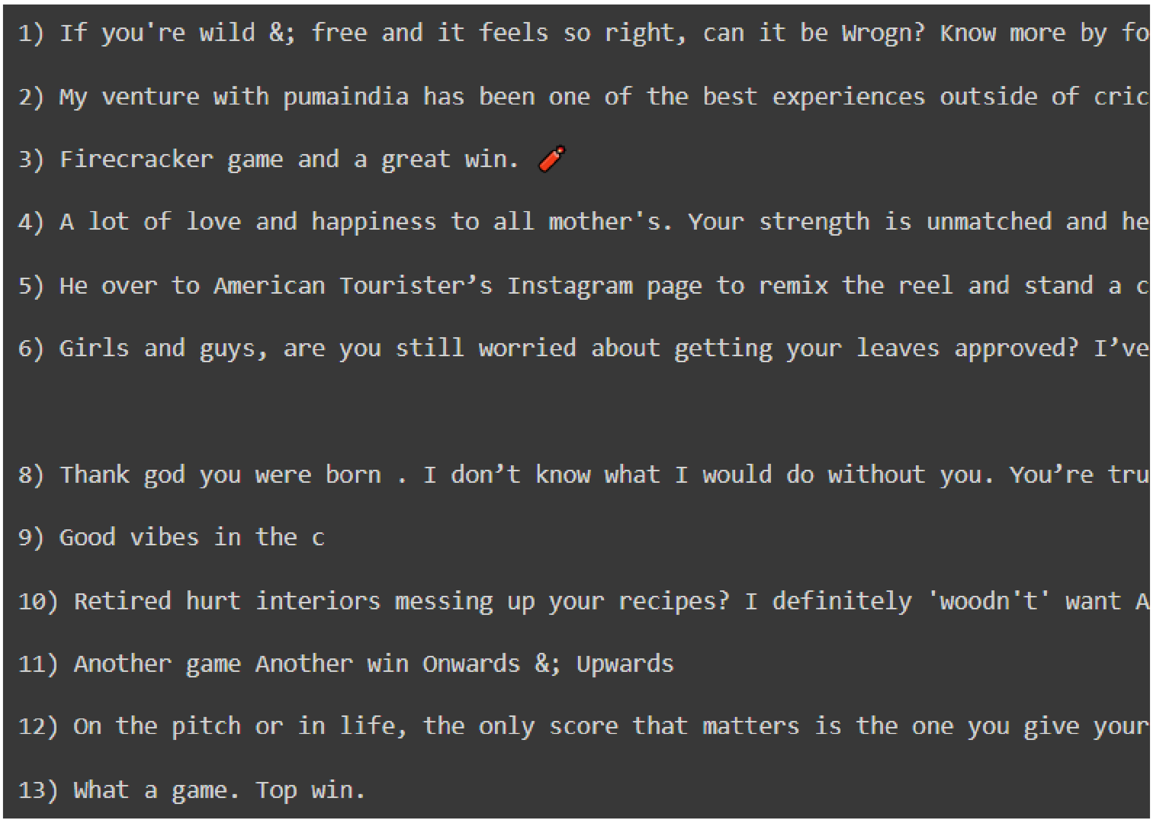
Getting sentiment analysis and labelling tweets positive, negative or neutral –



	Tweets	Subjectivity	Polarity	Analysis
0		0.000000	0.000000	Neutral
1	Who says the work can stop?	0.000000	0.000000	Neutral
2	If you're wild &; free and it feels so right, ...	0.467143	0.257143	Positive
3	My venture with pumaindia has been one of the ...	0.450000	0.600000	Positive
4	Firecracker game and a great win. 🧨	0.516667	0.400000	Positive
5	A lot of love and happiness to all mother's. Y...	0.600000	0.733333	Positive
6	He over to American Tourister’s Instagram page...	0.262500	0.287500	Positive
7	Here’s how you can enter:1) Follow 2) Make a r...	0.000000	0.000000	Neutral
8	Girls and guys, are you still worried about ge...	0.694444	0.250000	Positive
9		0.000000	0.000000	Neutral
10	Back to my favourite . With my favourite	0.000000	0.000000	Neutral
11	Nothing like the one8 experience indeed.He to ...	0.000000	0.100000	Positive
12	Thank god you were born . I don’t know what I ...	0.875000	0.825000	Positive
13	Good vibes in the c	0.600000	0.700000	Positive
14	Retired hurt interiors messing up your recipes...	0.300000	0.366667	Positive
15	Another game Another win Onwards &; Upwards	0.400000	0.200000	Positive
16	Keep ticking small boxes everyday.	0.500000	-0.225000	Negative
17	On the pitch or in life, the only score that m...	0.883333	0.241667	Positive
18	What a game. Top win.	0.433333	0.300000	Positive
19	If you're immersed in the joy of doing what yo...	0.600000	0.266667	Positive

Image – Sentiment labelling of @imVkohli tweets.

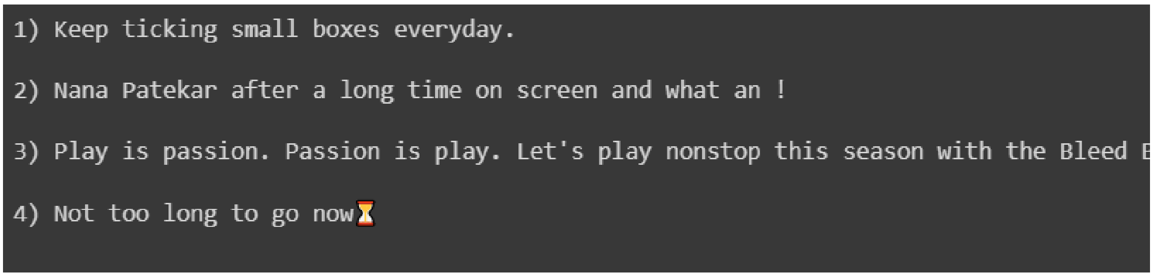
#Positive Tweets –



```
1) If you're wild &; free and it feels so right, can it be Wrogn? Know more by fo

2) My venture with pumaindia has been one of the best experiences outside of cric

3) Firecracker game and a great win. 🧨

4) A lot of love and happiness to all mother's. Your strength is unmatched and he

5) He over to American Tourister’s Instagram page to remix the reel and stand a c

6) Girls and guys, are you still worried about getting your leaves approved? I’ve


8) Thank god you were born . I don’t know what I would do without you. You’re tru

9) Good vibes in the c

10) Retired hurt interiors messing up your recipes? I definitely 'woodn't' want A

11) Another game Another win Onwards &; Upwards

12) On the pitch or in life, the only score that matters is the one you give your

13) What a game. Top win.
```

Image – Positive tweets

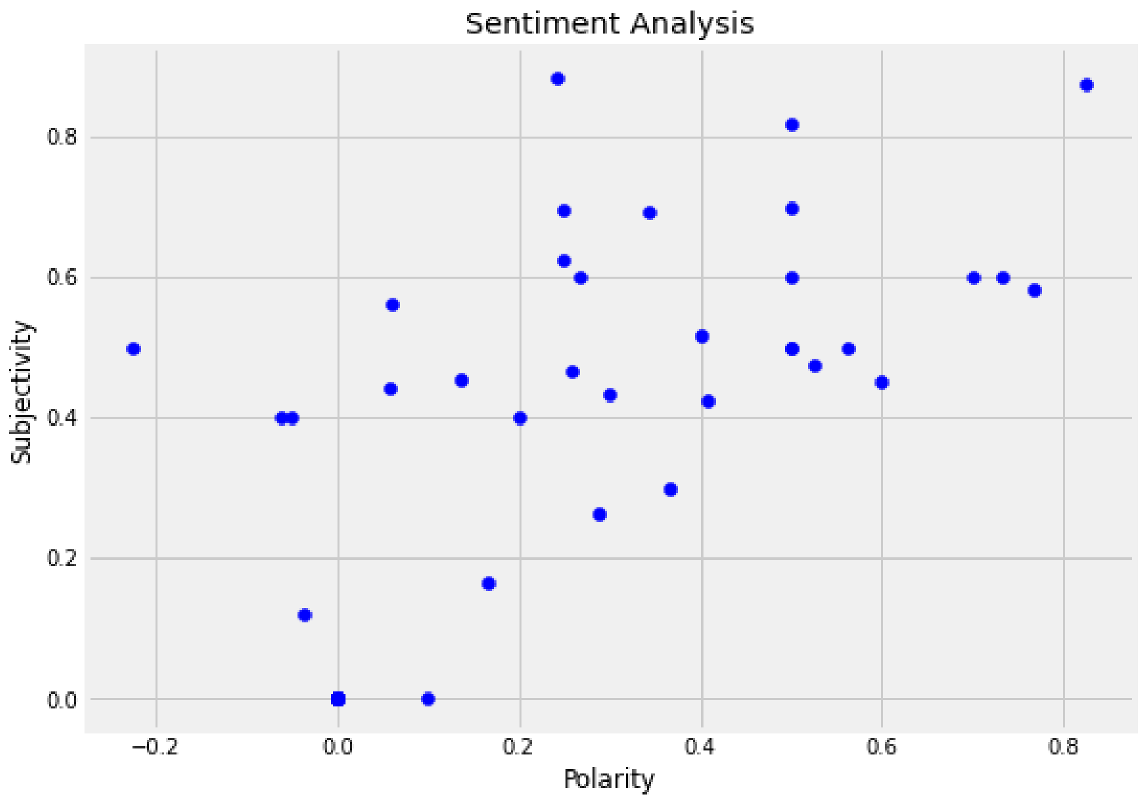
#Negative Tweets –



```
1) Keep ticking small boxes everyday.

2) Nana Patekar after a long time on screen and what an !

3) Play is passion. Passion is play. Let's play nonstop this season with the Bleed B

4) Not too long to go now⏳
```

Image – Negative Tweets

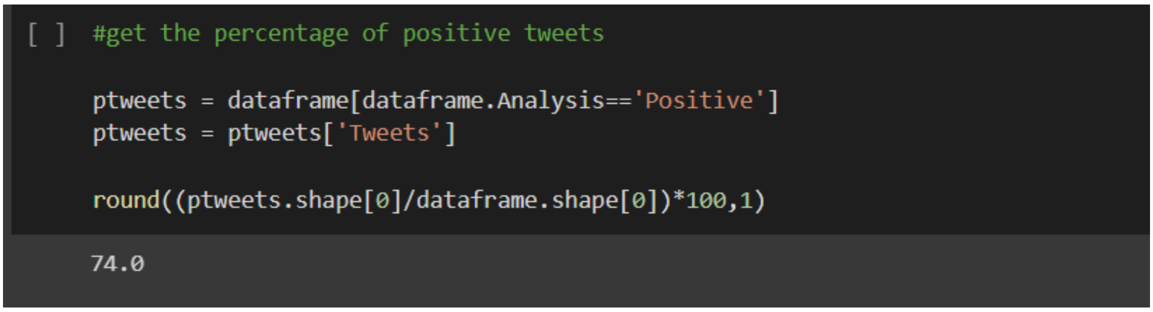
#Plotting polarity and subjectivity on a map –



Graph – Plot of subjectivity and polarity of @imVKohli tweets

Here we can see most tweets of Virat Kohli shows on the positive side of graph

#Percentage of positive tweets –

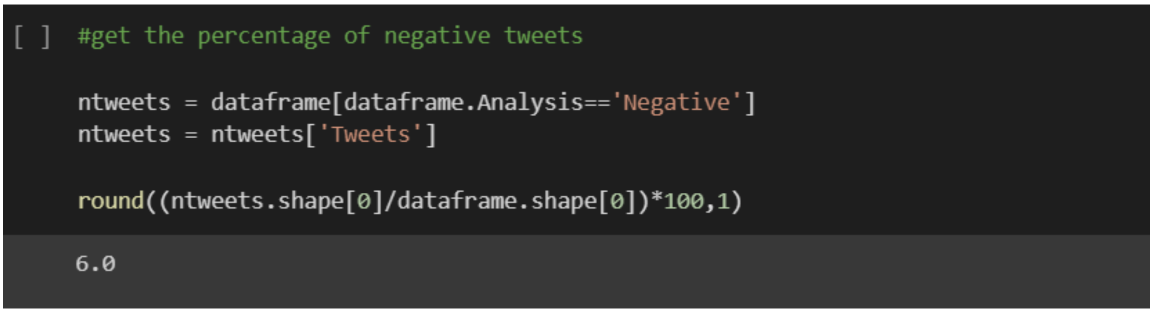


```
#get the percentage of positive tweets

ptweets = dataframe[dataframe.Analysis=='Positive']
ptweets = ptweets['Tweets']

round((ptweets.shape[0]/dataframe.shape[0])*100,1)
```

```
74.0
```

Image – Percentage of positive tweets

#Percentage of negative tweets –

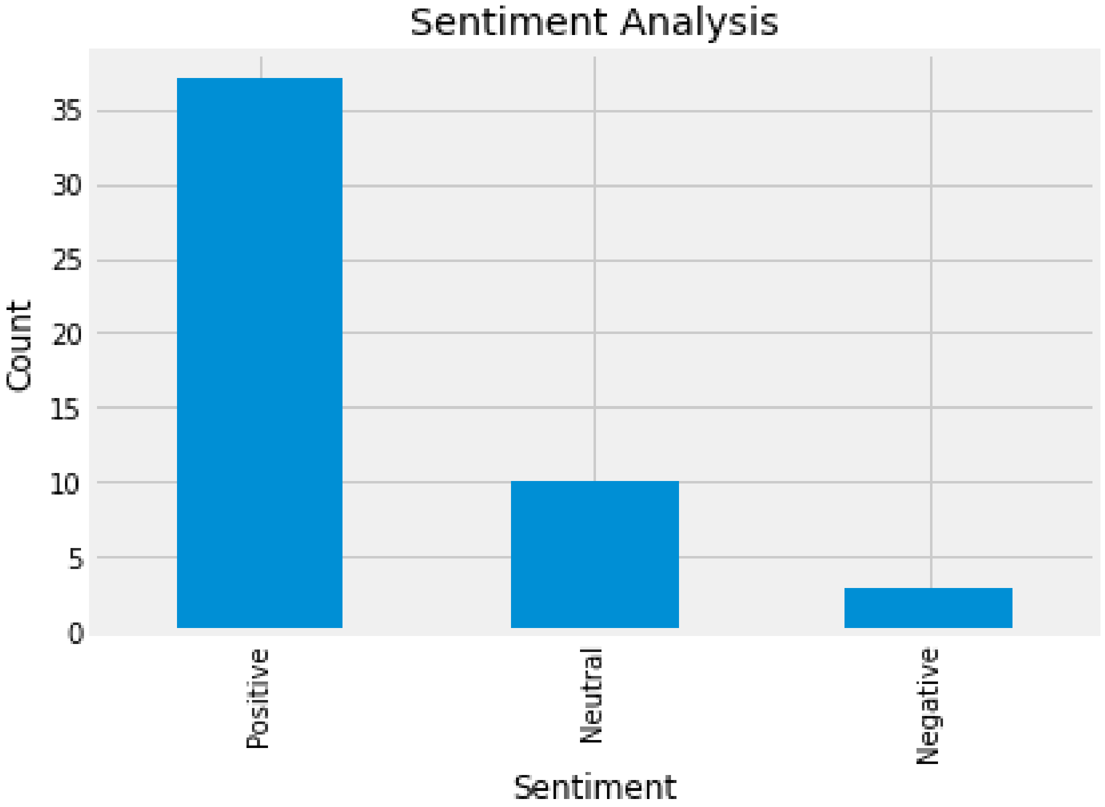


```
#get the percentage of negative tweets

ntweets = dataframe[dataframe.Analysis=='Negative']
ntweets = ntweets['Tweets']

round((ntweets.shape[0]/dataframe.shape[0])*100,1)
```

```
6.0
```

Image – Percentage of negative tweets

#Value counts of positive, negative and neutral element –



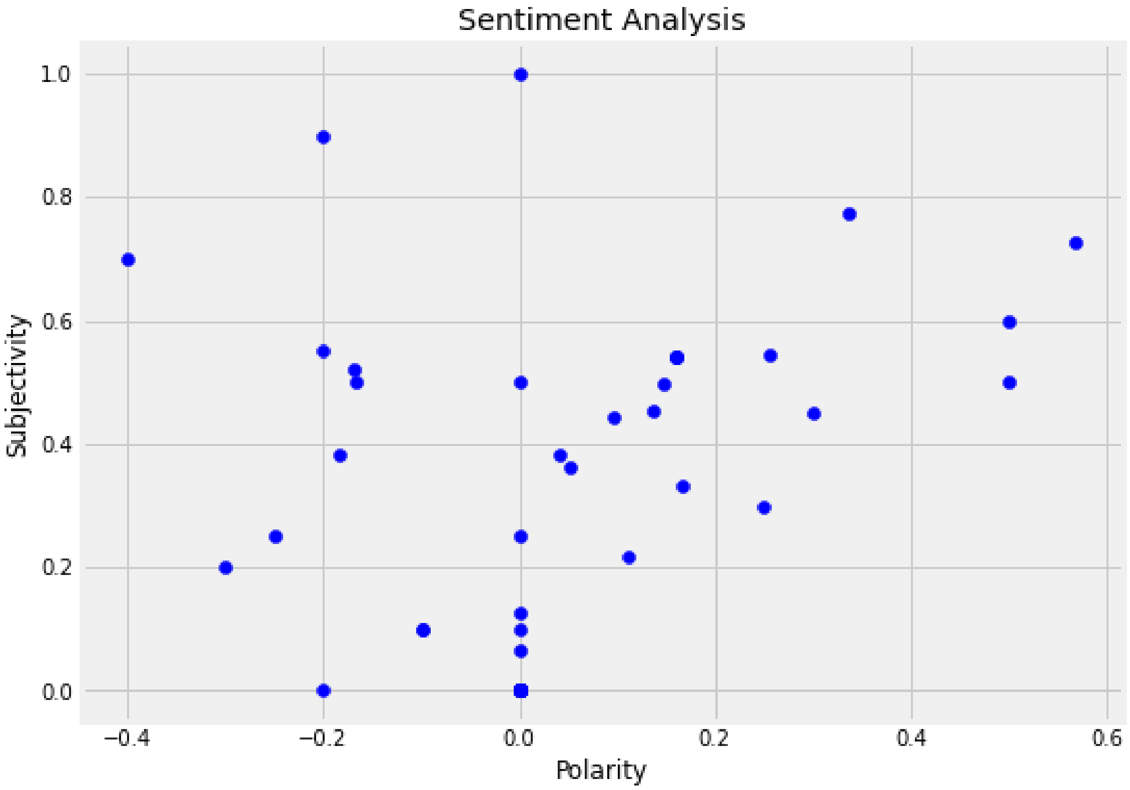
Graph – Shows count of positive, negative and neutral tweets

4.5 Result –

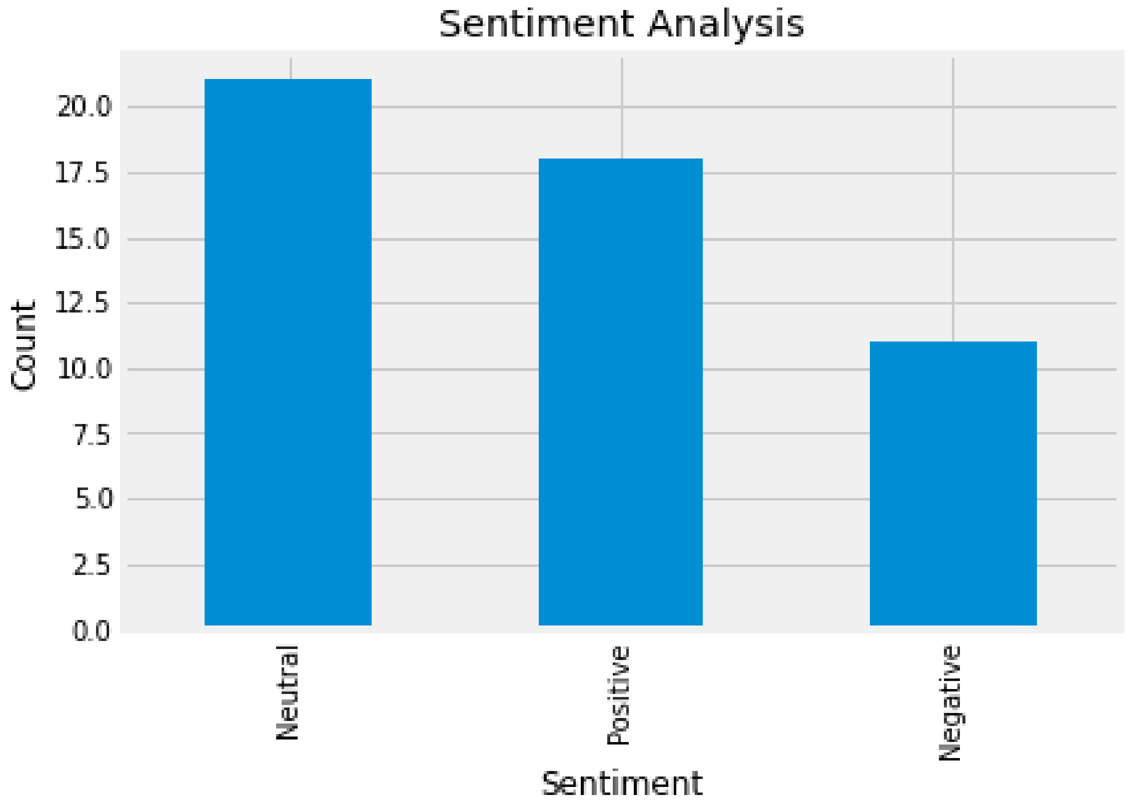
Sentiment Analysis of Virat Kohli’s twitter handle showed positive tweet percentage of 74% and negative tweet percentage of 6% while others are neutral.

I also tried this analysis on a
A) popular news channel **India Today** (@IndiaToday)

Positive Tweets - 36%
Negative Tweets – 22%



Graph – Plot of subjectivity and polarity of @IndiaToday tweets

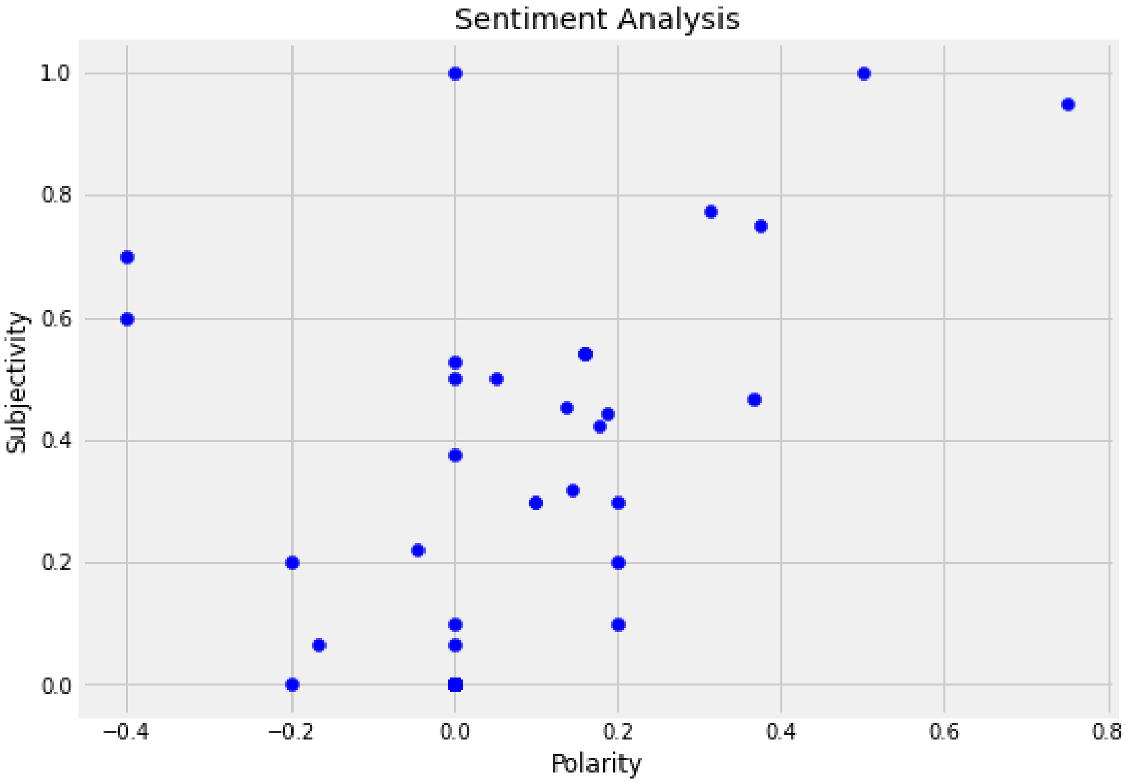


Graph – Shows count of positive, negative and neutral tweets of @IndiaToday

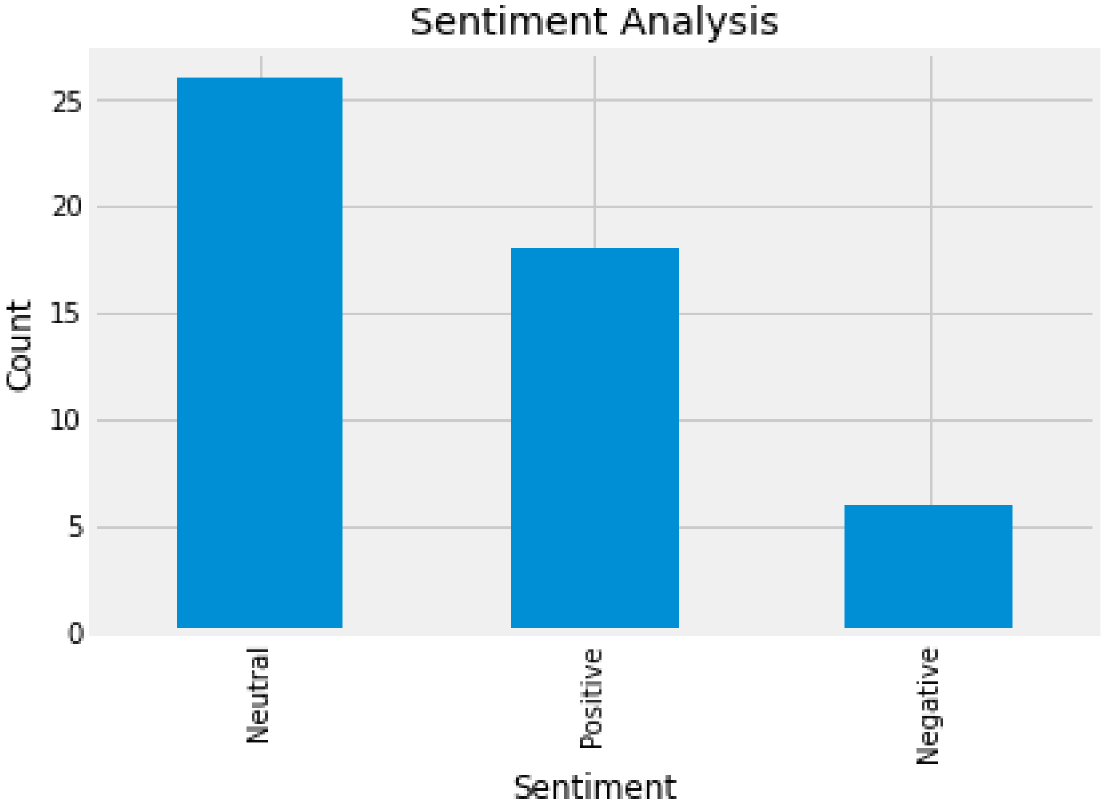
B) Another Online newspaper “**Times of India**” (@timesofindia)

Positive Tweets – 36%

Negative Tweets – 12%



Graph – Plot of subjectivity and polarity of @timeofindia tweets



Graph – Shows count of positive, negative and neutral tweets of @timesofIndia

CHAPTER 5 – PROJECT CONCLUSION

Conclusion

Emotional stress has become a leading concern all over the world. I have taken many lives and more and more people are getting affected by it. We analized sentiments of various tweeter handles to get what emotions they are inciting and what sentiments they are pointing to. In our project, we were able to identify and plot positive, negative and neutral tweets and show them in graphical format

Future Scope

In the future we can use various other classifiers to come with better accuracy and this model can be even further upgraded to GUI interface get live sentiment of people. This can be used with a graphic interface to deliver the same results with better accessibility.

REFERENCES

[1] Haller DM, Sanci LA, Sawyer SM, Patton GC. The identification of young people’s emotional distress: a study in primary care. Br J Gen Pract. 2009 Mar;59(560):e61-70. doi: 10.3399/bjgp09X419510. PMID: 19275825; PMCID: PMC2648934.

[2] World Health Organization. (n.d.). SDG Target 3.4 Noncommunicable diseases and mental health. World Health Organization. https://www.who.int/data/gho/data/themes/topics/sdg-target-3 4-noncommunicable-diseases-and-mental-health.

[3] Praveen. (2020, April 16). Emotions dataset for NLP. Kaggle. https://www.kaggle.com/praveengovi/emotions-dataset-for-nlp.

[4] El Alaoui, I., Gahi, Y., Messoussi, R. et al. A novel adaptable approach for sentiment analysis on big social data. J Big Data 5, 12 (2018). https://doi.org/10.1186/s40537-018-0120-0

[5] B. Seref and E. Bostanci, ”Sentiment Analysis using Naive Bayes and Complement Naive Bayes Classifier Algorithms on Hadoop Framework,” 2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), 2018, pp. 1-7, doi: 10.1109/ISMSIT.2018.8567243.

[6] S. Kaur, G. Sikka and L. K. Awasthi, ”Sentiment Analysis Approach Based on N-gram and KNN Classifier,” 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), 2018, pp. 1-4, doi: 10.1109/ICSCCC.2018.8703350

[7] Zainuddin, Nurulhuda & Selamat, Ali. (2014). Sentiment analysis using Support Vector Machine. I4CT 2014 - 1st International Conference on Computer, Communications, and Control Technology, Proceedings. 333- 337. 10.1109/I4CT.2014.6914200.

[8] Wankhade, Mayur & Chandra, A & Rao, Sekhara & Dara, Suresh & Kaushik, Baij. (2017). A Sentiment Analysis of Food Review using Logistic Regression. 2456-3307.

[9] Ramosaco, Miftar & Hasani, Vjollca & Dumi, Alba. (2015). Application of Logistic Regression in the Study of Students’ Performance Level (Case Study of Vlora University). Journal of Educational and Social Research. 10.5901/jesr.2015.v5n3p239.

[10] Singh, Gurinder & Kumar, Bhawna & Gaur, Loveleen & Tyagi, Akriti. (2019). Comparison between Multinomial and Bernoulli Na¨ıve Bayes for Text Classification. 593-596. 10.1109/ICACTM.2019.8776800.

[11] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

[12] The Bernoulli model - https://nlp.stanford.edu/IRbook/html/htmledition/the-bernoulli-model-1.html

[13] A. Verma and S. Mehta, ”A comparative study of ensemble learning methods for classification in bioinformatics,” 2017 7th International Conference on Cloud Computing, Data Science & Engineering - Confluence, 2017, pp. 155-158, doi: 10.1109/CONFLUENCE.2017.7943141.

[14] Bauer, E. & Kohavi, R. (1999). An empirical comparison of voting classification algorithms. Machine Learning, 36(1/2), 105–139.