



COVID19- Twitter Analysis

Online internship Project Report

By

Aakarsh Yadav
Btech EE(3rd year)
Manipal University Jaipur.

And

Sanjana Rao
Btech Computer Science(3rd year)
Cummins College of Engineering, Pune

May 2020 -June 2020

Preface

This project has a prime focus on Twitter a famous social media site. Twitter is a platform where people can send and receive messages or “tweets” globally. Within 280 character limits, users can share their posts which can include links to relevant websites, videos, images, also their thoughts and opinions on different subjects for that matter. Such huge data can come handy to extract information and make use of it.

This particular report is for task 3 i.e Indian Sentiment analysis after 3 lockdowns using Tweets. Sentiment analysis is a text analysis method that detects polarity (e.g. a *positive* or *negative* opinion) within the text, whether a whole document, paragraph, sentence, or clause. The broader objective of the project is to understand how to analyze twitter data, which includes gathering the dataset, cleaning it, extracting the required information, and drawing conclusions out of it.

Contents

Sr no.	Topic	Page no.
1.	Introduction -Objective -Methodology -Data Source -Problem to be solved	3 4
2.	Chapter 1 -Technologies used - Sentiment analysis	6 7
3.	Chapter 2 -The total process	8
4	Scope, Limitations, and Conclusion	13
5.	References	14

Introduction

1. Objective -

The main objective of task 3 is to perform sentiment analysis that is what is the reaction of citizens to the decision of lockdown imposed by the government to tackle the pandemic situation.

People make use of Twitter to express their feelings about the lockdowns, and this data about the tweets of masses is collected and sentiment analysis is performed. Whether people have positive opinions, negative opinions, or if they are neutral about the situation, this primarily needs to be extracted from the dataset available.

2. Methodology

The fundamental methodology of the project is shown in the flow diagram:

These are the basic implied steps that are performed and under the umbrella of these steps, different tasks are broken down to get the result.

As shown in the diagram the basic steps are Dataset collection, cleaning of data, performing Exploratory data analysis, and finally performing the sentiment analysis.



Flow Diagram

3.Data Source


- Hashtag_data.csv: This dataset was already provided, (which is scrapped from Twitter using Tweepy and Twitter API)

This CSV file contained 30 major rows which have information about userid, username, the tweets, retweets, hashtag, etc.

A total of 124385 columns was provided with all the necessary information mentioned above.

4. The Problem to be solved:

Coronavirus has taken over the world in just a few months causing havoc and leading to many disasters related to almost every sector. Along with the economy and physical health our mental health is also changing. Due to the worldwide lockdown and work from home rules, our brain and emotions



are experiencing something new. Amidst this lockdown, many people are missing their old lifestyle and many are happy to stay at home.

Pandemic and lockdown both have triggered people to show different views regarding them and thankfully we are in the age of technology to see the world reacting about the situation on social media. We are trying to find the reactions of citizens of India due to the pandemic and as well as the lockdowns from their reactions on twitter.

Chapter 1:

● Technologies Used:

-Jupyter Notebook: This open-source software is used for coding, testing, debugging purposes.

-Language: Python

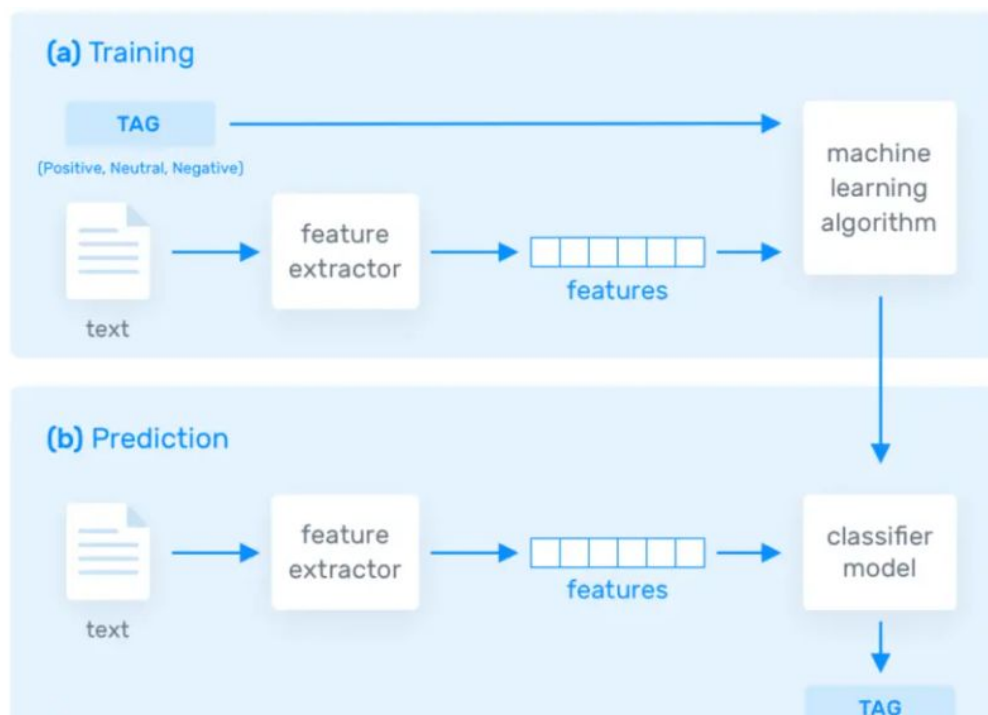
Imports used are shown in the table:

	Import name	Use
1.	Pandas	For handling data and perform functions on them
2.	Numpy	For mathematical operations
3.	Matplotlib.pyplot	For visualization and EDA
4.	Seaborn	For better informative statistical graphics
5.	Import CountVectorizer from above	converts a collection of text documents to a vector of term/token counts.
6.	NLTK	work with human language data for applying in statistical natural language processing (NLP).
7.	Vader	For Sentiment Analysis
8.	Import SentimentIntensityAnalyzer from above	Tells about positivity and negativity score
9	re	For string searching and manipulation.

- **Sentiment analysis :**

Sentiment analysis uses various Natural Language Processing (NLP) methods and algorithms: like Rule based Systems, Automatic based systems, and Hybrid systems

- Working - (Src - ref[1])



- a) Training and prediction process: The model learns to associate particular input to the corresponding tag
- b) Feature Extraction from the text- a machine learning step
- c) Classification Algorithms: This usually involves a statistical model like Naïve Bayes, Logistic Regression, Support Vector Machines, or Neural Networks.

Chapter 2:

2.1 The Total Process

The CSV data which was provided is read by pandas for performing further functions

→ Data Cleaning:

The next step is to clean and prepare the data such that the unnecessary features are removed

Many columns in the dataset were null objects and therefore they are of no use for our further coding, also some of the columns had missing values. Hence such columns are removed. Now we have only necessary metadata which would be properly used in further process.

We make use of Pandas to perform the above actions.

→ Exploratory Data Analysis :

Initially, we take a look at the tweets and the hashtags by creating a word cloud image(Makes use of Wordcloud package)

- Cleaning of text data:

The main part of the data is the "tweet" and the "hashtags", These two columns are textual data and unclean.

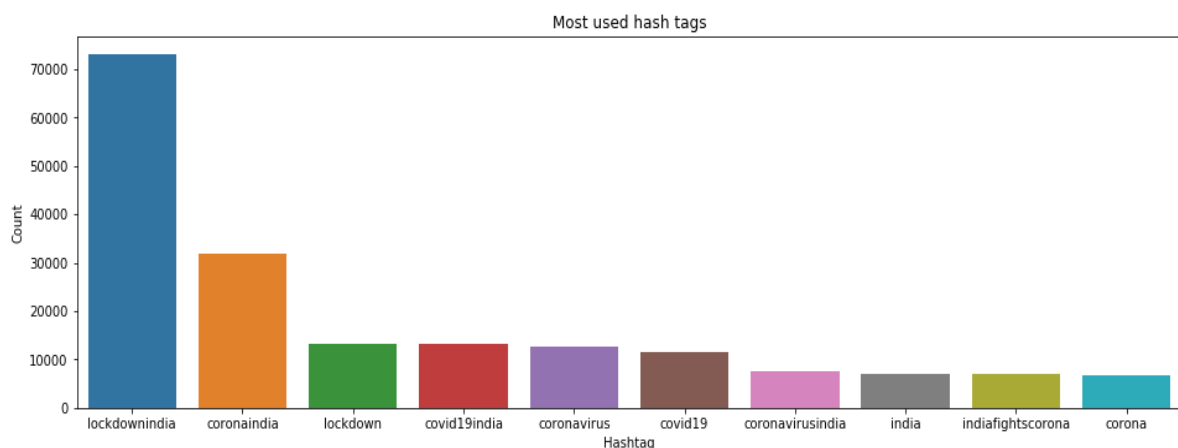
To get it ready for the sentiment analysis we need to perform a few important steps:-

1. Remove punctuations
2. Tokenization - Converting a sentence into a list of words
3. Remove stopwords
4. Lemmatization/stemming - Transforming any form of a word to its root word

The use of “re” package is done for the above steps.

- Some Visual representations: Converting the list of hashtags into dataframes we plot the tags according to their frequency count.

The graph is shown below. (use of matplotlib)



→ The Sentiment analysis

Using the SentimentIntensityAnalyzer from nltk.sentiment.vader majority of analysis is done.

1. Assignment of categorical values to the sentiment: This makes use of vader scores which is assigned as follows-

- If VADER score is more than 0.7 then it is a positive sentiment

- If VADER score is between 0 and 0.7 then it is a neutral sentiment

- If VADER score is less than 0.0 then it is a negative sentiment

2. Dates of the Lockdown :

All the lockdowns were assigned a particular period of time so hence we create dataframe of dates for all the lockdowns as follows:

- A. Janta Curfew = 22 MARCH 2020 for 1 day
- B. 1st Lockdown = 25 March 2020 to 14 APRIL 2020
- C. 2nd Lockdown = 15 APRIL 2020 to 3 MAY 2020
- D. 3rd Lockdown = 4 MAY 2020 to 17 May 2020
- E. 4th Lockdown = 18 MAY 2020 to 31 MAY 2020

F. 5th Lockdown = 1 JUNE 2020 to 30 JUNE 2020

This is performed using the pandas package.

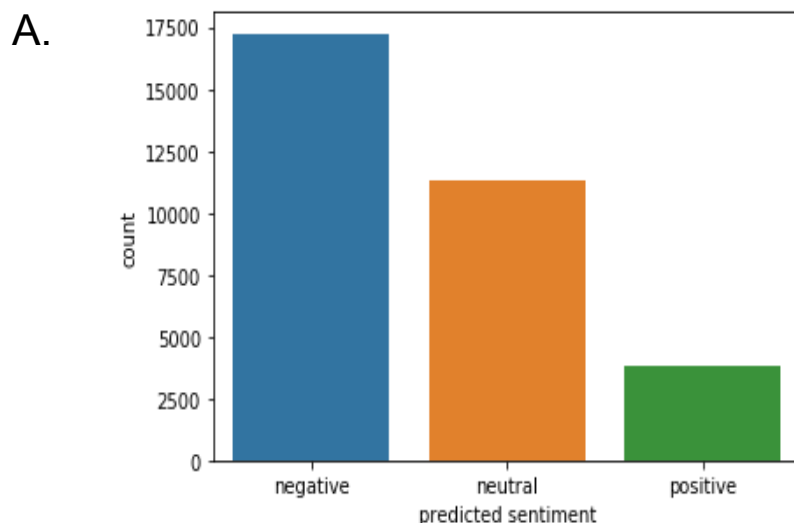
Now sentiment analysis on every lockdown is done

- A. i.e positive negative and neutral statements are analyzed with their frequency count,
- B. Along with that EDA of which hashtag is trending is also performed.
- C. The count of retweets and likes of each sentiment is also done.
- D. A wordcloud of hashtags and tweets is also visualized

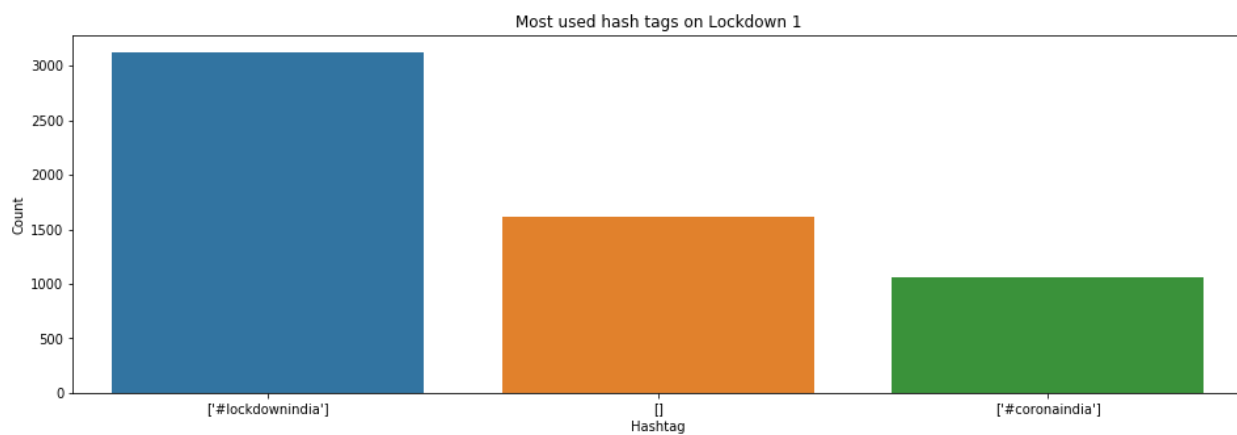
(Images of visualization of the above analysis are shown below)

All this is done with the help of nltk, seaborn, and pandas packages.

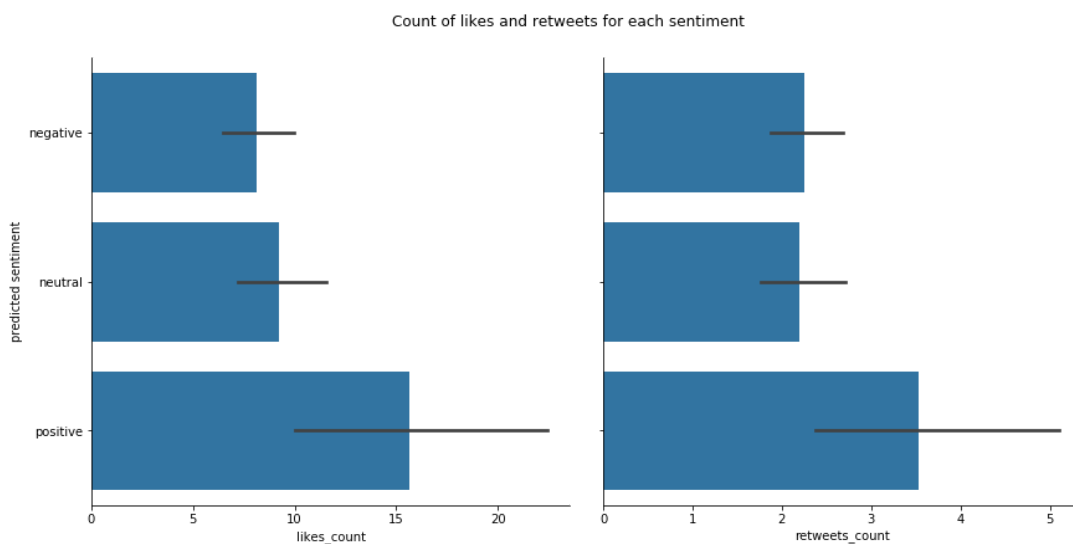
Example of lockdown 1:



B.



C.

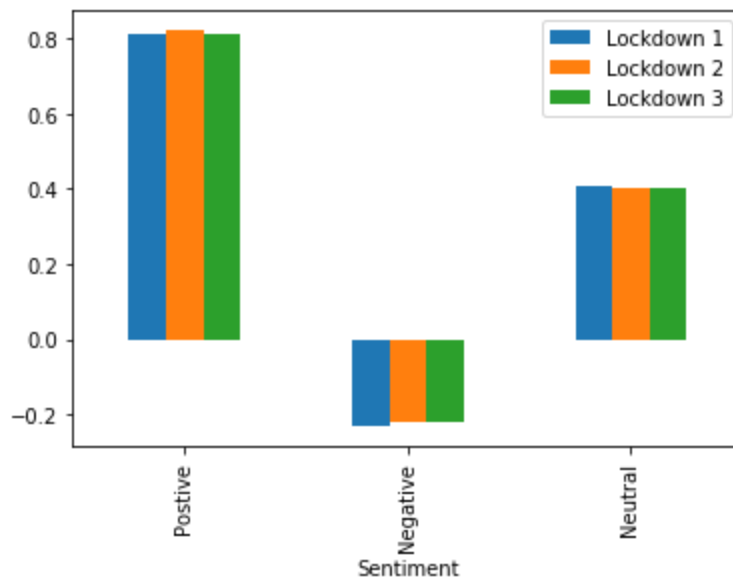


D.



• Comparisons

-All the 3 sentiments in the lockdowns are compared with each other:



(Use of numpy and matplotlib)

At the end we can conclude the following points:-

1. Throughout the first 3 lockdowns a number of tweets with the negative sentiments dominated.
2. Interaction with the tweets with each lockdown varied.
3. Though the number of interactions varied but the average sentiment remained the same.

Scope, Limitations, and Conclusion :

1. Scope:

Sentiment analysis is the interpretation and classification of emotions (positive, negative, and neutral) within text data using text analysis techniques. Sentiment analysis allows businesses to identify customer sentiment toward products, brands, or services in online conversations and feedback.

2. Limitations :

The following aspects are the limitations of Sentiment analysis:

- Incomplete Sentences and Irrelevant information
- Irony and sarcasm
- Emojis and use of characters
- Defining what is Neutral.

3. Conclusion:

So in this internship project, we learned how to perform analysis on Twitter data to extract relevant information and visualize it.

The main takeaways would be performing EDA and performing Sentiment analysis

References

- [1] <https://monkeylearn.com/sentiment-analysis/>
- [2] <https://www.kaggle.com/yaakarsh1011/ds4c-eda-of-covid-19-in-s-korea#1.-The-Lists-of-Data-Table>
- [3] <https://towardsdatascience.com/a-complete-exploratory-data-analysis-and-visualization-for-text-data-29fb1b96fb6a>