GALGOTIAS
UNIVERSITY
(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

CLASSIFIER FOR ACCURATE AND AUTOMATIC SENTIMENTNT CLASSIFICATION

A Project Report of Capstone Project - 2

Submitted by

KASHIF ZAIDI

(1613101330 / 16SCSE113018)

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

IN

Computer Science and Engineering

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING

Under the Supervision of

DR. SANJEEV KUMAR PRASAD,

Associate Professor

APRIL / MAY - 2020

SCHOOL OF COMPUTING AND SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

Certified that this project report **“CLASSIFIER FOR ACCURATE AND AUTOMATIC SNT CLASSIFICATION”** is the bonafide work of **“KASHIF ZAIDI(1613101330)”** who carried out the project work under my supervision**.**

SIGNATURE
Dr. MUNISH SHABARWAL
HEAD OF DEPARTMENT

School of Computing Science
& Engineering

SIGNATURE
Dr. SANJEEV KUMAR PRASAD
SUPERVISOR
Associate Professor
School of Computing Science
& Engineering

ABSTRACT

In this project, we investigate the utility of the linguistic features for detecting the snt of Twt messages. Twt is an online micro-blogging and social-networking platform which allows users to write short status updates of maximum length 140 characters. It is a rapidly expanding service with over 200 million registered users out of which 100 million are active users and half of them log on twt on a daily basis - generating nearly 250 million tweets per day. We evaluate the existing resources (lexical) as well as features that capture information about the informal and creative language used in microblogging. We crawled the political tweets during the general election in India, take report on Prime Minister’s account, analyzed the stock market values and further evaluate our proposed approach against all the topics. Due to this large amount of usage we hope to achieve a reflection of public snt by analyzing the snts expressed in the tweets. Analyzing the public snt is important for many applications. The aim of this project is to develop a functional classifier for accurate and automatic snt classification of an unknown tweet stream. Experimental results show the effectiveness of the proposed rules in determining the snt. Anlysis using social media is getting attention of many researchers to understand the public opinion and trend. In this project, we proposed the visualization of tweets based on single area. It will provide the polarity, number of retweets and then visualize them.

TABLE OF CONTENTS

LIST OF TABLES

TABLE OF FIGURES

TABLE OF ABBREVIATIONS

S. No	**Original Word**	**Used**
1	Twitter	twt
2	Sentiment	snt
3	Analysis	anlysis

4. INTRODUCTION

4.1 Overall Description

People share knowledge, experiences and thoughts with the world by using Social Media like blogs, forums, wikis, review sites, social networks, tweets and so on. This has changed the manner in which people communicate and influence social, political and economic behavior of other people in the Web 2.0. Indeed the Web 2.0 allows everyone having a voice, promising to boost human collaboration capabilities on a worldwide scale, enabling individuals to share opinions by means of read-write Web and user’s generated contents.

Micro-blogging websites have evolved to become a source of varied kinds of information. This is due to nature of micro-blogs on which people post real-time messages about their opinions on a variety of topics, discuss current issues, complain, and express positive snt for products they use in daily life. In fact, companies manufacturing such products have started to poll these micro-blogs to get a sense of general snt for their product.

Many times these companies study user reactions and reply to users on micro-blogs. One challenge is to build technology to detect and summarize an overall snt. In this paper, we look at one such popular micro-blog called Twt and build models for classifying “tweets” into positive, negative and neutral snt.

The snt anlysis is a complex process that involves 5 different steps to analyze snt data. These steps are:

- **Data collection:** the first step of snt anlysis consists of collecting data from user-generated content contained in blogs, forums, social networks. These data are disorganized, expressed in different ways by using different vocabularies, slangs, context of writing, etc. Manual anlysis is almost impossible. Therefore, text analytics and natural language processing are used to extract and classify.

- **Text preparation:** consists of cleaning the extracted data before anlysis. Non-textual contents and contents that are irrelevant for the anlysis are identified and eliminated.

- **Snt detection:** the extracted sentences of the reviews and opinions are examined. Sentences with subjective expressions (opinions, beliefs and views) are retained and sentences with objective communication (facts, factual information) are discarded.

- **Snt classification:** in this step, subjective sentences are classified in positive, negative, good, bad; like, dislike, but classification can be made by using multiple points.

- **Presentation of output:** the main objective of snt anlysis is to convert unstructured text into meaningful information. When the anlysis is finished, the text results are displayed on graphs like pie chart, bar chart and line graphs. Also time can be analyzed and can be graphically displayed constructing a snt time line with the chosen value (frequency, percentages, and averages) over time.

We build models for two classification tasks: a binary task of classifying snt into positive and negative classes and a 3-way task of classifying snt into positive, negative and neutral classes. We use manually annotated Twt data for our experiments. One advantage of this data, over previously used data-sets, is that the tweets are collected in a streaming fashion and therefore represent a true sample of actual tweets in terms of language use and content.

With more than 321 million active users, sending a daily average of 500 million Tweets, Twt allows businesses to reach a broad audience and connect with customers without intermediaries. On the downside, it’s harder for brands to quickly detect negative content, and if it goes viral you might end up with an unexpected PR crisis on your hands. This is one of the reasons why *social listening* — monitoring conversation and feedback in social media — has become a crucial process in social media marketing.

Monitoring Twt allows companies to understand their audience, keep on top of what’s being said about their brand and their competitors, and discover new trends in the industry. Are users talking positively or negatively about a product? Well, that’s exactly what snt anlysis determines.

Text understanding is a significant problem to solve. Some machine learning techniques, including various supervised and unsupervised algorithms, are being utilized. There are different approaches to generate summary. One approach could be rank the importance of sentences within the text and then generate summary for the text based on the importance numbers. There is another approach called end-to-end generative models. In some domain like image recognition, speech recognition, language translation, and question-answering, the end-to-end method performs better.

Twt nowadays is one of the popular social media which according to the statistics currently has over 300 million accounts. Twt is the rich source to learn about people’s opinion and sntal anlysis. For each tweet it is important to determine the snt of the tweet whether is it positive, negative, or neutral. Another challenge with twt is only 140 characters is the limitation of each tweet which cause people to use phrases and works which are not in language processing. Recently twt has extended the text limitations to 280 characters per each tweet.

Figure 4-1: Twt Logo

4.2 Purpose

This project addresses the problem of snt anlysis in twt; that is classifying tweets according to the snt expressed in them: positive, negative or neutral. Twt is an online micro-blogging and social-networking platform which allows users to write short status updates of maximum length 140 characters. Due to this large amount of usage we hope to achieve a reflection of public snt by analyzing the snts expressed in the tweets. Analyzing the public snt is important for many applications such as firms trying to find out the response of their products in the market, predicting political elections and predicting socioeconomic phenomena like stock exchange.

Snt Anlysis Dataset Twt has a number of applications:

Business: Companies use Twt Snt Anlysis to develop their business strategies, to assess customers' feelings towards products or brand, how people respond to their campaigns or product launches and also why consumers are not buying certain products.

Figure 4-2: Twt Business

Politics: In politics Snt Anlysis Dataset Twt is used to keep track of political views, to detect consistency and inconsistency between statements and actions at the government level. Snt Anlysis Dataset Twt is also used for analyzing election results.

Figure 4-3: Twt Politics

Public Actions: Twt Snt Anlysis also is used for monitoring and analysing social phenomena, for predicting potentially dangerous situations and determining the general mood of the blogosphere.

4.3 Motivation and Scope

We have chosen to work with twt since we feel it is a better approximation of public snt as opposed to conventional internet articles and web blogs. The reason is that the amount of relevant data is much larger for twt, as compared to traditional blogging sites. More over the response on twt is more prompt and also more general (since the number of users who tweet is substantially more than those who write web blogs on a daily basis).[1] Snt anlysis of public is highly critical in macro-scale socioeconomic phenomena like predicting the stock market rate of a particular firm. This could be done by analyzing overall public snt towards that firm with respect to time and using economics tools for finding the correlation between public snt and the firm's stock market value. Firms can also estimate how well their product is responding in the market, which areas of the market is it having a favourable response and in which a negative response (since twt allows us to download stream of geo-tagged tweets for particular locations. If firms can get this information they can analyze the reasons behind geographically differentiated response, and so they can market their product in a more optimized manner by looking for appropriate solutions like creating suitable market segments. Predicting the results of popular political elections and polls is also an emerging application to snt anlysis.

5. LITERATURE SURVEY

Snt anlysis has been handled as a Natural Language Processing task at many levels of granularity. Micro-blog data like Twt, on which users post real time reactions to and opinions about “everything”, poses newer and different challenges.

They use tweets ending in positive emoticons like “😊” “😊” as positive and negative emoticons like “☹” “☹” as negative. They build models using Naïve Bayes and Support Vector Machines (SVM), and they report SVM outperforms other classifiers.

The Snt classification is a task of classifying a target unit in a document to positive (favorable) or negative (unfavorable) class. There are three main classification levels:

- document level: classifies an opinion document as expressing a positive or negative opinion or snt. It considers the whole document a basic information unit.

- sentence-level: classifies snt expressed in each sentence. If the sentence is subjective it classifies it in positive or negative opinions.

- aspect-level: classifies the snt with respect to the specific aspects of entities. Users can give different opinions for different aspects of the same entity.

SENTIMENT CLASSIFICATION APPROACHES		SENTIMENT CLASSIFICATION APPROACHES	SENTIMENT CLASSIFICATION APPROACHES
Machine learning	Bayesian Networks Naïve Bayes Classification Maximum Entropy Neural Networks Support Vector Machine	Term presence and frequency Part of speech information Negations Opinion words and phrases	**ADVANTAGES:** the ability to adapt and create trained models for specific purposes and contexts **LIMITATIONS:** the low applicability to new data because it is necessary the availability of labeled data that could be costly or even prohibitive
Lexicon based	Dictionary based approach Novel Machine Learning Approach Corpus based approach Ensemble Approaches	Manual construction, Corpus-based Dictionary based	**ADVANTAGES:** wider term coverage **LIMITATIONS:** finite number of words in the lexicons and the assignation of a fixed snt orientation and score to words
Hybrid	Machine learning Lexicon based	Snt lexicon constructed using public resources for initial snt detection Snt words as features in machine learning method	**ADVANTAGES:** lexicon/learning symbiosis, the detection and measurement of snt at the concept level and the lesser sensitivity to changes in topic domain **LIMITATIONS:** noisy reviews

Table 5-1: Snt Classification

6. EXISTING MODEL

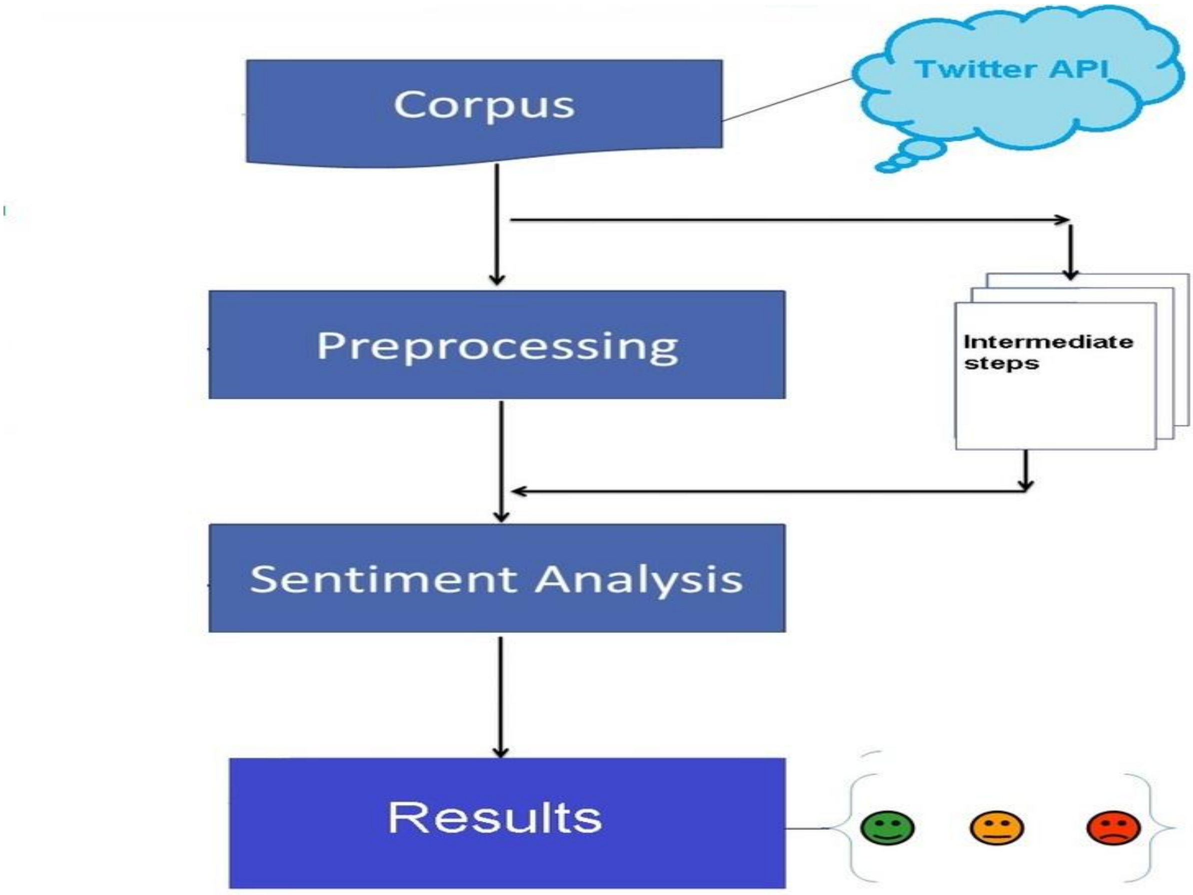
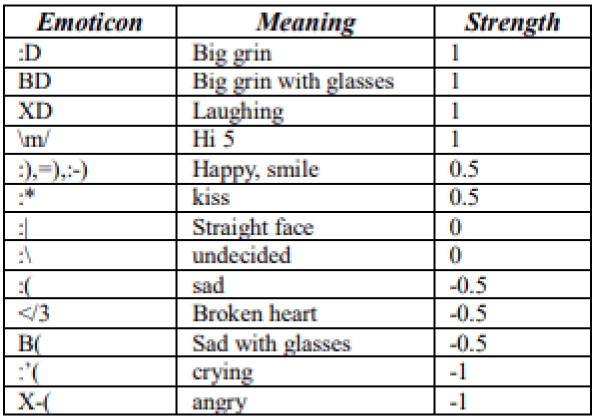


Figure 6-1: Existing Model Flow Chart of Twt Snt

Corpus: In linguistics, a **corpus** or **text corpus** is a large and structured set of texts (nowadays usually electronically stored and processed). In corpus linguistics, they are used to do statistical anlysis and hypothesis testing, checking occurrences or validating linguistic rules within a specific language territory.

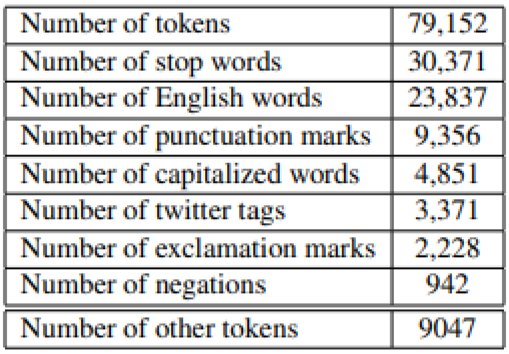
Preprocessing: Data processing involves Tokenization which is the process of splitting the tweets into individual words called tokens. Tokens can be split using whitespace or punctuation characters. It can be unigram or bigram depending on the classification model used. The bag-of-words model is one of the most extensively used model for classification. It is based on the fact of assuming text to be classified as a bag or collection of individual words with no link or interdependence. The simplest way to incorporate this model in our project is by using unigrams as features. It is just a collection of individual words in the text to be classified, so, we split each tweet using whitespace.[2]

Pre-processing takes place as follows: a) replace all the emoticons with a their snt polarity by looking up the emoticon dictionary, b) replace all URLs with a tag ||U||, c) replace targets (e.g. "@John") with tag ||T||, d) replace all negations (e.g. not, no, never, n't, cannot) by tag "NOT", and e) replace a sequence of repeated characters by three characters, for example, convert cooooooooool to coool. We do not replace the sequence by only two characters since we want to differentiate between the regular usage and emphasized usage of the word. Data preprocessing is an important tool for Data Mining (DM) algorithm. Twt data is an unstructured data set it is a collection of information from people entered his/her feelings, opinion, attitudes, products review, emotions, etc. This type of information is growing day by day in the internet. May companies want to analyze customers opinions which like the product and the services. The Proposed work to analyses the twt trending information and collect various different information form the users. It improves the accuracy of Twt data. This work easy to identify the people reaction or opinion. Additionally, improve the better performance for data preprocessing tool.



Emoticon	*Meaning*	*Strength*
:D	Big grin	1
BD	Big grin with glasses	1
XD	Laughing	1
\m/	Hi 5	1
:),=),:-)	Happy, smile	0.5
:*	kiss	0.5
:|	Straight face	0
:\	undecided	0
:(	sad	-0.5
</3	Broken heart	-0.5
B(	Sad with glasses	-0.5
:'(	crying	-1
X-(	angry	-1

Table 6-1: Emoticons



Number of tokens	79,152
Number of stop words	30,371
Number of English words	23,837
Number of punctuation marks	9,356
Number of capitalized words	4,851
Number of twitter tags	3,371
Number of exclamation marks	2,228
Number of negations	942
Number of other tokens	9047

Table 6-2: Classification

Intermediate Steps: Number of steps are involved in this to prepare data for anlysis which are as follows

- Training
- Classification

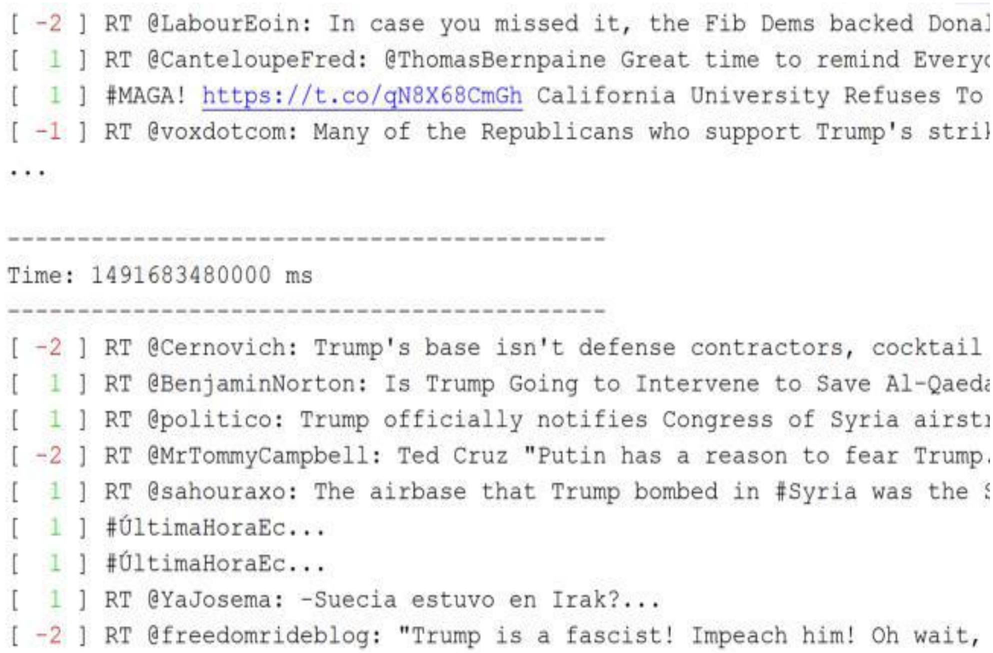
Snt Anlysis: For discovering the polarity, we used a simple algorithm of counting positive and negative words in a tweet. For both, positive and negative words, different lists were made. Next step is to compare every word in a tweet against both these list. If the current word matches a word in positive list, then a score of 1 is incremented and if a negative word is found then it is decremented. More positive words lead to higher snt score. However, Standford NLP can be used to predict accurate snt anlysis which provide complex algorithms to predict it.[4]

Analyze Tweets makes it simple to understand what people think (positively or negatively) about a certain keyword. You might be curious about what people think about a restaurant, hotel, or shopping mall. Or, maybe you want to better understand what the positive and negative topics regarding a certain political candidate. If people are tweeting about this keyword, then Analyze Tweets can help you categorize that conversation.

Analyze Tweets is especially valuable when there is a large amount of tweets around a subject.

An example of this would be to analyze the hashtag #MyCompanyBlackFridayDeal. Using Analyze Tweets, I might be able to see which deals customers are reacting positively or negatively to. Or, maybe they're just complaining about the lines at the store being too long because the deals are so good!

The output contains a list of tweets in real time along with their snt score on the left-hand side. The first tweet has score of -2 which is due to two negative keywords. Next two tweets are positive as they contain keywords like “good” and “great. Both these words are in the positive words list. It is to be noted that if a tweet has a score of 0, then it is ignored from final output. The problem with neutral tweets is that they serve no purpose as they don’t convey any snt towards the product.



```
[ -2 ] RT @LabourEoin: In case you missed it, the Fib Dems backed Donal
[  1 ] RT @CanteloupeFred: @ThomasBernpaine Great time to remind Every
[  1 ] #MAGA! https://t.co/qN8X68CmGh California University Refuses To
[ -1 ] RT @voxdotcom: Many of the Republicans who support Trump's stri
...

-------------------------------------------
Time: 1491683480000 ms
-------------------------------------------
[ -2 ] RT @Cernovich: Trump's base isn't defense contractors, cocktail
[  1 ] RT @BenjaminNorton: Is Trump Going to Intervene to Save Al-Qaed
[  1 ] RT @politico: Trump officially notifies Congress of Syria airst
[ -2 ] RT @MrTommyCampbell: Ted Cruz "Putin has a reason to fear Trump.
[  1 ] RT @sahouraxo: The airbase that Trump bombed in #Syria was the
[  1 ] #ÚltimaHoraEc...
[  1 ] #ÚltimaHoraEc...
[  1 ] RT @YaJosema: -Suecia estuvo en Irak?...
[ -2 ] RT @freedomrideblog: "Trump is a fascist! Impeach him! Oh wait,
```

Figure 6-2: Anlysis based on number of words

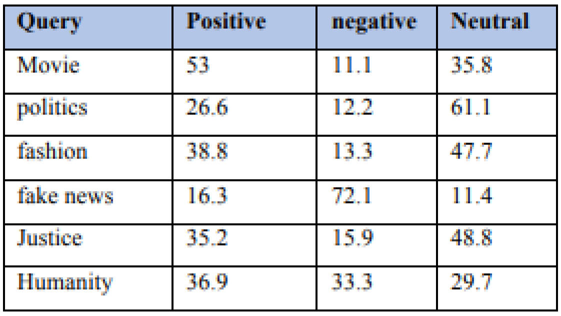
Let’s say a company has just launched a new product feature and you notice a sharp increase in mentions on Twt. However, receiving tons of mentions does not *necessarily* mean a good thing. Are customers tweeting more because they are expressing good things about this new product feature? Or, are customers actually complaining about the feature having lots of bugs? Performing Twt snt anlysis can be an excellent way to understand the tone of those mentions and obtain real-time insights on how users are perceiving your new product.

Thanks to snt anlysis, companies can understand the reputation of their brand. By analyzing social media posts, product reviews, customer feedback, or NPS responses (among other sources of unstructured business data), they can be aware of how their customers *feel* about their product. They can also track specific topics and get relevant insights on *how* people are talking about those topics.

Some more applications of snt anlysis[7]

1. Applications that use Reviews from Websites: Today Internet has a large collection of reviews and feedbacks on almost everything. This includes product reviews, feedbacks on political issues, comments about services, etc. Thus, there is a need for a snt anlysis system that can extract snts about a particular product or services. It will help us to automate in provision of feedback or rating for the given product, item, etc. This would serve the needs of both the users and the vendors.
2. Applications as a Sub-component Technology A snt predictor system can be helpful in recommender systems as well. The recommender system will not recommend items that receive a lot of negative feedback or fewer ratings. In online communication, we come across abusive language and other negative elements. These can be detected simply by identifying a highly negative snt and correspondingly taking action against it.
3. Applications in Business Intelligence It has been observed that people nowadays tend to look upon reviews of products which are available online before they buy them. And for many businesses, the online opinion decides the success or failure of their product. Thus, Snt Anlysis plays an important role in businesses. Businesses also wish to extract snt from the online reviews in order to improve their products and in turn their reputation and help in customer satisfaction.
4. Applications across Domains: Recent researches in sociology and other fields like medical, sports have also been benefitted by Snt Anlysis that show trends in human emotions especially on social media.
5. Applications in Smart Homes Smart homes are supposed to be the technology of the future. In future entire homes would be networked and people would be able to control any part of the home using a tablet device. Recently there has been lot of research going on Internet of Things (IoT). Snt Anlysis would also find its way in IoT. Like for example, based on the current snt or emotion of the user, the home could alter its ambiance to create a soothing and peaceful environment.

<u>Results:</u> Emotions, opinions and snts play an important role in all human life. Mining such opinions termed as snt anlysis [6]. Performing task of Snt anlysis & polarity classification is a challenging task. We did snt anlysis by using “Dictionary Based approach”. This approach uses a predefined dictionary of positive and Negative words. Senti Word net is a standard dictionary used by most researchers today for snt anlysis. Task of Polarity classification we mean the reviews collected are classified depending upon the emotions expressed as Positive, Negative and Neutral.



Query	Positive	negative	Neutral
Movie	53	11.1	35.8
politics	26.6	12.2	61.1
fashion	38.8	13.3	47.7
fake news	16.3	72.1	11.4
Justice	35.2	15.9	48.8
Humanity	36.9	33.3	29.7

Table 6-3: Snt Anlysis Results

7. PROPOSED MODEL

We categorize the analytics into 3 sections:

- Crawling, cleaning data and labeling un-structured data by using/mapping known English words from various sources.
- Applying Natural Language based classifiers used for text processing to train tweets and predict moods.
- Applying standard machine learning algorithms and deep learning to do multi-class mood classification

The objective is to highlight mechanisms for labeling tweets, and classifying and summarizing them from different viewpoints.

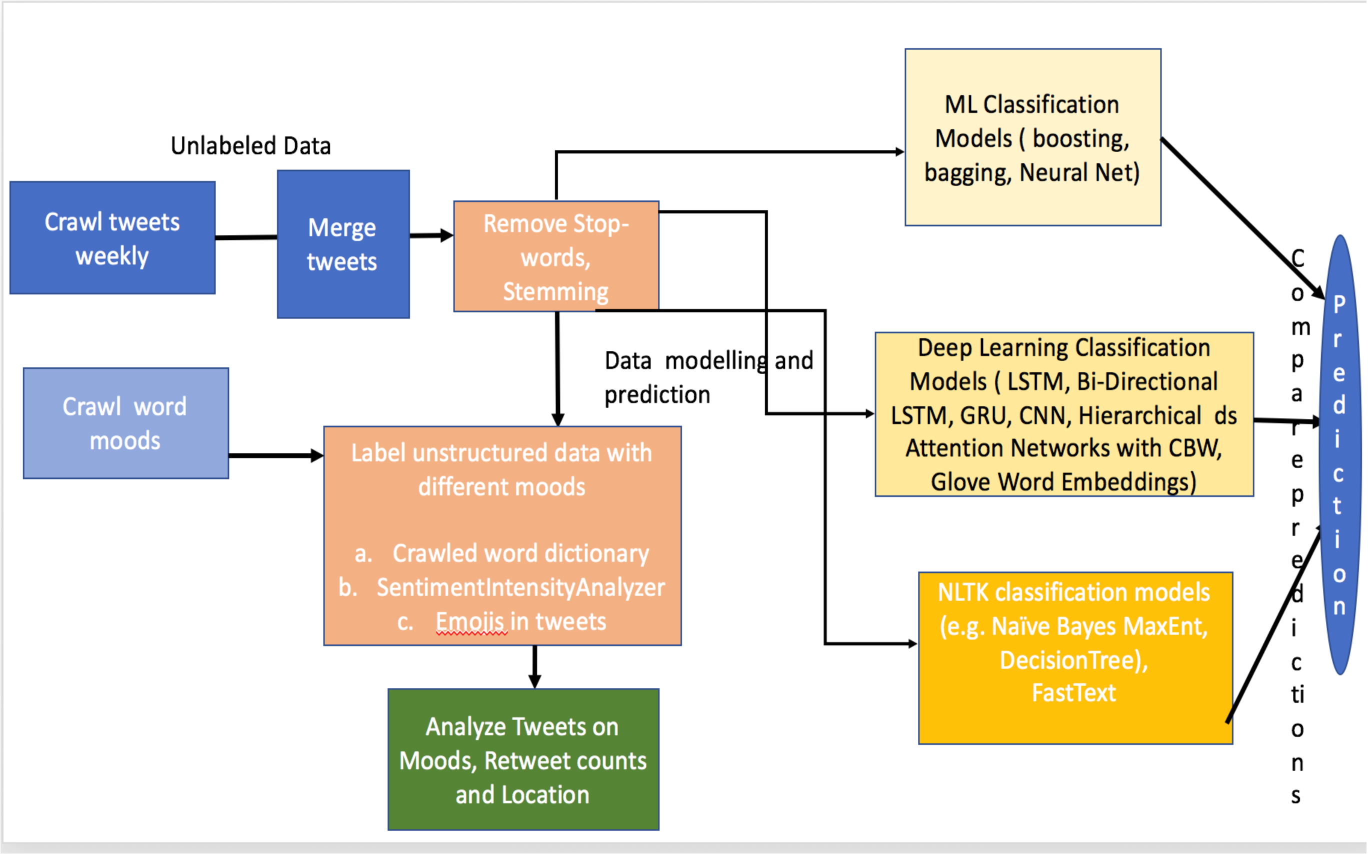

Unlabeled Data
Crawl tweets weekly
Merge tweets
Remove Stop-words, Stemming
ML Classification Models (boosting, bagging, Neural Net)
Data modelling and prediction
Crawl word moods
Label unstructured data with different moods
a. Crawled word dictionary
b. SentimentIntensityAnalyzer
c. Emojis in tweets
Deep Learning Classification Models (LSTM, Bi-Directional LSTM, GRU, CNN, Hierarchical ds Attention Networks with CBW, Glove Word Embeddings)
NLTK classification models (e.g. Naïve Bayes MaxEnt, DecisionTree), FastText
Analyze Tweets on Moods, Retweet counts and Location
Compare predictions
Prediction

Figure 7-1: Proposed Model of Twt Snt

- **Twt Data Crawler:** Users on Twt are generating about half billion tweets everyday. Some of these tweets are available to researchers and developers through Twt's public APIs. An open source library called *Tweepy* is used to collect different types of data from Twt and build your own Twt data crawler.

  Crawlers can be used to collects a user's profile information from Twt given the user's Twt ID, a user's social network information given the user's ID and the tweets using a set of specified keywords and a geo-location based criteria.

  Open Authentication (OAuth) is an open standard for authentication that is adopted by Twt to provide access to the protected information. OAuth provides a safer alternative to traditional authentication approaches using a three-way handshake[2].

  The authentication of API requests on Twt is done through OAuth. Note that Twt APIs can only be accessed by registered. In order to register for an application, the user need to have a Twt account. If the user already has one, they can just use it. If not, then user can go ahead and sign up one at Twt. After that, the user needs to bind his Twt account with the application you registered. Once the user finish the binding process, he will get the keys and tokens for his application.

- **ML Classification:** Machine learning is a field of study and is concerned with algorithms that learn from examples. Classification is a task that requires the use of machine learning algorithms that learn how to assign a class label to examples from the problem domain. An easy to understand example is classifying emails as “*spam*” or “*not spam*.”

  There are many different types of classification tasks that you may encounter in machine learning and specialized approaches to modeling that may be used for each.

  - Boosting

  - Bagging

- **NLTK Classification:** Detecting patterns is a central part of Natural Language Processing. Words ending in -*ed* tend to be past tense verbs. Frequent use of *will* is indicative of news text. These observable patterns — word structure and word frequency — happen to correlate with particular aspects of meaning, such as tense and topic. This classifier helps in such a way that which aspects of form to associate with which aspects of meaning.

    - Document Classification
    - Exploiting Context
    - Sequence Classification
    - Recognizing Textual Entailment
    - Scaling up to Large Datasets
    - Sentence Segmentation

8. IMPLEMENTATION

The project involves the usage of Apache Hadoop to analyze tweets. The objective is to find the polarity of the words (in tweets) retrieved.

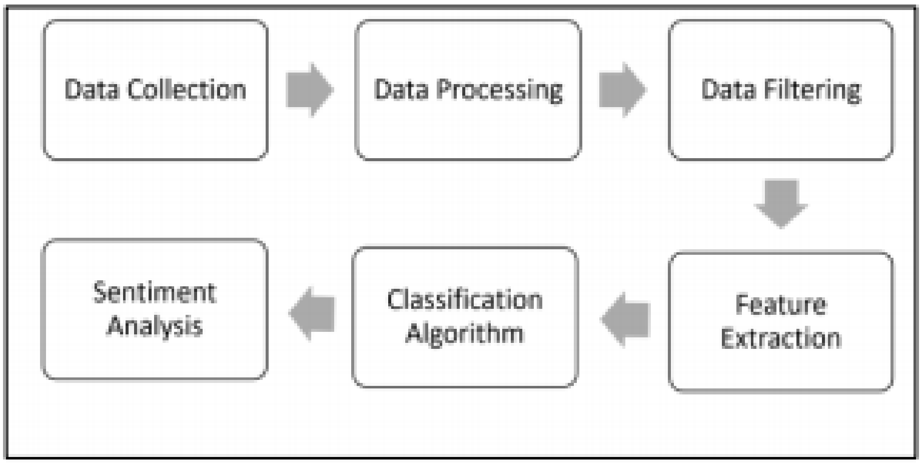
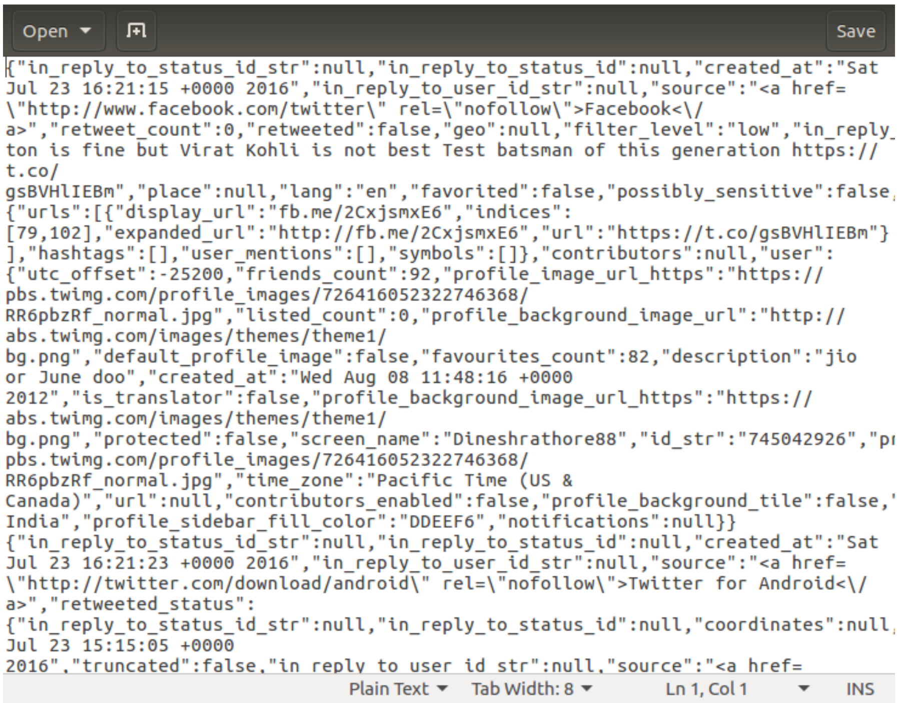


Figure 8-1: Framework for Twt Anlysis

Each step in the framework involves several sub-tasks:

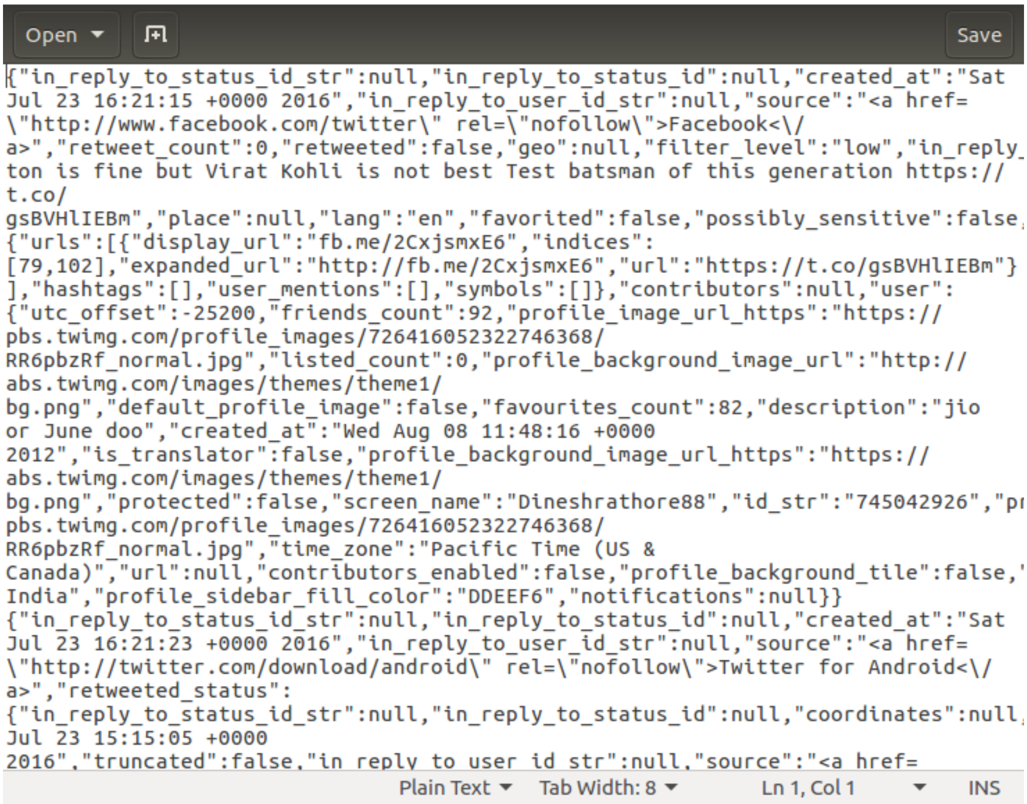
- **Data collection:** Data in the form of raw tweets is retrieved by using the Apache flume and twt API and token keys. The API requires us to register a developer account with Twt and fill in parameters such as consumerKey, consumerSecret, accessTokenaccess, and TokenSecret. This API allows to get all random tweets or filter data by using keywords. Filters supports to retrieve tweets which match a specific criterion defined by the developer. We used this to retrieve tweets related to specific keywords which are taken as input from users. Initially, we set at least set an application name and mode. We execute the program in local mode instead of cluster [3].



```
{"in_reply_to_status_id_str":null,"in_reply_to_status_id":null,"created_at":"Sat
Jul 23 16:21:15 +0000 2016","in_reply_to_user_id_str":null,"source":"<a href=
\"http://www.facebook.com/twitter\" rel=\"nofollow\">Facebook<\/
a>","retweet_count":0,"retweeted":false,"geo":null,"filter_level":"low","in_reply_
ton is fine but Virat Kohli is not best Test batsman of this generation https://
t.co/
gsBVHlIEBm","place":null,"lang":"en","favorited":false,"possibly_sensitive":false,
{"urls":[{"display_url":"fb.me/2CxjsmxE6","indices":
[79,102],"expanded_url":"http://fb.me/2CxjsmxE6","url":"https://t.co/gsBVHlIEBm"}
],"hashtags":[],"user_mentions":[],"symbols":[]},"contributors":null,"user":
{"utc_offset":-25200,"friends_count":92,"profile_image_url_https":"https://
pbs.twimg.com/profile_images/726416052322746368/
RR6pbzRf_normal.jpg","listed_count":0,"profile_background_image_url":"http://
abs.twimg.com/images/themes/theme1/
bg.png","default_profile_image":false,"favourites_count":82,"description":"jio
or June doo","created_at":"Wed Aug 08 11:48:16 +0000
2012","is_translator":false,"profile_background_image_url_https":"https://
abs.twimg.com/images/themes/theme1/
bg.png","protected":false,"screen_name":"Dineshrathore88","id_str":"745042926","pr
pbs.twimg.com/profile_images/726416052322746368/
RR6pbzRf_normal.jpg","time_zone":"Pacific Time (US &
Canada)","url":null,"contributors_enabled":false,"profile_background_tile":false,"
India","profile_sidebar_fill_color":"DDEEF6","notifications":null}}
{"in_reply_to_status_id_str":null,"in_reply_to_status_id":null,"created_at":"Sat
Jul 23 16:21:23 +0000 2016","in_reply_to_user_id_str":null,"source":"<a href=
\"http://twitter.com/download/android\" rel=\"nofollow\">Twitter for Android<\/
a>","retweeted_status":
{"in_reply_to_status_id_str":null,"in_reply_to_status_id":null,"coordinates":null,
Jul 23 15:15:05 +0000
2016","truncated":false,"in_reply_to_user_id_str":null,"source":"<a href=
```

Figure 8-2: Data Collected using Apache Flume from Twt

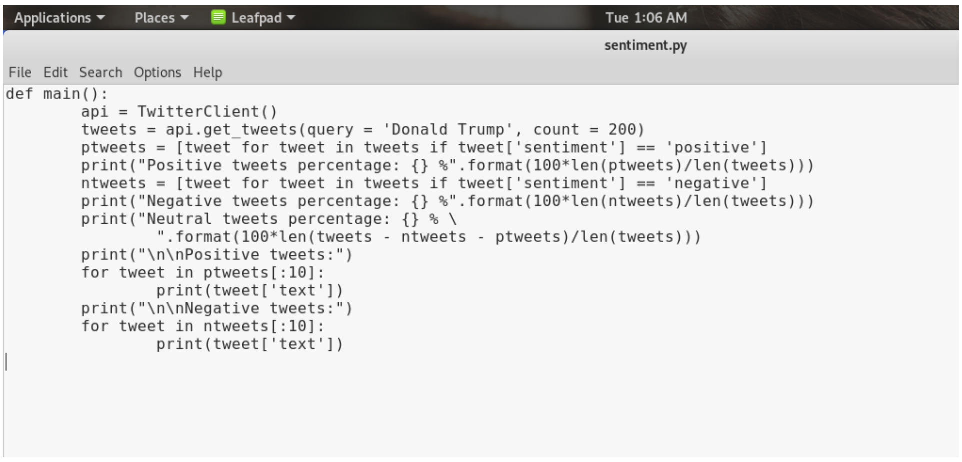
- **<u>Data Processing:</u>** Data processing involves Tokenization which is the process of splitting the tweets into individual words called tokens. Tokens can be split using whitespace or punctuation characters. It can be unigram or bigram depending on the classification model used. The bag-ofwords model is one of the most extensively used model for classification. It is based on the fact of assuming text to be classified as a bag or collection of individual words with no link or interdependence. The simplest way to incorporate this model in our project is by using unigrams as features. It is just a collection of individual words in the text to be classified, so, we split each tweet using whitespace. For example, the tweet “Met aziz today !!” is split from each whitespace as follows. {‘Met’,’ Aziz’,’ !!’ } The next step in data processing is normalization by conversion of tweet into lowercase. Tweets are normalized by converting it to lowercase which makes its comparison with an dictionary easier[5].



```
{"in_reply_to_status_id_str":null,"in_reply_to_status_id":null,"created_at":"Sat
Jul 23 16:21:15 +0000 2016","in_reply_to_user_id_str":null,"source":"<a href=
\"http://www.facebook.com/twitter\" rel=\"nofollow\">Facebook<\/
a>","retweet_count":0,"retweeted":false,"geo":null,"filter_level":"low","in_reply_
ton is fine but Virat Kohli is not best Test batsman of this generation https://
t.co/
gsBVHlIEBm","place":null,"lang":"en","favorited":false,"possibly_sensitive":false,
{"urls":[{"display_url":"fb.me/2CxjsmxE6","indices":
[79,102],"expanded_url":"http://fb.me/2CxjsmxE6","url":"https://t.co/gsBVHlIEBm"}
],"hashtags":[],"user_mentions":[],"symbols":[]},"contributors":null,"user":
{"utc_offset":-25200,"friends_count":92,"profile_image_url_https":"https://
pbs.twimg.com/profile_images/726416052322746368/
RR6pbzRf_normal.jpg","listed_count":0,"profile_background_image_url":"http://
abs.twimg.com/images/themes/theme1/
bg.png","default_profile_image":false,"favourites_count":82,"description":"jio
or June doo","created_at":"Wed Aug 08 11:48:16 +0000
2012","is_translator":false,"profile_background_image_url_https":"https://
abs.twimg.com/images/themes/theme1/
bg.png","protected":false,"screen_name":"Dineshrathore88","id_str":"745042926","pr
pbs.twimg.com/profile_images/726416052322746368/
RR6pbzRf_normal.jpg","time_zone":"Pacific Time (US &
Canada)","url":null,"contributors_enabled":false,"profile_background_tile":false,"
India","profile_sidebar_fill_color":"DDEEF6","notifications":null}}
{"in_reply_to_status_id_str":null,"in_reply_to_status_id":null,"created_at":"Sat
Jul 23 16:21:23 +0000 2016","in_reply_to_user_id_str":null,"source":"<a href=
\"http://twitter.com/download/android\" rel=\"nofollow\">Twitter for Android<\/
a>","retweeted_status":
{"in_reply_to_status_id_str":null,"in_reply_to_status_id":null,"coordinates":null,
Jul 23 15:15:05 +0000
2016","truncated":false,"in_reply_to_user_id_str":null,"source":"<a href=
```

Figure 8-3: Cleaned Tweets

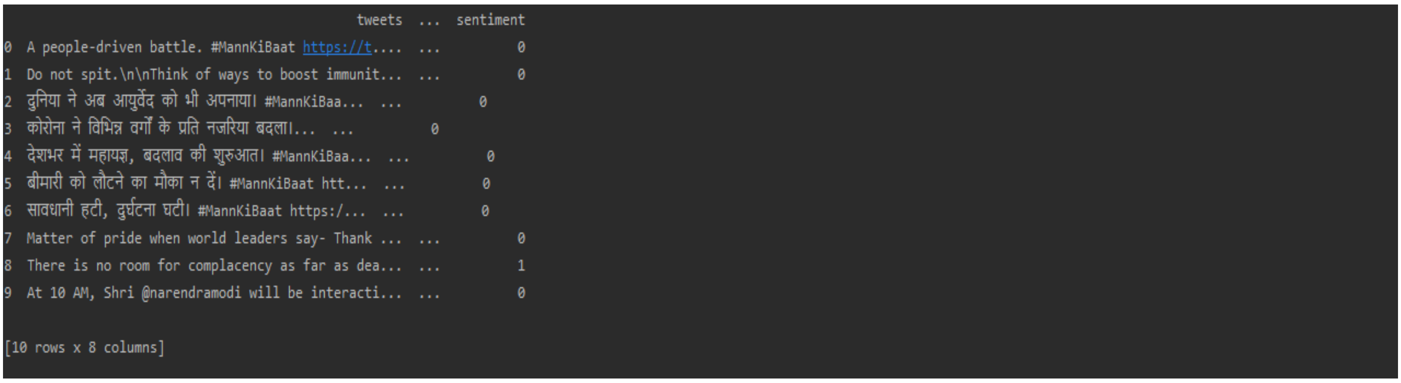
- **<u>Data Filtering:</u>** A tweet acquired after data processing still has a portion of raw information in it which we may or may not find useful for our application. Thus, these tweets are further filtered by removing stop words, numbers and punctuations. Stop words: For example, tweets contain stop words which are extremely common words like “is”, “am”, “are” and holds no additional information[6]. These words serve no purpose and this feature is implemented using a list stored in stop file.dat. We then compare each word in a tweet with this list and delete the words matching the stop list.



```
def main():
        api = TwitterClient()
        tweets = api.get_tweets(query = 'Donald Trump', count = 200)
        ptweets = [tweet for tweet in tweets if tweet['sentiment'] == 'positive']
        print("Positive tweets percentage: {} %".format(100*len(ptweets)/len(tweets)))
        ntweets = [tweet for tweet in tweets if tweet['sentiment'] == 'negative']
        print("Negative tweets percentage: {} %".format(100*len(ntweets)/len(tweets)))
        print("Neutral tweets percentage: {} % \
                ".format(100*len(tweets - ntweets - ptweets)/len(tweets)))
        print("\n\nPositive tweets:")
        for tweet in ptweets[:10]:
                print(tweet['text'])
        print("\n\nNegative tweets:")
        for tweet in ntweets[:10]:
                print(tweet['text'])
```

Figure 8-4: Snt Anlysis Code

- **<u>Snt Anlysis</u>:** Snt anlysis is done by using custom algorithm which finds polarity as below. Finding polarity: For discovering the polarity, we used a simple algorithm of counting positive and negative words in a tweet[3]. For both, positive and negative words, different lists were made. Next step is to compare every word in a tweet against both these list. If the current word matches a word in positive list, then a score of 1 is incremented and if a negative word is found then it is decremented. More positive words lead to higher snt scores.



```
                                         tweets  ...  sentiment
0  A people-driven battle. #MannKiBaat https://t....  ...          0
1  Do not spit.\n\nThink of ways to boost immunit...  ...          0
2  दुनिया ने अब आयुर्वेद को भी अपनाया। #MannKiBaa...  ...          0
3  कोरोना ने विभिन्न वर्गों के प्रति नजरिया बदला।...  ...          0
4  देशभर में महायज्ञ, बदलाव की शुरुआत। #MannKiBaa...  ...          0
5  बीमारी को लौटने का मौका न दें। #MannKiBaat htt...  ...          0
6  सावधानी हटी, दुर्घटना घटी। #MannKiBaat https:/...  ...          0
7  Matter of pride when world leaders say- Thank ...  ...          0
8  There is no room for complacency as far as dea...  ...          1
9  At 10 AM, Shri @narendramodi will be interacti...  ...          0

[10 rows x 8 columns]
```

Figure 8-5: Polarity of Tweets

Class Diagram

Class diagram is a static diagram. It represents the static view of an application. Class diagram is not only used for visualizing, describing, and documenting different aspects of a system but also for constructing executable code of the software application.

Class diagram describes the attributes and operations of a class and also the constraints imposed on the system. The class diagrams are widely used in the modeling of objectoriented systems because they are the only UML diagrams, which can be mapped directly with object-oriented languages.

Class diagram shows a collection of classes, interfaces, associations, collaborations, and constraints. It is also known as a structural diagram.

Purpose of Class Diagrams

The purpose of class diagram is to model the static view of an application. Class diagrams are the only diagrams which can be directly mapped with object-oriented languages and thus widely used at the time of construction.

UML diagrams like activity diagram, sequence diagram can only give the sequence flow of the application, however class diagram is a bit different. It is the most popular UML diagram in the coder community.

The purpose of the class diagram can be summarized as −

- Anlysis and design of the static view of an application.
- Describe responsibilities of a system.
- Base for component and deployment diagrams.
- Forward and reverse engineering.

How to Draw a Class Diagram?

Class diagrams are the most popular UML diagrams used for construction of software applications. It is very important to learn the drawing procedure of class diagram.

Class diagrams have a lot of properties to consider while drawing but here the diagram will be considered from a top level view.

Class diagram is basically a graphical representation of the static view of the system and represents different aspects of the application. A collection of class diagrams represent the whole system.

The following points should be remembered while drawing a class diagram −

- The name of the class diagram should be meaningful to describe the aspect of the system.
- Each element and their relationships should be identified in advance.
- Responsibility (attributes and methods) of each class should be clearly identified
- For each class, minimum number of properties should be specified, as unnecessary properties will make the diagram complicated.
- Use notes whenever required to describe some aspect of the diagram. At the end of the drawing it should be understandable to the developer/coder.
- Finally, before making the final version, the diagram should be drawn on plain paper and reworked as many times as possible to make it correct.

Where to Use Class Diagrams?

Class diagram is a static diagram and it is used to model the static view of a system. The static view describes the vocabulary of the system.

Class diagram is also considered as the foundation for component and deployment diagrams. Class diagrams are not only used to visualize the static view of the system but they are also used to construct the executable code for forward and reverse engineering of any system.

Generally, UML diagrams are not directly mapped with any object-oriented programming languages but the class diagram is an exception.

Class diagram clearly shows the mapping with object-oriented languages such as Java, C++, etc. From practical experience, class diagram is generally used for construction purpose.

In a nutshell it can be said, class diagrams are used for −

- Describing the static view of the system.
- Showing the collaboration among the elements of the static view.
- Describing the functionalities performed by the system.
- Construction of software applications using object oriented languages.

Perspectives of Class Diagram

The choice of perspective depends on how far along you are in the development process. During the formulation of a domain model, for example, you would seldom move past the conceptual perspective. Anlysis models will typically feature a mix of conceptual and specification perspectives. Design model development will typically start with heavy emphasis on the specification perspective, and evolve into the implementation perspective.

A diagram can be interpreted from various perspectives:

- Conceptual: represents the concepts in the domain
- Specification: focus is on the interfaces of Abstract Data Type (ADTs) in the software
- Implementation: describes how classes will implement their interfaces

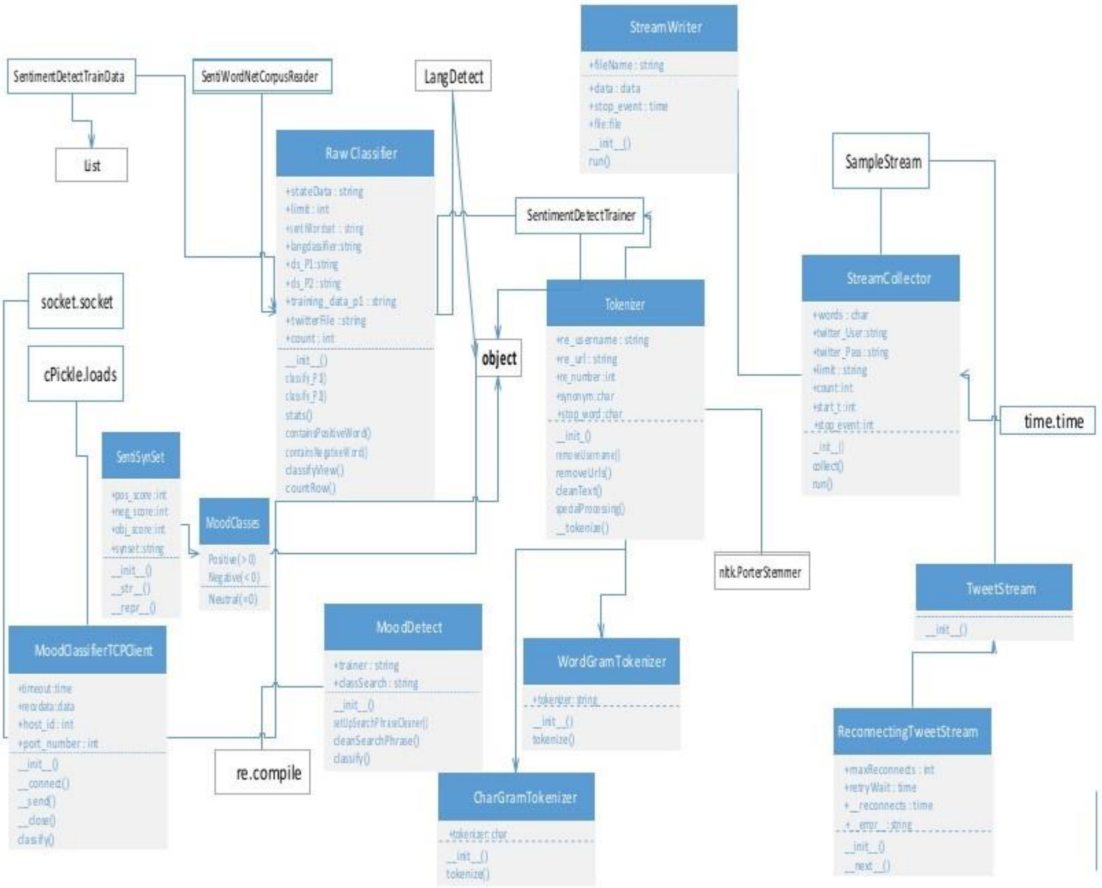


Figure 8-6: Class Diagram of Proposed Model

Activity Diagram

Activity diagram is defined as a UML diagram that focuses on the execution and flow of the behavior of a system instead of implementation. It is also called **object-oriented flowchart**. Activity diagrams consist of activities that are made up of actions which apply to behavioral modeling technology.

Components of Activity Diagram

Activities

It is a behaviour that is divided into one or more actions. Activities are a network of nodes connected by edges. There can be action nodes, control nodes, or object nodes. Action nodes represent some action. Control nodes represent the control flow of an activity. Object nodes are used to describe objects used inside an activity. Edges are used to show a path or a flow of execution. Activities start at an initial node and terminate at a final node.

Activity partition/swim lane

An activity partition or a swim lane is a high-level grouping of a set of related actions. A single partition can refer to many things, such as classes, use cases, components, or interfaces.

If a partition cannot be shown clearly, then the name of a partition is written on top of the name of an activity.

Fork and Join nodes

Using a fork and join nodes, concurrent flows within an activity can be generated. A fork node has one incoming edge and numerous outgoing edges. It is similar to one too many decision parameters. When data arrives at an incoming edge, it is duplicated and split across numerous outgoing edges simultaneously. A single incoming flow is divided into multiple parallel flows.

A join node is opposite of a fork node as It has many incoming edges and a single outgoing edge. It performs logical AND operation on all the incoming edges. This helps you to synchronize the input flow across a single output edge.

Pins

An activity diagram that has a lot of flows gets very complicated and messy.

Pins are used to clearing up the things. It provides a way to manage the execution flow of activity by sorting all the flows and cleaning up messy thins. It is an object node that represents one input to or an output from an action.

Both input and output pins have precisely one edge.

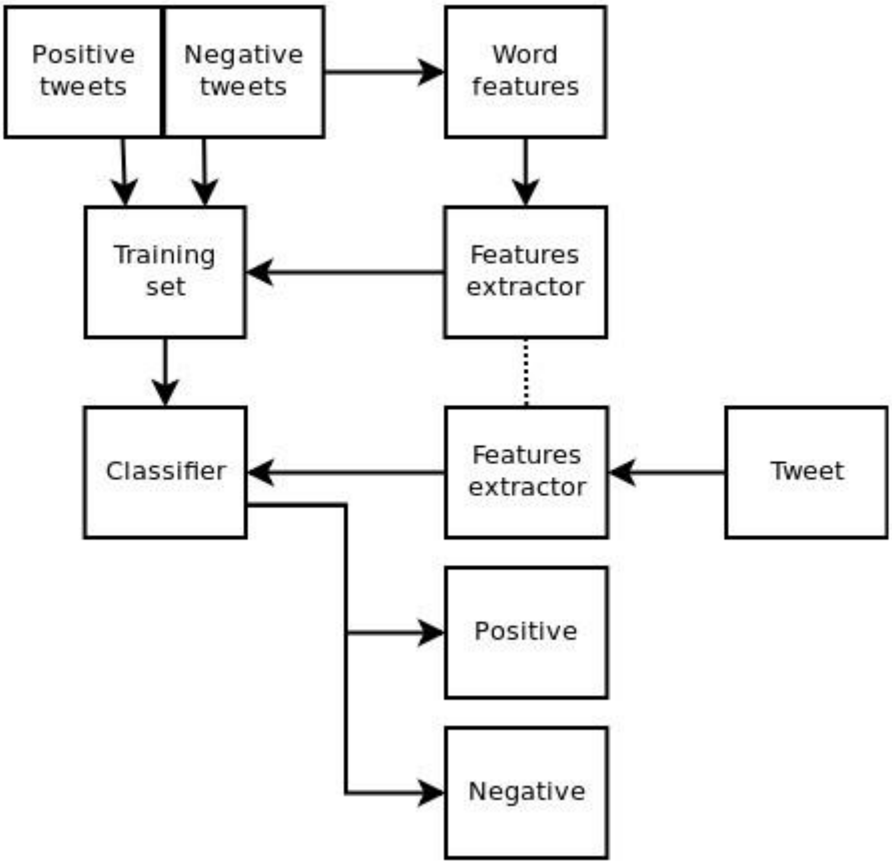


Figure 8-7: Activity Diagram of Proposed Model

Data Flow Diagram

A picture is worth a thousand words. A Data Flow Diagram (DFD) is a traditional way to visualize the information flows within a system. A neat and clear DFD can depict a good amount of the system requirements graphically. It can be manual, automated, or a combination of both.

It shows how information enters and leaves the system, what changes the information and where information is stored. The purpose of a DFD is to show the scope and boundaries of a system as a whole. It may be used as a communications tool between a systems analyst and any person who plays a part in the system that acts as the starting point for redesigning a system.

It is usually beginning with a context diagram as level 0 of the DFD diagram, a simple representation of the whole system. To elaborate further from that, we drill down to a level 1 diagram with lower-level functions decomposed from the major functions of the system. This could continue to evolve to become a level 2 diagram when further anlysis is required. Progression to levels 3, 4 and so on is possible but anything beyond level 3 is not very common. Please bear in mind that the level of detail for decomposing a particular function depending on the complexity that function.

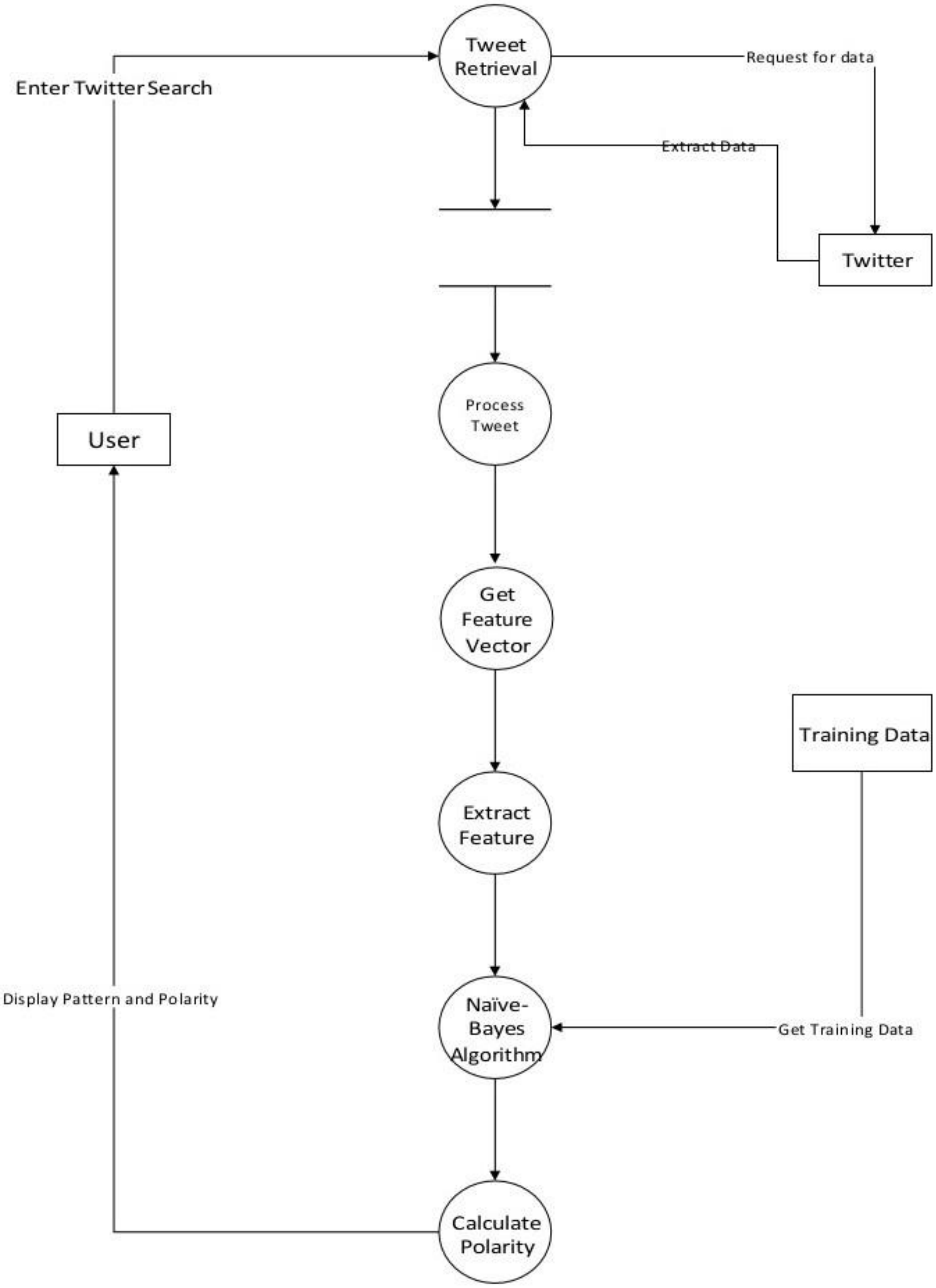


Figure 8-8: Data Flow Diagram of Proposed Model

Sequence Diagram

UML sequence diagrams model the flow of logic within your system in a visual manner, enabling you both to document and validate your logic, and are commonly used for both anlysis and design purposes. Sequence diagrams are the most popular UML artifact for dynamic modeling, which focuses on identifying the behavior within your system. Other dynamic modeling techniques include activity diagramming, communication diagramming, timing diagramming, and interaction overview diagramming. Sequence diagrams, along with class diagrams and physical data models are in my opinion the most important design-level models for modern business application development.

Sequence diagrams are typically used to model:

1. **Usage scenarios**. A usage scenario is a description of a potential way your system is used. The logic of a usage scenario may be part of a use case, perhaps an alternate course. It may also be one entire pass through a use case, such as the logic described by the basic course of action or a portion of the basic course of action, plus one or more alternate scenarios. The logic of a usage scenario may also be a pass through the logic contained in several use cases. For example, a student enrolls in the university, and then immediately enrolls in three seminars.

2. **The logic of methods**. Sequence diagrams can be used to explore the logic of a complex operation, function, or procedure. One way to think of sequence diagrams, particularly highly detailed diagrams, is as visual object code.

3. **The logic of services**. A service is effectively a high-level method, often one that can be invoked by a wide variety of clients. This includes web-services as well as business transactions implemented by a variety of technologies such as CICS/COBOL or CORBA-compliant object request brokers (ORBs).

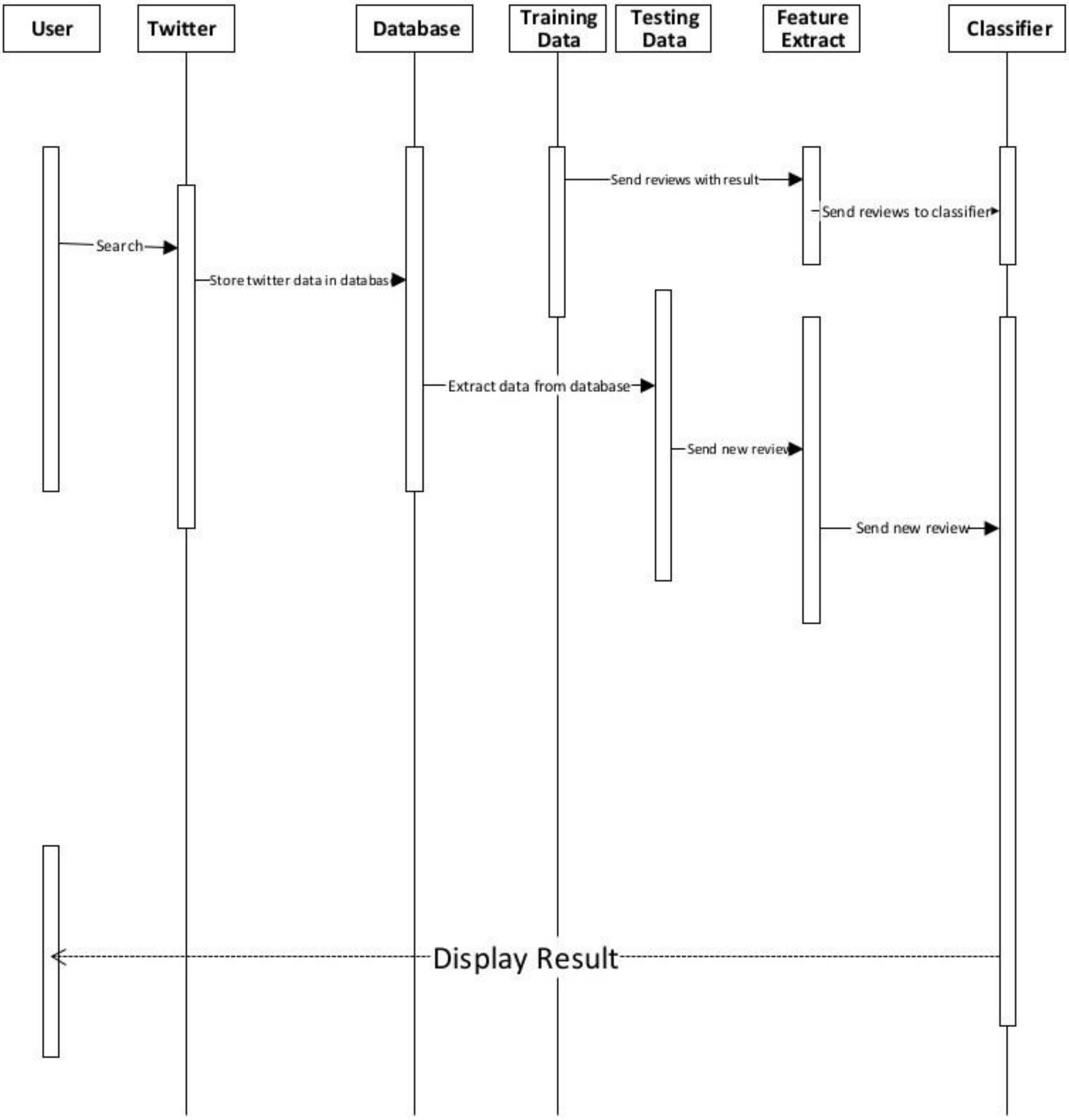


Figure 8-9: Sequence Diagram of Proposed Model

Use Case Diagram

Use case diagrams are usually referred to as behavior diagrams used to describe a set of actions (use cases) that some system or systems (subject) should or can perform in collaboration with one or more external users of the system (actors). Each use case should provide some observable and valuable result to the actors or other stakeholders of the system.

Use case diagrams are in fact twofold - they are both behavior diagrams, because they describe behavior of the system, and they are also structure diagrams - as a special case of class diagrams where classifiers are restricted to be either actors or use cases related to each other with associations.

Purpose of Use Case Diagrams

The purpose of use case diagram is to capture the dynamic aspect of a system. However, this definition is too generic to describe the purpose, as other four diagrams (activity, sequence, collaboration, and Statechart) also have the same purpose. We will look into some specific purpose, which will distinguish it from other four diagrams.

Use case diagrams are used to gather the requirements of a system including internal and external influences. These requirements are mostly design requirements. Hence, when a system is analyzed to gather its functionalities, use cases are prepared and actors are identified.

When the initial task is complete, use case diagrams are modelled to present the outside view.

In brief, the purposes of use case diagrams can be said to be as follows −

- Used to gather the requirements of a system.
- Used to get an outside view of a system.
- Identify the external and internal factors influencing the system.
- Show the interaction among the requirements are actors.

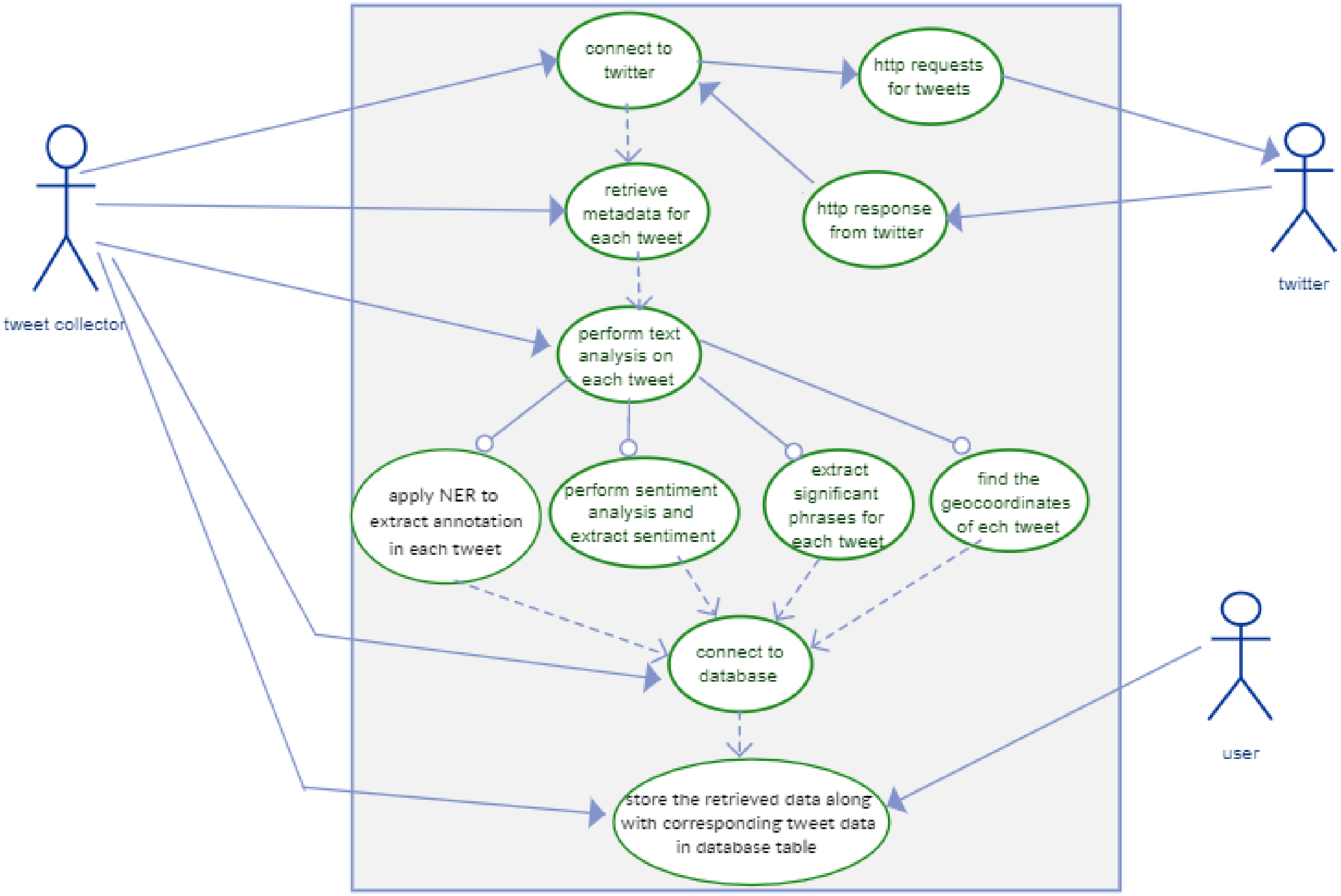

connect to twitter
http requests for tweets
retrieve metadata for each tweet
http response from twitter
perform text analysis on each tweet
apply NER to extract annotation in each tweet
perform sentiment analysis and extract sentiment
extract significant phrases for each tweet
find the geocoordinates of ech tweet
connect to database
store the retrieved data along with corresponding tweet data in database table
tweet collector
twitter
user

Figure 8-10: Use Case Diagram of Proposed Model

System Flow Diagram

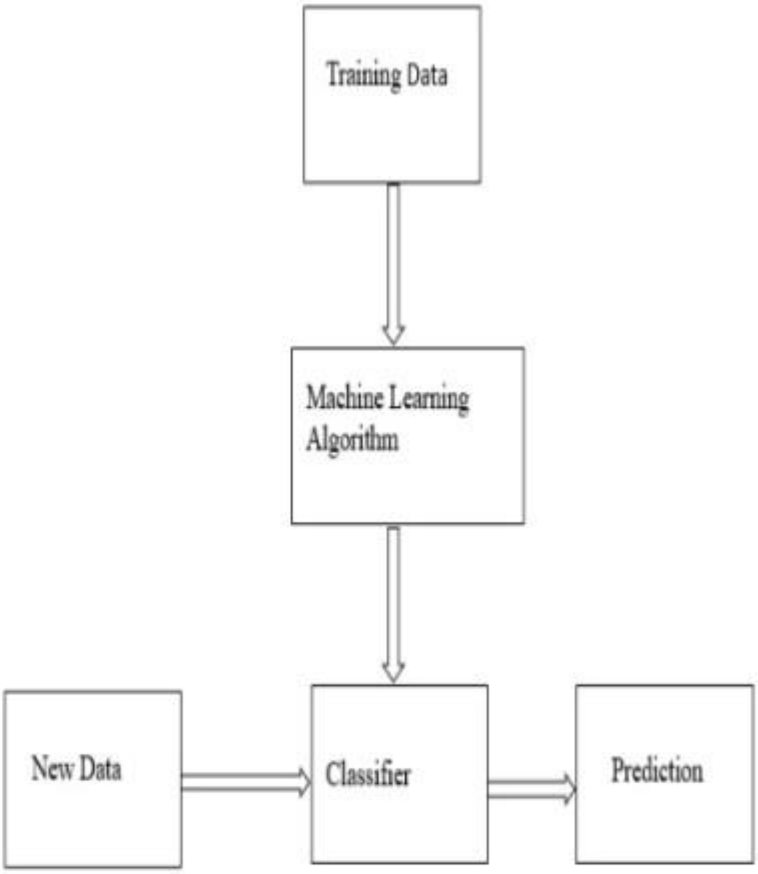
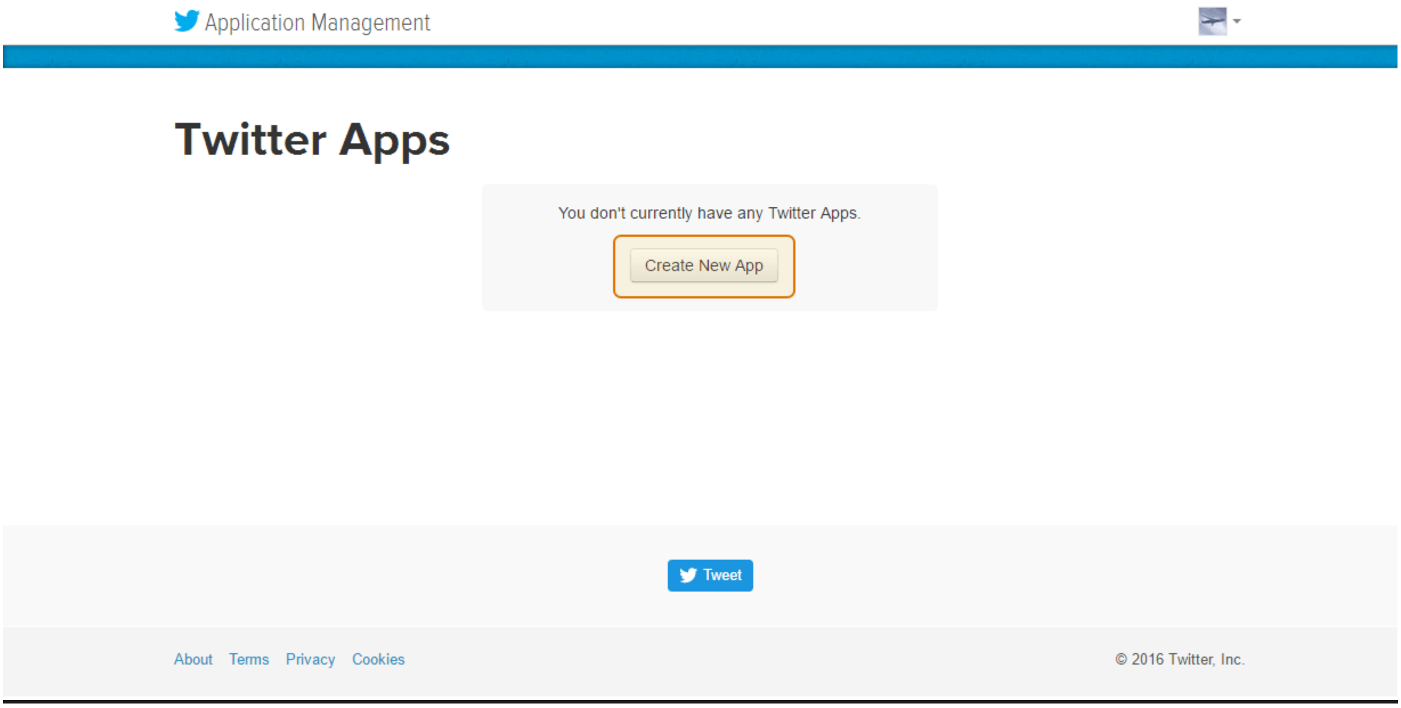


Figure 8-11: System Flow Chart of Proposed Model

9. OUTPUT / SCREENSHOTS

a) Creating Developer Account:



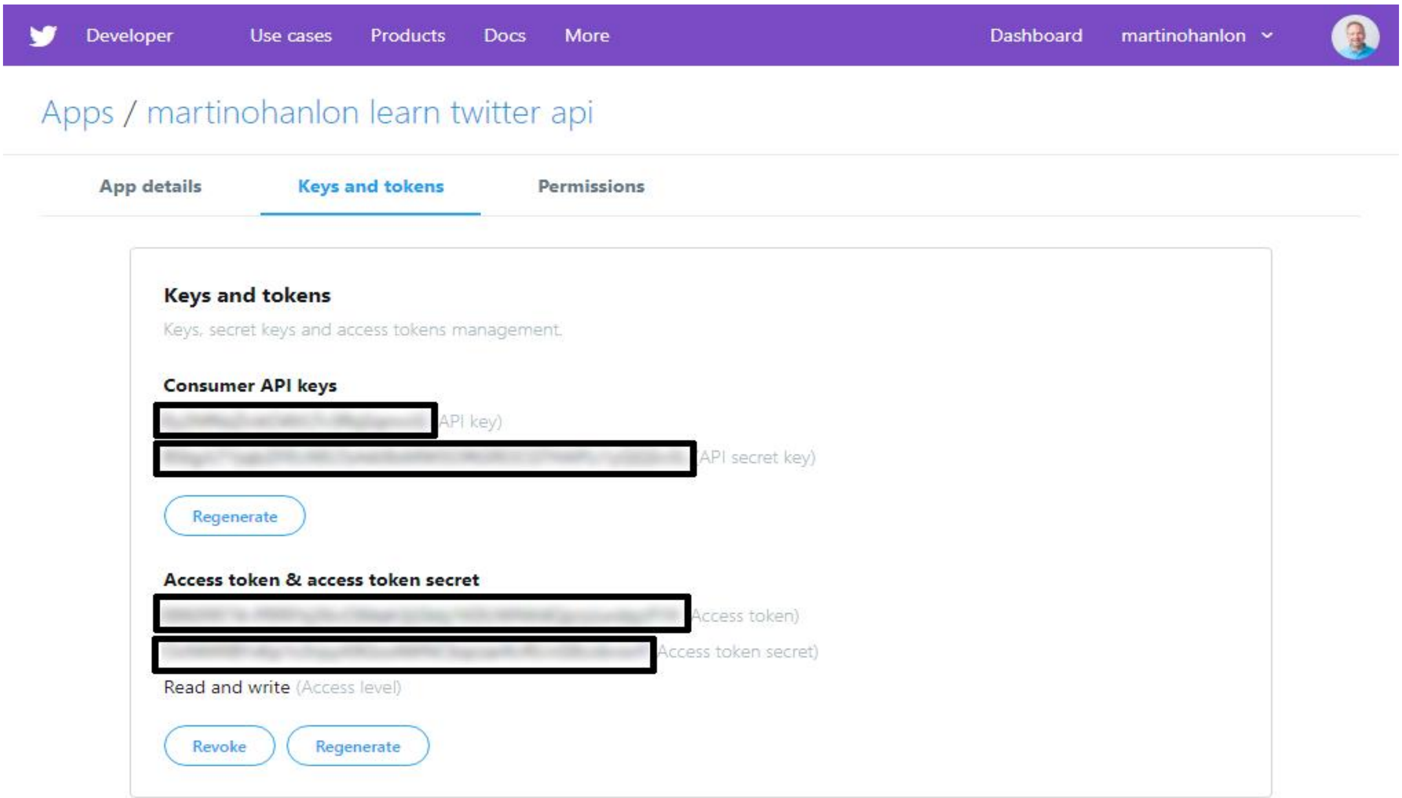
Twitter Apps

You don't currently have any Twitter Apps.

© 2016 Twitter, Inc.

Figure 9-1: Creating New App Through Developer Account

b) Getting Token Keys:



Apps / martinohanlon learn twitter api

Keys and tokens

Keys, secret keys and access tokens management.

Consumer API keys

[redacted] (API key)

[redacted] (API secret key)

Access token & access token secret

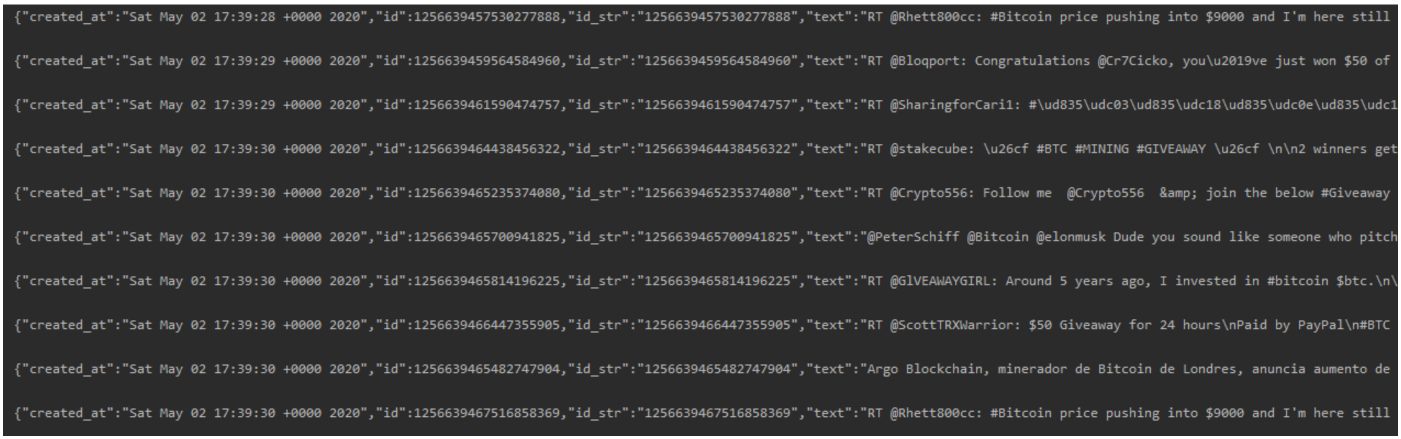
[redacted] (Access token)

[redacted] (Access token secret)

Read and write (Access level)

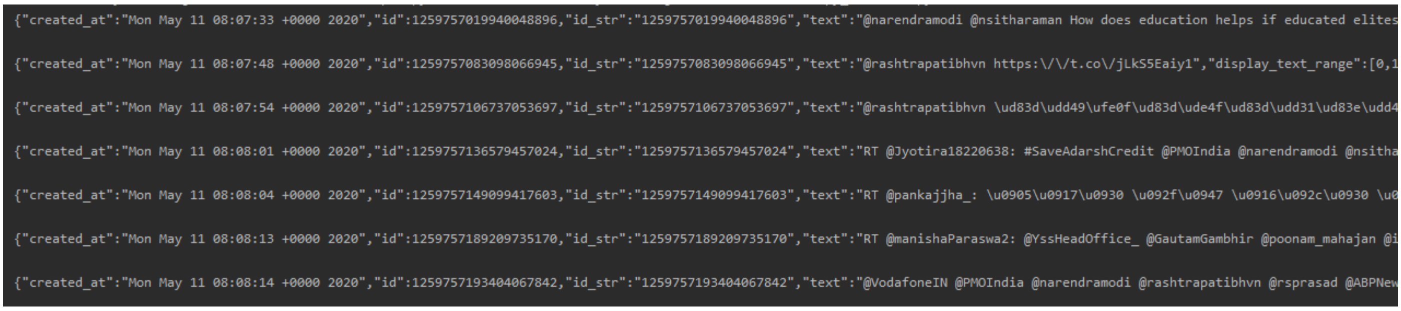
Figure 9-2: Getting the Access Keys to Developer Account

c) Streaming of Tweets:



```
{"created_at":"Sat May 02 17:39:28 +0000 2020","id":1256639457530277888,"id_str":"1256639457530277888","text":"RT @Rhett800cc: #Bitcoin price pushing into $9000 and I'm here still

{"created_at":"Sat May 02 17:39:29 +0000 2020","id":1256639459564584960,"id_str":"1256639459564584960","text":"RT @Bloqport: Congratulations @Cr7Cicko, you\u2019ve just won $50 of

{"created_at":"Sat May 02 17:39:29 +0000 2020","id":1256639461590474757,"id_str":"1256639461590474757","text":"RT @SharingforCari1: #\ud835\udc03\ud835\udc18\ud835\udc0e\ud835\udc1

{"created_at":"Sat May 02 17:39:30 +0000 2020","id":1256639464438456322,"id_str":"1256639464438456322","text":"RT @stakecube: \u26cf #BTC #MINING #GIVEAWAY \u26cf \n\n2 winners get

{"created_at":"Sat May 02 17:39:30 +0000 2020","id":1256639465235374080,"id_str":"1256639465235374080","text":"RT @Crypto556: Follow me  @Crypto556  &amp; join the below #Giveaway

{"created_at":"Sat May 02 17:39:30 +0000 2020","id":1256639465700941825,"id_str":"1256639465700941825","text":"@PeterSchiff @Bitcoin @elonmusk Dude you sound like someone who pitch

{"created_at":"Sat May 02 17:39:30 +0000 2020","id":1256639465814196225,"id_str":"1256639465814196225","text":"RT @GlVEAWAYGIRL: Around 5 years ago, I invested in #bitcoin $btc.\n\

{"created_at":"Sat May 02 17:39:30 +0000 2020","id":1256639466447355905,"id_str":"1256639466447355905","text":"RT @ScottTRXWarrior: $50 Giveaway for 24 hours\nPaid by PayPal\n#BTC

{"created_at":"Sat May 02 17:39:30 +0000 2020","id":1256639465482747904,"id_str":"1256639465482747904","text":"Argo Blockchain, minerador de Bitcoin de Londres, anuncia aumento de

{"created_at":"Sat May 02 17:39:30 +0000 2020","id":1256639467516858369,"id_str":"1256639467516858369","text":"RT @Rhett800cc: #Bitcoin price pushing into $9000 and I'm here still
```

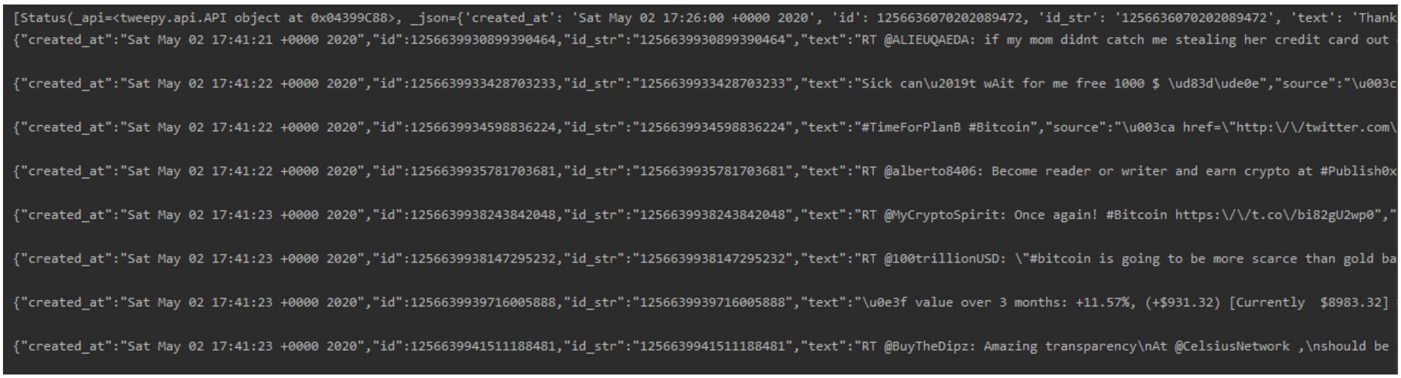
Figure 9-3: Tweets Streaming of @bitcoin on the Output Screen



```
{"created_at":"Mon May 11 08:07:33 +0000 2020","id":1259757019940048896,"id_str":"1259757019940048896","text":"@narendramodi @nsitharaman How does education helps if educated elites

{"created_at":"Mon May 11 08:07:48 +0000 2020","id":1259757083098066945,"id_str":"1259757083098066945","text":"@rashtrapatibhvn https:\/\/t.co\/jLkS5Eaiy1","display_text_range":[0,1

{"created_at":"Mon May 11 08:07:54 +0000 2020","id":1259757106737053697,"id_str":"1259757106737053697","text":"@rashtrapatibhvn \ud83d\udd49\ufe0f\ud83d\ude4f\ud83d\udd31\ud83e\udd4

{"created_at":"Mon May 11 08:08:01 +0000 2020","id":1259757136579457024,"id_str":"1259757136579457024","text":"RT @Jyotira18220638: #SaveAdarshCredit @PMOIndia @narendramodi @nsitha

{"created_at":"Mon May 11 08:08:04 +0000 2020","id":1259757149099417603,"id_str":"1259757149099417603","text":"RT @pankajjha_: \u0905\u0917\u0930 \u092f\u0947 \u0916\u092c\u0930 \u0

{"created_at":"Mon May 11 08:08:13 +0000 2020","id":1259757189209735170,"id_str":"1259757189209735170","text":"RT @manishaParaswa2: @YssHeadOffice_ @GautamGambhir @poonam_mahajan @i

{"created_at":"Mon May 11 08:08:14 +0000 2020","id":1259757193404067842,"id_str":"1259757193404067842","text":"@VodafoneIN @PMOIndia @narendramodi @rashtrapatibhvn @rsprasad @ABPNew
```

Figure 9-4: Tweets Streaming of @rashtrapatibhvn on the Output Screen

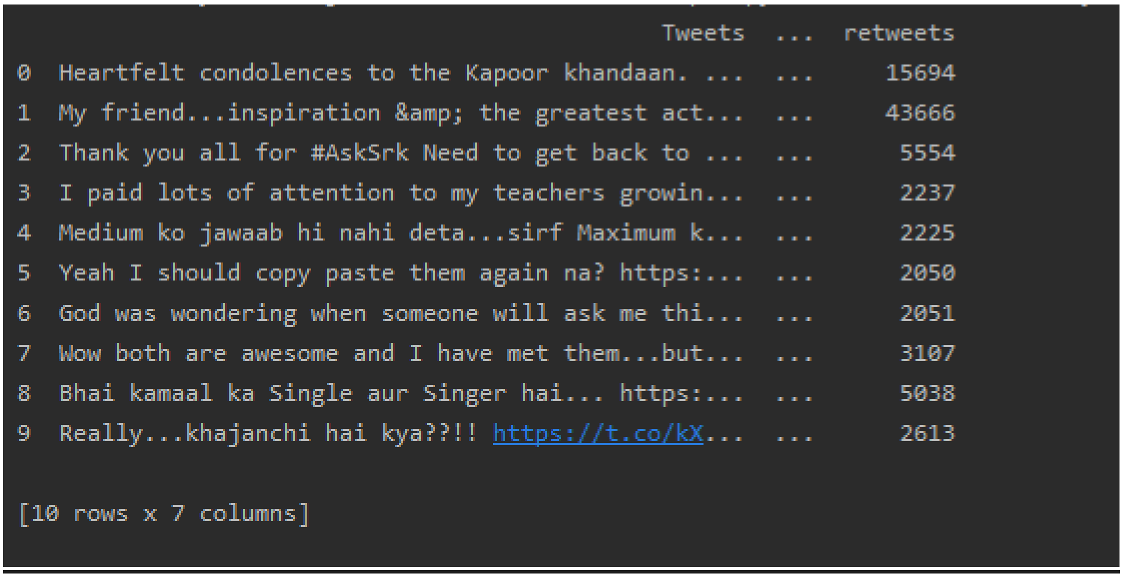
d) Pagination:



```
[Status(_api=<tweepy.api.API object at 0x04399C88>, _json={'created_at': 'Sat May 02 17:26:00 +0000 2020', 'id': 1256636070202089472, 'id_str': '1256636070202089472', 'text': 'Thank
{"created_at":"Sat May 02 17:41:21 +0000 2020","id":1256639930899390464,"id_str":"1256639930899390464","text":"RT @ALIEUQAEDA: if my mom didnt catch me stealing her credit card out

{"created_at":"Sat May 02 17:41:22 +0000 2020","id":1256639933428703233,"id_str":"1256639933428703233","text":"Sick can\u2019t wAit for me free 1000 $ \ud83d\ude0e","source":"\u003c

{"created_at":"Sat May 02 17:41:22 +0000 2020","id":1256639934598836224,"id_str":"1256639934598836224","text":"#TimeForPlanB #Bitcoin","source":"\u003ca href=\"http:\/\/twitter.com\

{"created_at":"Sat May 02 17:41:22 +0000 2020","id":1256639935781703681,"id_str":"1256639935781703681","text":"RT @alberto8406: Become reader or writer and earn crypto at #Publish0x

{"created_at":"Sat May 02 17:41:23 +0000 2020","id":1256639938243842048,"id_str":"1256639938243842048","text":"RT @MyCryptoSpirit: Once again! #Bitcoin https:\/\/t.co\/bi82gU2wp0","

{"created_at":"Sat May 02 17:41:23 +0000 2020","id":1256639938147295232,"id_str":"1256639938147295232","text":"RT @100trillionUSD: \"#bitcoin is going to be more scarce than gold ba

{"created_at":"Sat May 02 17:41:23 +0000 2020","id":1256639939716005888,"id_str":"1256639939716005888","text":"\u0e3f value over 3 months: +11.57%, (+$931.32) [Currently  $8983.32]

{"created_at":"Sat May 02 17:41:23 +0000 2020","id":1256639941511188481,"id_str":"1256639941511188481","text":"RT @BuyTheDipz: Amazing transparency\nAt @CelsiusNetwork ,\nshould be
```

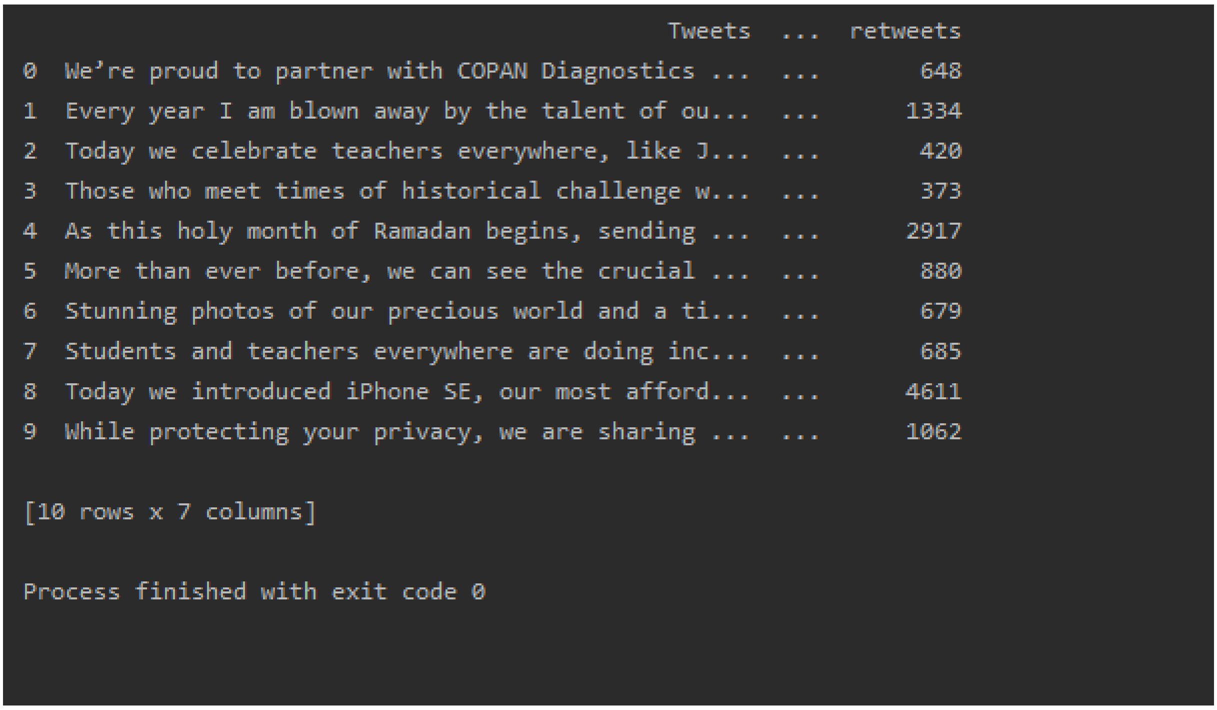
Figure 9-5: Paging

e) Analyze and Visualize:



```
                                               Tweets  ...  retweets
0  Heartfelt condolences to the Kapoor khandaan. ...  ...     15694
1  My friend...inspiration &amp; the greatest act...  ...     43666
2  Thank you all for #AskSrk Need to get back to ...  ...      5554
3  I paid lots of attention to my teachers growin...  ...      2237
4  Medium ko jawaab hi nahi deta...sirf Maximum k...  ...      2225
5  Yeah I should copy paste them again na? https:...  ...      2050
6  God was wondering when someone will ask me thi...  ...      2051
7  Wow both are awesome and I have met them...but...  ...      3107
8  Bhai kamaal ka Single aur Singer hai... https:...  ...      5038
9  Really...khajanchi hai kya??!! https://t.co/kX...  ...      2613

[10 rows x 7 columns]
```

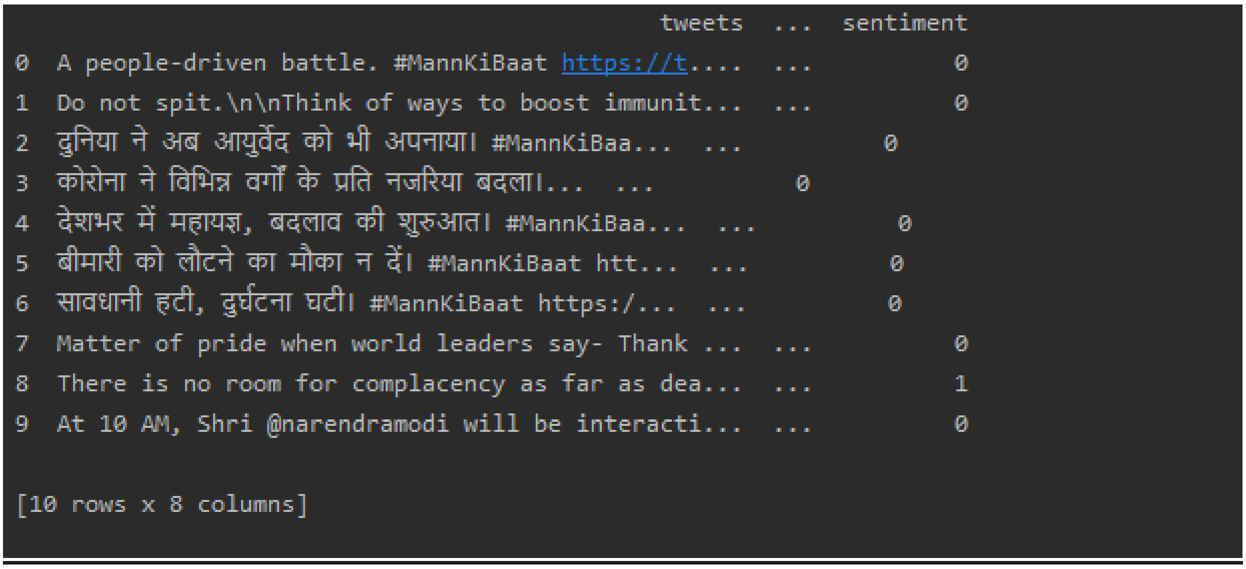
Figure 9-6: Visualization of @iamsrk Tweets



```
                                               Tweets  ...  retweets
0  We’re proud to partner with COPAN Diagnostics ...  ...       648
1  Every year I am blown away by the talent of ou...  ...      1334
2  Today we celebrate teachers everywhere, like J...  ...       420
3  Those who meet times of historical challenge w...  ...       373
4  As this holy month of Ramadan begins, sending ...  ...      2917
5  More than ever before, we can see the crucial ...  ...       880
6  Stunning photos of our precious world and a ti...  ...       679
7  Students and teachers everywhere are doing inc...  ...       685
8  Today we introduced iPhone SE, our most afford...  ...      4611
9  While protecting your privacy, we are sharing ...  ...      1062

[10 rows x 7 columns]

Process finished with exit code 0
```

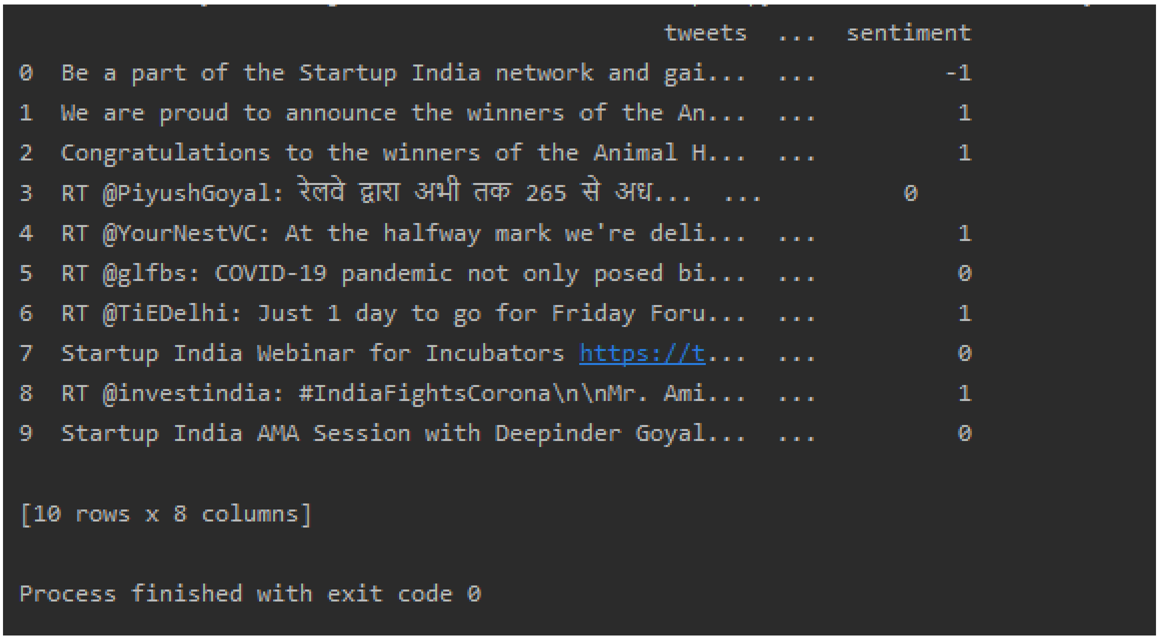
Figure 9-7: Visualization of @tim_cook Tweets

f) Snt Anlysis:



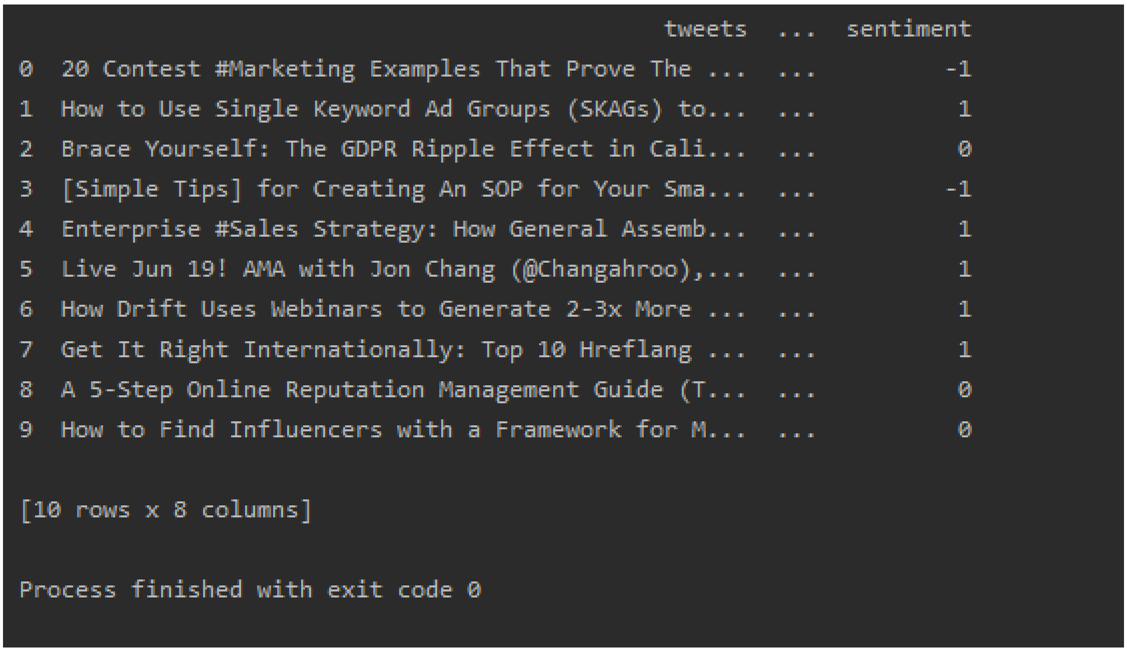
```
                                              tweets  ...  sentiment
0  A people-driven battle. #MannKiBaat https://t....  ...          0
1  Do not spit.\n\nThink of ways to boost immunit...  ...          0
2  दुनिया ने अब आयुर्वेद को भी अपनाया। #MannKiBaa...  ...          0
3  कोरोना ने विभिन्न वर्गों के प्रति नजरिया बदला।...  ...          0
4  देशभर में महायज्ञ, बदलाव की शुरुआत। #MannKiBaa...  ...          0
5  बीमारी को लौटने का मौका न दें। #MannKiBaat htt...  ...          0
6  सावधानी हटी, दुर्घटना घटी। #MannKiBaat https:/...  ...          0
7  Matter of pride when world leaders say- Thank ...  ...          0
8  There is no room for complacency as far as dea...  ...          1
9  At 10 AM, Shri @narendramodi will be interacti...  ...          0

[10 rows x 8 columns]
```

Figure 9-8: Polarity of @PMOIndia Tweets



```
                                              tweets  ...  sentiment
0  Be a part of the Startup India network and gai...  ...         -1
1  We are proud to announce the winners of the An...  ...          1
2  Congratulations to the winners of the Animal H...  ...          1
3  RT @PiyushGoyal: रेलवे द्वारा अभी तक 265 से अध...  ...          0
4  RT @YourNestVC: At the halfway mark we're deli...  ...          1
5  RT @glfbs: COVID-19 pandemic not only posed bi...  ...          0
6  RT @TiEDelhi: Just 1 day to go for Friday Foru...  ...          1
7  Startup India Webinar for Incubators https://t...  ...          0
8  RT @investindia: #IndiaFightsCorona\n\nMr. Ami...  ...          1
9  Startup India AMA Session with Deepinder Goyal...  ...          0

[10 rows x 8 columns]

Process finished with exit code 0
```

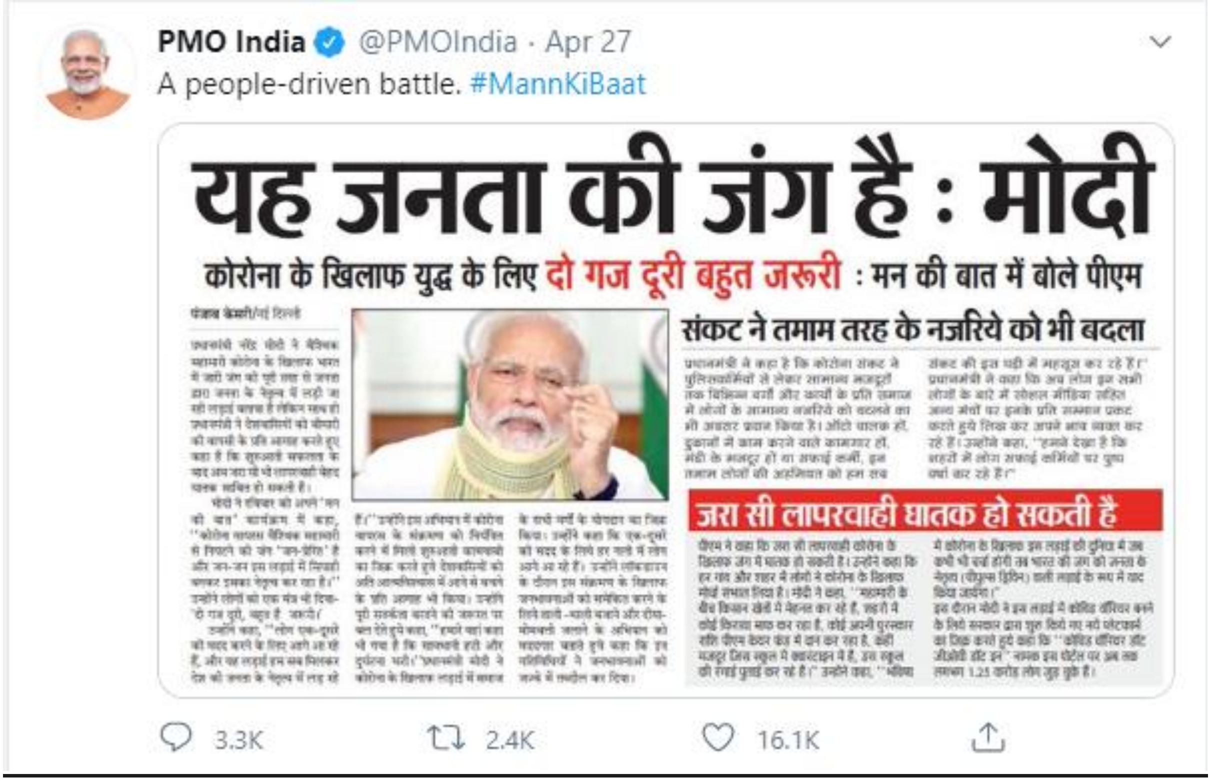
Figure 9-9: Polarity of @startupindia Tweets



```
                                              tweets  ...  sentiment
0  20 Contest #Marketing Examples That Prove The ...  ...         -1
1  How to Use Single Keyword Ad Groups (SKAGs) to...  ...          1
2  Brace Yourself: The GDPR Ripple Effect in Cali...  ...          0
3  [Simple Tips] for Creating An SOP for Your Sma...  ...         -1
4  Enterprise #Sales Strategy: How General Assemb...  ...          1
5  Live Jun 19! AMA with Jon Chang (@Changahroo),...  ...          1
6  How Drift Uses Webinars to Generate 2-3x More ...  ...          1
7  Get It Right Internationally: Top 10 Hreflang ...  ...          1
8  A 5-Step Online Reputation Management Guide (T...  ...          0
9  How to Find Influencers with a Framework for M...  ...          0

[10 rows x 8 columns]

Process finished with exit code 0
```

Figure 9-10: Polarity of @GrowthHackerSMB Tweets

g) <u>Tweets</u>:



PMO India @PMOIndia · Apr 27

A people-driven battle. #MannKiBaat

यह जनता की जंग है : मोदी

कोरोना के खिलाफ युद्ध के लिए दो गज दूरी बहुत जरूरी : मन की बात में बोले पीएम

पंजाब केसरी/नई दिल्ली

[illegible]

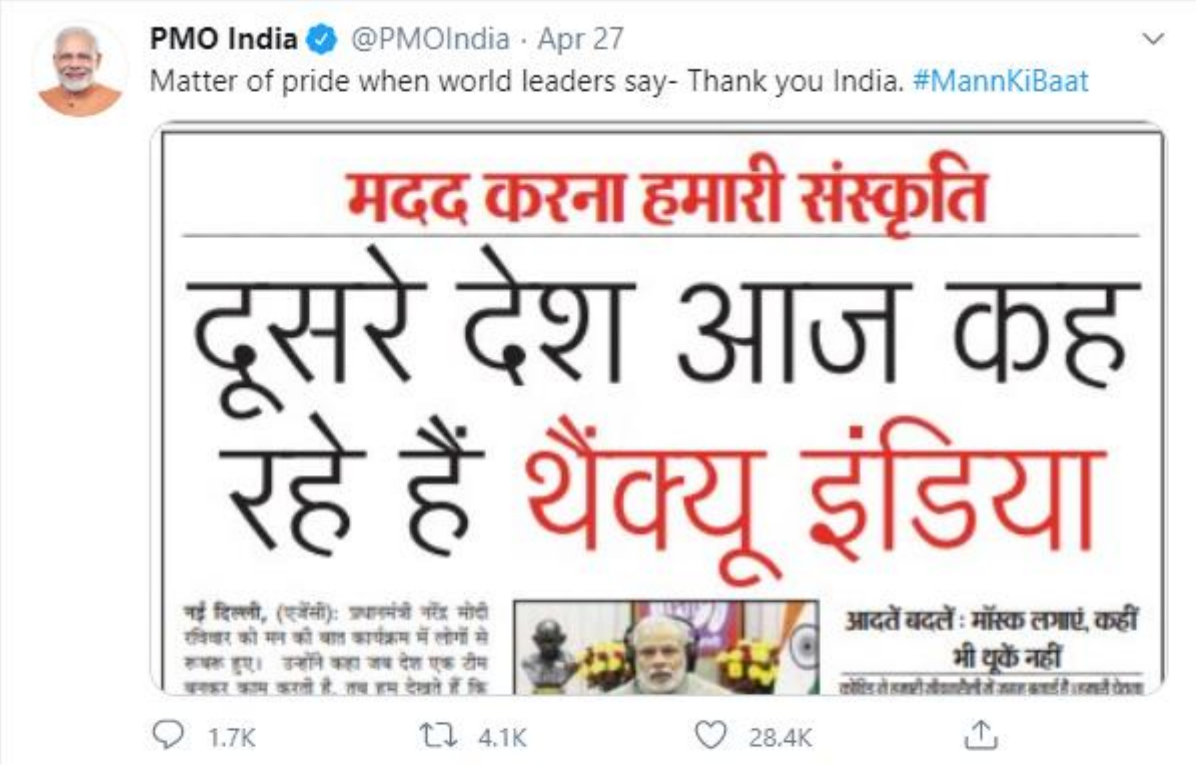
संकट ने तमाम तरह के नजरिये को भी बदला

[illegible]

जरा सी लापरवाही घातक हो सकती है

[illegible]

3.3K 2.4K 16.1K



PMO India @PMOIndia · Apr 27

Matter of pride when world leaders say- Thank you India. #MannKiBaat

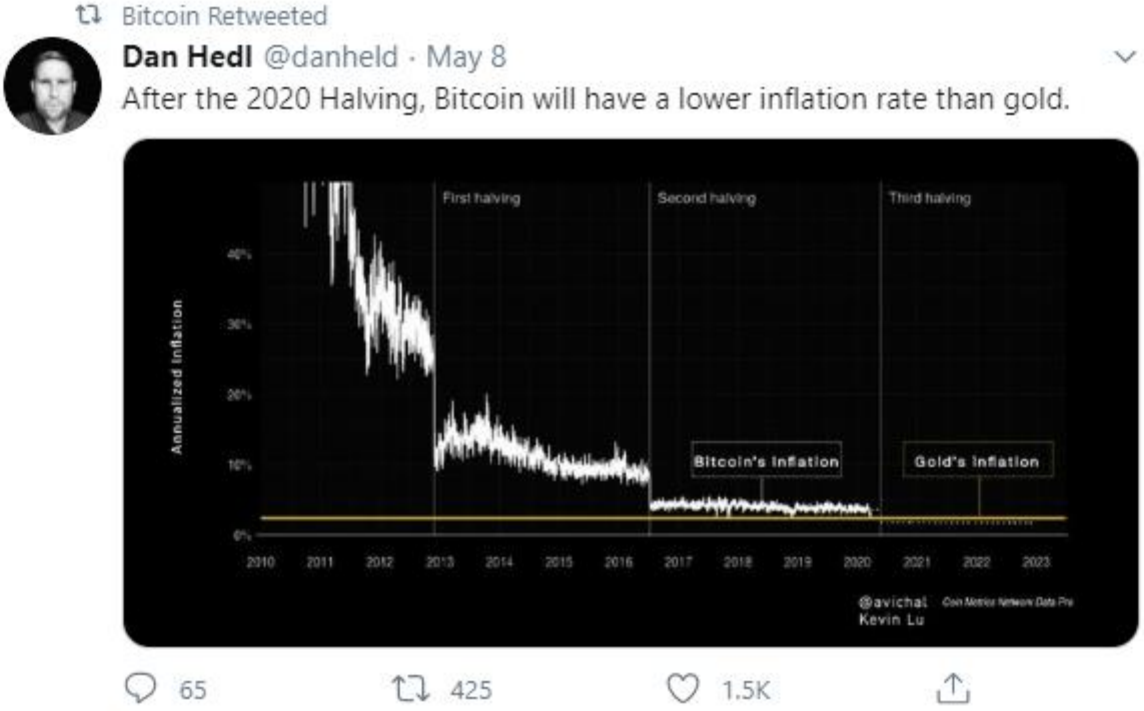
मदद करना हमारी संस्कृति

दूसरे देश आज कह रहे हैं थैंक्यू इंडिया

नई दिल्ली, (एजेंसी): प्रधानमंत्री नरेंद्र मोदी रविवार को मन की बात कार्यक्रम में लोगों से रूबरू हुए। उन्होंने कहा जब देश एक टीम बनकर काम करती है, तब हम देखते हैं कि

आदतें बदलें : मॉस्क लगाएं, कहीं भी थूकें नहीं

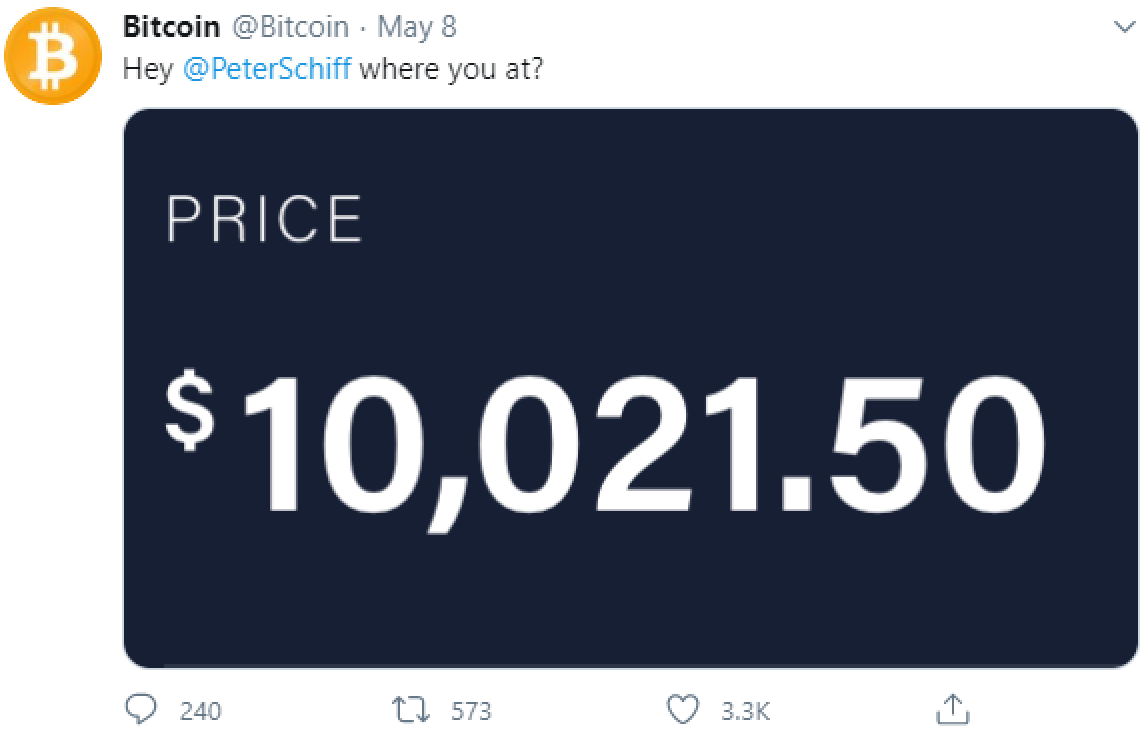
[illegible]



Bitcoin Retweeted

Dan Hedl @danheld · May 8

After the 2020 Halving, Bitcoin will have a lower inflation rate than gold.



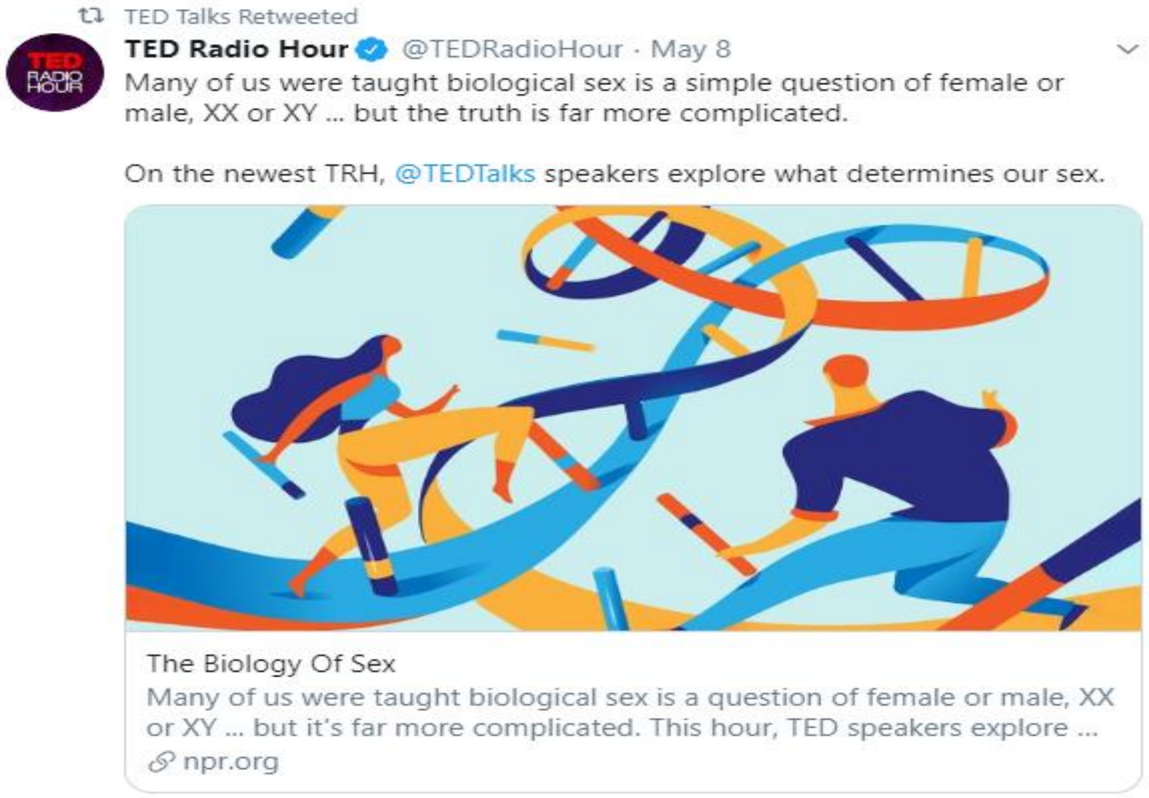
Bitcoin @Bitcoin · May 8

Hey @PeterSchiff where you at?

PRICE

$10,021.50

240 573 3.3K



TED Talks Retweeted

TED Radio Hour @TEDRadioHour · May 8

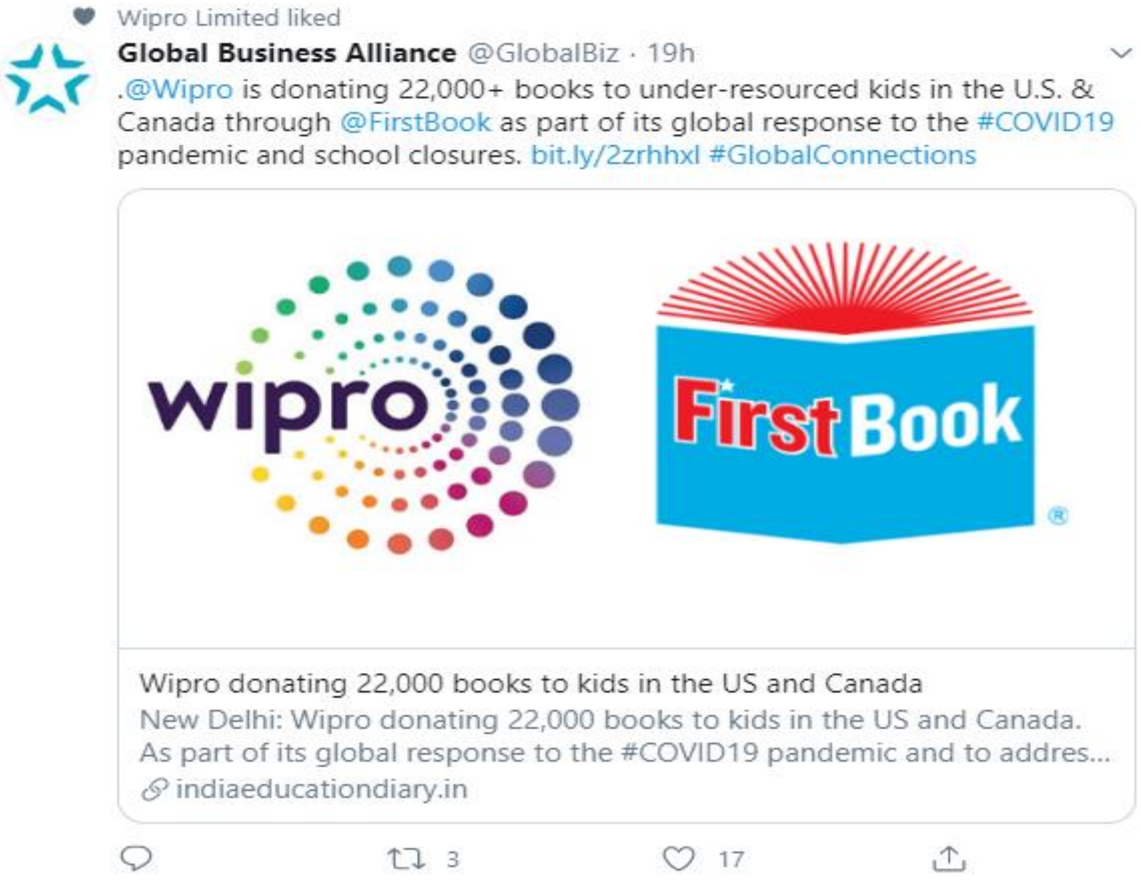
Many of us were taught biological sex is a simple question of female or male, XX or XY ... but the truth is far more complicated.

On the newest TRH, @TEDTalks speakers explore what determines our sex.

The Biology Of Sex

Many of us were taught biological sex is a question of female or male, XX or XY ... but it's far more complicated. This hour, TED speakers explore ...

npr.org



Wipro Limited liked

Global Business Alliance @GlobalBiz · 19h

.@Wipro is donating 22,000+ books to under-resourced kids in the U.S. & Canada through @FirstBook as part of its global response to the #COVID19 pandemic and school closures. bit.ly/2zrhhxl #GlobalConnections

Wipro donating 22,000 books to kids in the US and Canada

New Delhi: Wipro donating 22,000 books to kids in the US and Canada. As part of its global response to the #COVID19 pandemic and to addres...

indiaeducationdiary.in

10. FUTURE WORK

From future perspective, we would like to extend this project by implementing some machine learning algorithms for applications like election results, product ratings, movies' outcomes and running the project on clusters to expand its functionalities. Moreover, we would like to make a web application for users to input keywords and get analyzed results. In this project, we have worked only with unigram models, but we would like to extend it to bigram and further which will increase linkage between the data and provide accurate snt anlysis results. Computation of overall tweet score can be done for a single keyword which can provide an overall snt of the public regarding a topic.

Also a future challenge is in applying snt classification approaches and tools for snt anlysis of posts in social media is to overcome the ambiguity that actually represents particular problem since it is not easily make use of co-reference information. Typically the analyzed posts contain irony and sarcasm, which are particularly difficult to detect. So an evolution of approaches and tools is required to overcome this limitation.

11. CONCLUSION

Twt is a source of vast unstructured and noisy data sets that can be processed to locate interesting patterns and trends. Apache Spark proved prolific in extracting live streams of data and has further capability to store batches of data in HDFS and other major conventional storages. The processing capabilities of Spark makes the project flexible to further extend to multiple nodes, thereby supporting distributed computing. Real-time data anlysis makes it possible for business organizations to keep track of their services and generates opportunities to promote, advertise and improve from time to time.

Our heartfelt appreciation goes to Professor Hari Om Sharan with regards to his feedback across the course of the project from the initial proposal up to the conclusion and for the valuable lessons learned along the way including collaboration within a group and the challenges involved in a large-scale software development efforts.

.

12. REFERENCES

[1] Fu, T., Abbasi, A., Zeng, D., Chen, H. 2012. Sntal Spidering: Leveraging Opinion Information in Focused Crawlers. ACM Transactions on Information Systems

[2] Bollen, J., Mao, H., Zeng, X. 2011. Twt mood predicts the stock market. Journal of Computational Science

[3] Cao, J., Zeng. K., Wang. H., Cheng, J., Qiao, F., Wen, D., Gao, Y. 2014. Web-Based Traffic Snt Anlysis: Methods and Applications. ITS(15) 2, April 2014

[4] Gonçalves, P., Benevenuto, F., Cha, M. 2013. Panas-t: A psychometric scale for measuring snts on twt.

[5] Mullen, T., & Collier, N. 2004. Snt Anlysis using Support Vector Machines with Diverse Information Sources.

[6] M.Lovelin Ponn Felciah,R.Anbuselvi, ”A Study on Snt Anlysis of Social Media Reviews,” IEEE Second Conference on Innovations in Information Embedded & Communication Systems, 2015

[7] Vishal A. Kharde, S.S. Sonawane “Snt Anlysis of Twt Data: A Survey of Techniques” Volume 139 – No.11, April 2016