

# TWITTER SENTIMENT ANALYSIS 2020

---

2020, MAY 26

---

Authored by: VAIBHAV PANDEY  
SID: 18103072



---

## ABSTRACT

In this project we have analyzed the sentiments of the general public regarding the current political scenario of the country with the help of the tweets made by them. Twitter is extensively used by people to voice their political opinions and question the various policies of the government and even the opposition. We have exploited this fact and analyzed these sentiments of the general people. We have used Twitter Application Programming Interface (API) to stream real-time tweets into our MySQL Database. We have developed a Deep Learning model. Relevant data has been mined and after required data cleaning with the help of libraries like snowball Stemmer and nltk, we have tested and trained our model. This model was used to analyze real-time tweets. Analysis revealed that around 43% of the tweets portrayed positive opinion, 21% of people had negative sentiments while 36% public had neutral views regarding the government. Also, regarding the Opposition, our analysis revealed 31% positive, 24% negative and 45% neutral sentiment trends.

---

# Contents

	Page
1) Introduction.....	4
2) Twitter Sentiment Analysis Process	
a) Tweets Extraction	
i) Set up Developers Account.....	5-7
ii) Data Streaming Into Database.....	7-9
b) Preprocessing and Deep Learning Model	
i) Understanding the training dataset.....	10
ii) Preprocessing the dataset.....	10-11
iii) Text Cleaning.....	11
iv) Stemming/Lemmatization.....	12
v) Train-Test Split.....	12
vi) Word2Vec Model.....	12-14
vii) Text Tokenization.....	14
viii) Padding.....	14
ix) Label Encoding.....	14
x) Embedding Label Construction.....	15
xi) Istm Model Architecture.....	15
xii) Model Compilation.....	16
xiii) Callbacks.....	16
xiv) Training Time.....	16-17
c) Visualization of Statistics generated by model.....	18
3) Conclusion.....	19
4) References.....	20

---

# INTRODUCTION

Today we are living in the era of social media and Big Data. People use social media to voice their views and opinions about each and every topic under the sun like sports, culture and politics as well. This data can help us analyze the opinion of the general public, about their joys and woes, and can also in turn help the government in framing their future policies. We are using twitter for the political analysis which is a great platform for getting the sentiments of the people on the current political scenario in the country.

Twitter has been extensively used by various firms for data mining. Prashant Kishor is a renowned political strategist of India. He is credited for formulating the successful marketing and advertising campaign for Mr. Narendra Modi for the 2014 General Elections. He along with his other group members formed the I-PAC (Political Action Committee for India) in 2015. I-PAC has played great role in shaping the results of Bihar Elections of 2015 ,Punjab Assembly Elections of 2017, Andhra Pradesh Assembly Elections of 2019 and many more. I-PAC along with its other mediums, has extensively used Twitter Analysis to schedule successive political campaigns. Also in the Facebook-Cambridge Analytica data scandal personal data of millions of people was harvested without their consent and was used for political advertisement purposes. Many e-commerce sites also employ sentiment analysis using product review system to improve their future marketing and advertising techniques.

We have analyzed millions of tweets made by the Indian people in which they have voiced their opinion about general current politics and schemes. We have deconstructed their thoughts and feelings on the basis of their tweets. Thus, in turn, obtained a clearer picture of the mindset of the masses. This analysis will help us to surface the actual political feelings of the people of the country. We can successfully deduce the feelings of contentment and resentment of the general public through this analysis.

---

# TWITTER SENTIMENT ANALYSIS PROCESS

The analysis process consists mainly of three sub-processes:

- Tweets Extraction
- Tweets Preprocessing and further Processing
- Data Visualization

## Tweets Extraction

This comprises the data collection part of the project. Data collection is the process of gathering quantitative and qualitative information on specific variables with the aim of evaluating outcomes gleaned or actionable insights. Good data collection requires a clear process to ensure the data you collect is clean, consistent, and reliable. Establishing that process, however, can be tricky. It involves taking stock of your objectives, identifying your data requirements, deciding on a method of data collection, and finally organizing a data collection plan that synthesizes the most important aspects of your program.

### 1) Set-up for the Developers Account

In this project we have used Twitter Application Programming Interface (API) to collect real time tweets made by the people. The application we created is named 'Analysis1.0' as shown in the below image. To create a Twitter API we first need to set-up a Twitter Developers Account. To setup such an account requires authorization by Twitter. Twitter scrutinizes applications for such an account on a

---

case by case basis. After completing the rather tedious authorization process, we obtain the account's credentials :

- Access Token
- Access Token Secret
- Consumer Key
- Consumer Key Secret


These credentials are the authorization keys which are required to be supplied in our Data Extraction model.

This developers account allows us to create a twitter API which in turn allows us to stream real-time tweets made by the public. Our application is linked to my Twitter Account and has been used to collect the required tweets for the project.

**App details**

Edit ▾

Details and URLs

 **App icon**  
App icon is default, click edit to upload.

**App Name**  
Analysis1.0

**Description**  
Analysis app.

**Website URL**  
<https://twitter.com/vpntluk>

**Sign in with Twitter**  
Enabled

**Callback URL**  
<https://twitter.com/>

**Terms of service URL**  
None

**Privacy policy URL**  
None

**Organization name**  
NJ & VP Co.

**Organization website URL**  
None

**App usage**  
We want to use twitter for sentiment analysis of Indian people. Since twitter is a very good platform for expressing one's opinion so we can use the tweets made by people for sentiment analysis.

## 2) Data Streaming Into Database

Streaming tweets into database requires the following two things :

- A MySQL Database
- The Tweepy and mysql-connector

---

## Setting up MySQL Database

We need to set-up a database before we can start streaming real-time tweets. We have used MySQL database which is the most popular database around the world. The schema for the database includes :

- username : VARCHAR(255)
- created\_at : VARCHAR(45)
- tweet : TEXT
- retweet\_count : INT(11)
- location : VARCHAR(100)

## The Tweepy and mysql-connector

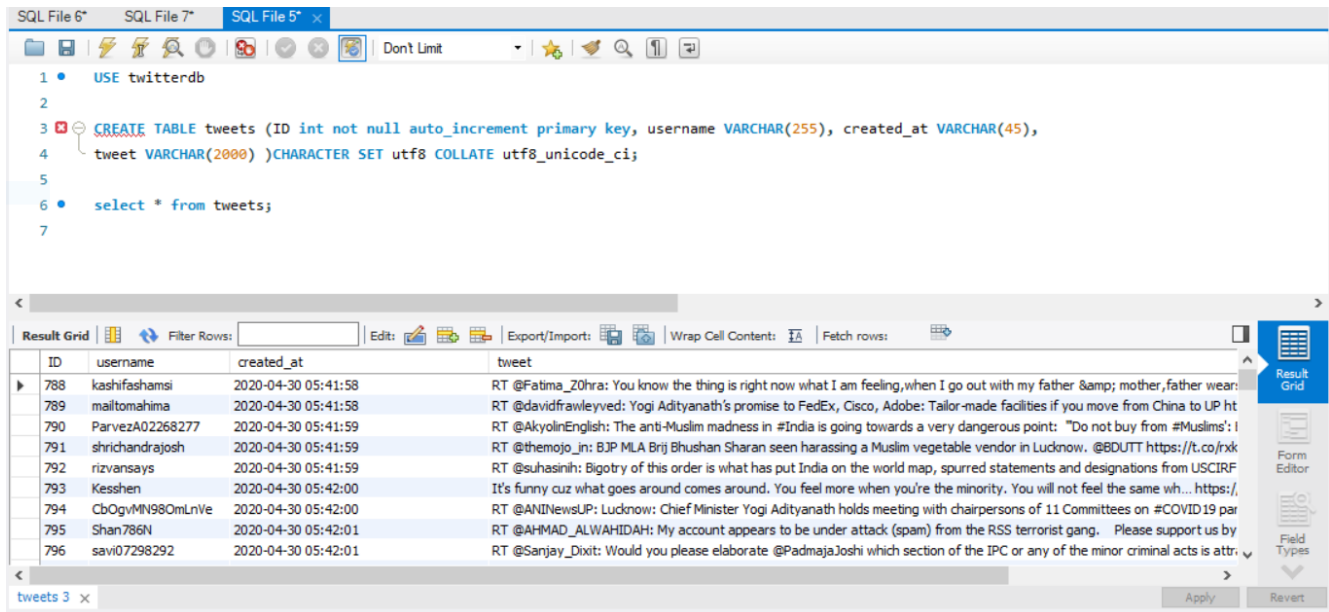
Now we want our code to do the following tasks :

- We want to create a class that allows us to connect to the Twitter API. This is achieved with the help of Tweepy library. After importing this library we need to complete the following tasks to successfully connect to Twitter API:
  - 1.Create a class inheriting from 'StreamListener'.
  - 2.Instantiate an object from this class.
  - 3.Use this object to connect to the API.
- We also need to create some code that connects to our database and reads the data into the correct columns. This can be successfully completed by importing the mysql-connector and using it to connect to our database which we have already created.

After setting all the keywords which we are looking for in our tweets, all we need to do is make use of the functions of the object used for connecting to the API to stream the tweets data. This data is received in the form of a json object which has many key-value fields. This objects contains all the required data like username, created\_at and the tweet text. Thereafter the information extracted from this object is stored into our database.



An instance of the database is shown below. The name of our database is 'twitterdb' and the table used for storing the incoming data is named 'tweets'.



```
1 • USE twitterdb
2
3 CREATE TABLE tweets (ID int not null auto_increment primary key, username VARCHAR(255), created_at VARCHAR(45),
4 tweet VARCHAR(2000) )CHARACTER SET utf8 COLLATE utf8_unicode_ci;
5
6 • select * from tweets;
7
```

ID	username	created_at	tweet
788	kashifashamsi	2020-04-30 05:41:58	RT @Fatima_Zohra: You know the thing is right now what I am feeling,when I go out with my father & mother,father wear:
789	mailtomahima	2020-04-30 05:41:58	RT @davidfrawleyved: Yogi Adityanath's promise to FedEx, Cisco, Adobe: Tailor-made facilities if you move from China to UP ht
790	ParvezA02268277	2020-04-30 05:41:59	RT @AkyolinEnglish: The anti-Muslim madness in #India is going towards a very dangerous point: "Do not buy from #Muslims": I
791	shrichandrajosh	2020-04-30 05:41:59	RT @themojo_in: BJP MLA Brij Bhushan Sharan seen harassing a Muslim vegetable vendor in Lucknow. @BDUTT https://t.co/rxk
792	rizvansays	2020-04-30 05:41:59	RT @suhasinih: Bigotry of this order is what has put India on the world map, spurred statements and designations from USCIRF
793	Kesshen	2020-04-30 05:42:00	It's funny cuz what goes around comes around. You feel more when you're the minority. You will not feel the same wh... https://
794	CbOgvMN98OmLnVe	2020-04-30 05:42:00	RT @ANINewsUP: Lucknow: Chief Minister Yogi Adityanath holds meeting with chairpersons of 11 Committees on #COVID19 par
795	Shan786N	2020-04-30 05:42:01	RT @AHMAD_ALWAHIDAH: My account appears to be under attack (spam) from the RSS terrorist gang. Please support us by
796	savi07298292	2020-04-30 05:42:01	RT @Sanjay_Dixit: Would you please elaborate @PadmajaJoshi which section of the IPC or any of the minor criminal acts is attr

With this, we have successfully collected all the relevant data required for our analysis. Now we need to process this data to obtain the required sentiment trends.

---

# Preprocessing and Deep Learning Model

## 1 . Understanding the training dataset

The dataset is encoded in format "ISO-8859-1" and it has six columns namely:

- target
- ids
- date
- flag
- user
- Text

For reading the dataset pandas library of python has been used.

★ Basic definition of pandas library is as follows:

Pandas is a high-level data manipulation tool developed by Wes McKinney. It is built on the NumPy package and its key data structure is called the Data Frame.

The dataset used has 1,600,000 rows. In the dataset target column has 3 types of sentiments which are:

- Negative
- Neutral
- Positive

## 2 . Preprocessing the dataset

Before applying any machine learning algorithm data preprocessing needs to be done to make the data fit for training.

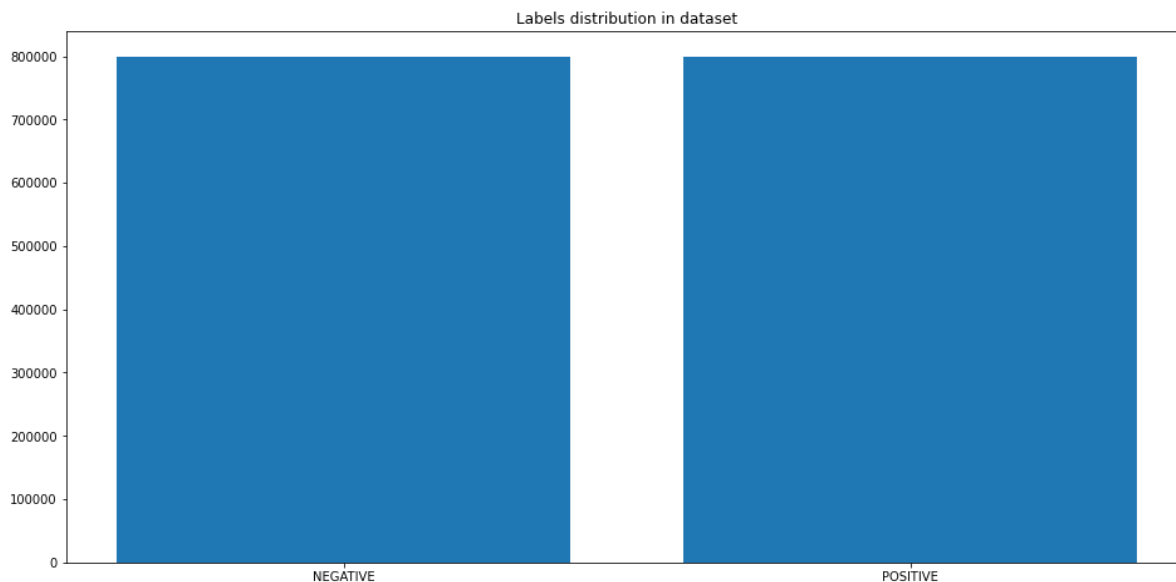
Steps:

- Checking imbalance in dataset

- Text cleaning
- Stemming/Lemmatization
- Train-Test Split

Checking imbalance in dataset :

- In this particular step we need to count the frequency of the Positive and Negative sentiments. If a particular sentiment is in excess then it can hamper the accuracy of the model. Luckily the dataset was balanced and had equal proportion of Positive and Negative sentiments.



### 3. Text Cleaning:

In our dataset many of the tweets have words like @, #, http?0-9 etc. These do not convey any meaning. Also, there are many stop words like is, the, are, they etc. These too don't convey any meaning and hence have been removed from the dataset. Also, all capital letters have been converted into small letters.

This has been done using libraries:

- nltk
- stop words
- re
- snowball Stemmer

---

## 4 . Stemming / Lemmatization:

### ★ Basic definition of stemming:

Stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words known as a lemma.

This has been done to reducing the words to their root form.

## 5 . Train-Test Split:

In this project an 80-20 split rule has been followed and also random state has been kept as 42 for better results. This has been done using train test split module of sklearn library.

## 6 . Word2Vec Model

Some common terminologies which will be used further are described below.

### ★ Understanding Word2Vec:

Word2Vec is one of many different word embedding techniques. In turn, word embedding is one of the most popular representation of document vocabulary.

Thus, hierarchy is as follows:

- Document vocabulary representation -> Word embedding -> Word2Vec

### ★ Understanding word embedding:

Word embedding is vector representation of a word

### ★ Understanding dictionary:

Dictionary of a corpus of text is a data structure which consists of all unique words.

So, Word2Vec is a two-layer neural network that takes as its input a large corpus of text and produces a vector space, typically of several hundred

dimensions, with each unique word in the corpus being assigned a corresponding vector in the space

A hand-drawn diagram illustrating word vectors. On the left, a vertical column of boxes represents dimensions: 'Royalty', 'Masculinity', 'Femininity', 'Age', and an ellipsis. To the right, four vertical columns represent words: 'King', 'Queen', 'Woman', and 'Princess', each with corresponding numerical values in the dimension boxes.

	King	Queen	Woman	Princess
Royalty	0.99	0.99	0.02	0.98
Masculinity	0.99	0.05	0.01	0.02
Femininity	0.05	0.93	0.999	0.94
Age	0.7	0.6	0.5	0.1

Next in this project the word2Vec model has been built.

The parameters in it have been set as:

➤ **size=300**

size is the number of dimensions (N) of the N- dimensional space that genism Word2Vec maps the words onto.

➤ **window=7**

It is maximum distance between the current and predicted word within a sentence.

➤ **min\_count=10**

min\_count is for pruning the internal dictionary. So, all words having frequency less than 10 have been omitted from dictionary;

➤ **workers=8**

Specifies number of cores used, it has been set to 8 to fasten processing.

---

Using this Word2Vec model vocabulary has been built based on tf-idf . Tf-idf stands for term frequency-inverse document frequency. The goal of using tf-idf is to scale down the impact of tokens that occur very frequently in a given corpus and that are hence empirically less informative than features that occur in a small fraction of the training corpus.

- $tf(t)$ = the term frequency is the number of times the term appears in the document
- $idf(d, t)$  = the document frequency is the number of documents 'd' that contain term 't'

On basis of the above parameters the vocabulary has learnt 30520 unique words.

This is a pretrained model. For better results we trained the model on the dataset for 32 epochs.

## 7 . Text Tokenization

Tokenization is the process where a text corpus is vectorized by turning each text into a vector where coefficient for each token is based on tf-idf .

## 8 . Padding

Input data for a deep learning model must be a single tensor so samples that are shorter than the longest item have been padded with placeholder and longer ones have been truncated to length 300.

## 9 . Label encoding

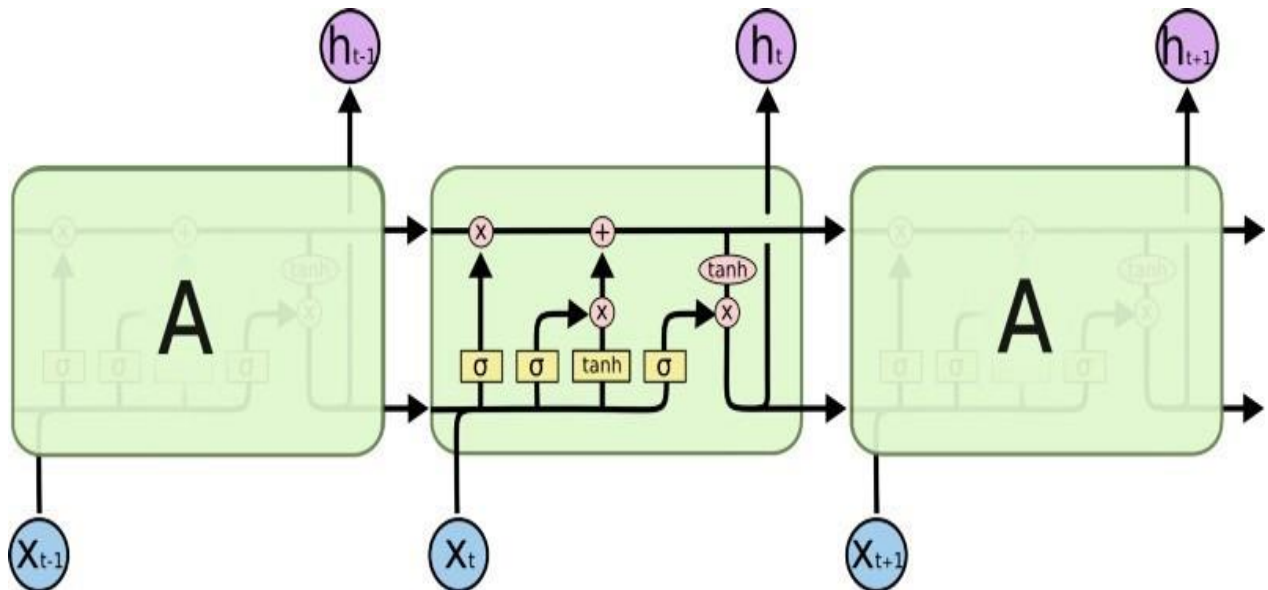
This has been performed on target column since deep learning model can't be fed values POSITIVE, NEGATIVE, NEUTRAL

## 10 . Embedding layer construction

Embedding layer has dimensions of  $290572 \times 300$  as each word has dimension of 300 and there are 290572 words in training dataset.

## 11 . LSTM Model architecture

The model is sequential and on top of the embedding layer LSTM layer has been added with output dimensionality of 100 and finally on top of it a dense layer with a single unit has been added with activation as sigmoid. In the LSTM and embedding layer dropout layer with dropout rate = 0.2 has been added to avoid overfitting.



The model has 87,332,101 parameters. The embedding layer parameters have been set as non-trainable so the total trainable parameters in the model are 160,501.

## 12 . Model Compilation

The loss function in the model has been defined as `binary_crossentropy` as result is either positive or negative and metrics has been defined as accuracy. The batch size of the model has been set as 1024.

➤ Model: "sequential\_1"

```
_____ Layer (type) Output Shape Param #
=====
===== embedding_1 (Embedding) (None, 300, 300) 87171600
_____
_____ dropout_1 (Dropout) (None, 300, 300) 0
_____
_____ lstm_1 (LSTM) (None, 100) 160400
_____
_____ dense_1 (Dense) (None, 1) 101
=====
===== Total params: 87,332,101
===== Trainable params: 160,501
===== Non-trainable params: 87,171,600
```

## 13 . Callbacks

Callbacks have been used to avoid overfitting of the model. The learning rate has also been adjusted on the fly for accurate results. The callbacks which have been used are:

- Early Stopping
- ReduceLROnPlateau

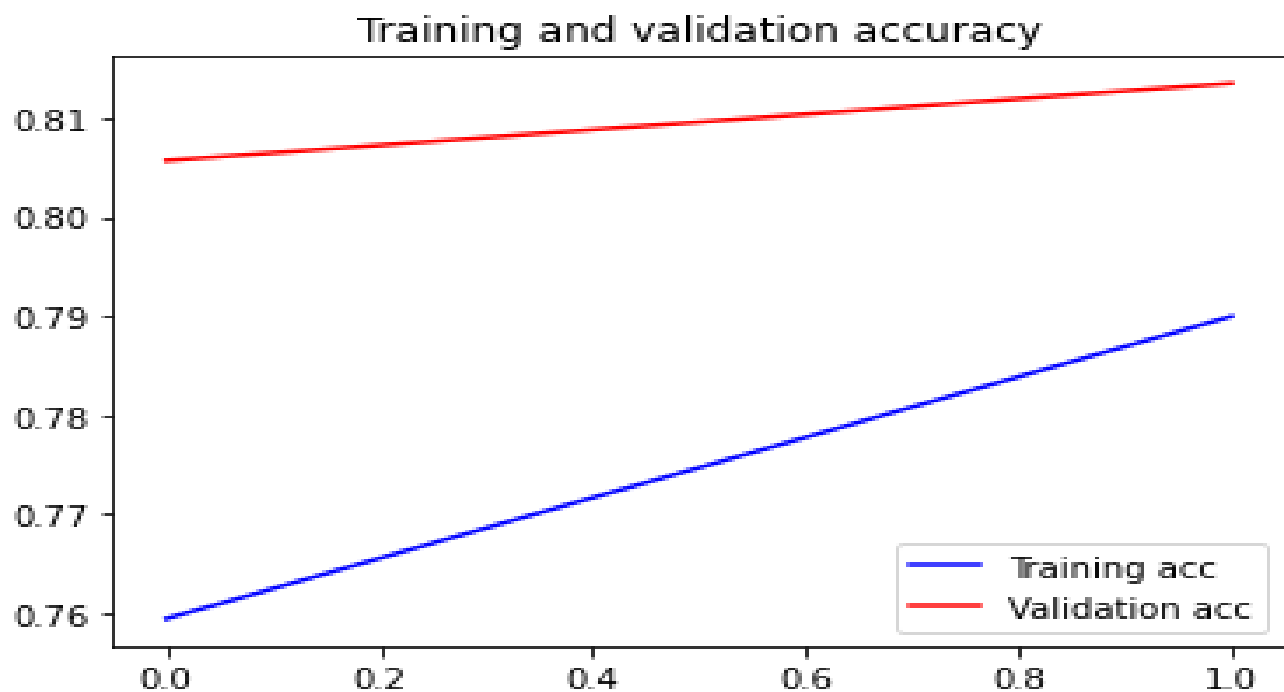
## 14 . Training time

The model took 169 minutes for training.



- Predicting sentiments
- The threshold for Positive, Negative and Neutral sentiments are:
  - $\text{score} \leq 0.4 \rightarrow \text{NEGATIVE}$
  - $0.4 < \text{score} < 0.7 \rightarrow \text{NEUTRAL}$
  - $\text{score} \geq 0.7 \rightarrow \text{POSITIVE}$

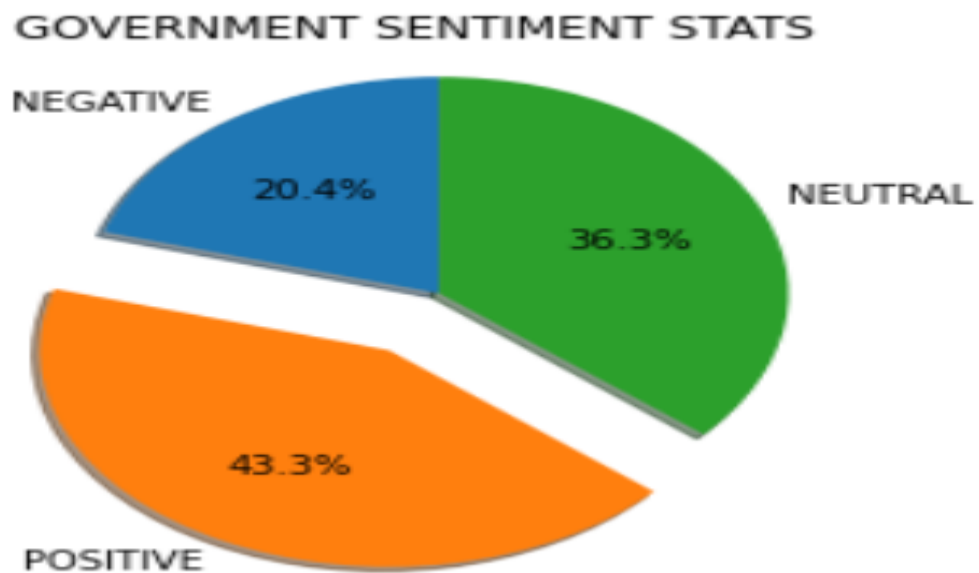
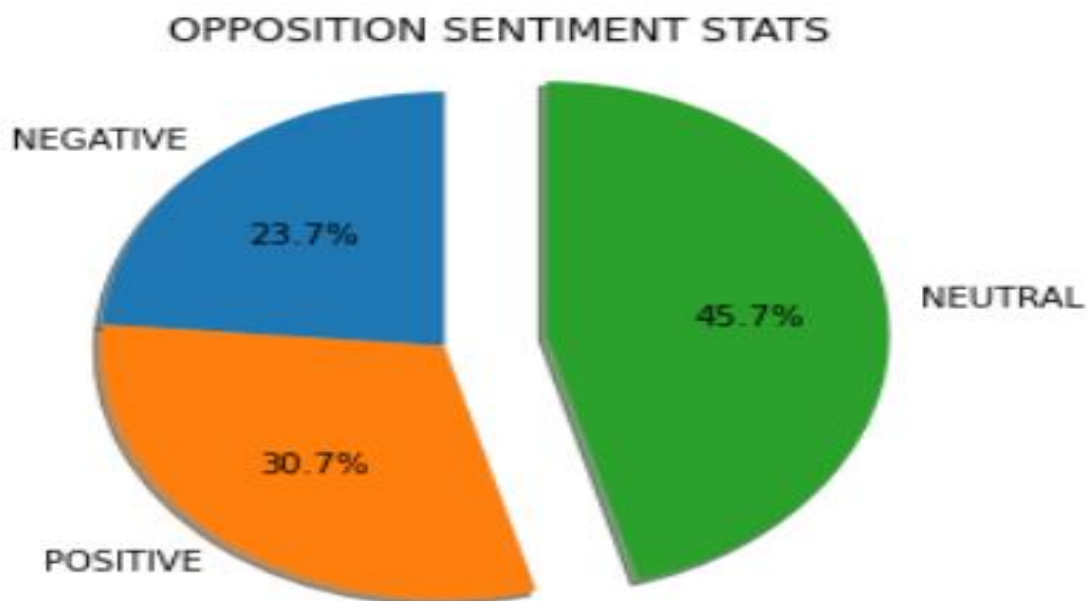
The model has an accuracy of 82% on the testing data.



---

## Visualization of statistics generated by model

---



---

## CONCLUSION

We used twitter data to analyze the political sentiments of the public. A robust Deep Learning model together with a real-time dataset has been the highlight of our project. We have successfully developed a model which can reflect the feelings of resentment or contentment of the public regarding not just political issues but concerning any topic of public interest. Sentiment Analysis has already become a powerful tool for firms like e-commerce sites. These sites make use of such analysis to improve their customer experience and user interface. Sentiment analysis is also extensively used for Product Analytics and Reputation Management through social media monitoring.

In current times, where data is considered the ‘new oil’ of the modern world, the goal is to transform data into information and information into insight. Data coupled with machine learning technology can help us achieve wonders. Machine learning has helped us achieve things which were considered impossible just until recent past and it is still evolving. We can expect an even deeper personalization of these machine learning models in the future which may even change our way of living itself and as Pearl Zhu, author of the “Digital Master” book series, has rightly said, “We are moving slowly into an era where big data is the starting point, not the end.”

---

## REFERENCES

1. Twitter API : <https://developer.twitter.com/en>
2. Insights for deep Learning Model : B. Pang, L. Lee, S. Vaidyanathan. Thumbs up? Sentiment Classification using Machine Learning Techniques, 2002.
3. G. P. Gang Leng and T. M. McGinnity. An on-line algorithm for creating self-organizing fuzzy neural networks. *Neural Networks*, 17(10):1477–1493.
4. Streaming tweets into database : <https://towardsdatascience.com/streaming-twitter-data-into-a-mysql-database-d62a02b050d6>
5. Training and testing dataset : [www.kaggle.com/kazanova/sentiment140](http://www.kaggle.com/kazanova/sentiment140)
6. Rnn and lstm guide for sentiment analysis, detailed step by step insights : <https://medium.com/@lamiae.hana/a-step-by-step-guide-on-sentiment-analysis-with-rnn-and-lstm-3a293817e314>
7. lstm model : <https://towardsdatascience.com/sentiment-analysis-using-lstm-step-by-step-50d074f09948>