

# INTRODUCTION

## 1.1 OVERVIEW :

Sentiment Analysis is the automated process of analysing text data and sorting it into sentiments positive, negative or neutral. Performing Sentiment Analysis on data from Twitter using machine learning can help companies understand how people are talking about their brand. As we all know, we can use the twitter data to perform analysis and figure out the trends and sentiments of the current world. With more than 321 million active users, sending a daily average of 500 million Tweets, Twitter allows businesses to reach wide set of 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.

From the above context, we can think of why not use the data to predict the current sentiments and trends with regards to a particular topic of interest, which can be used for many purposes like to cater the needs of the people and regarding the relevance of the product/topic in the online community. By this project, we can filter out the genuine sentiments of the people from such data, helping various organizations/communities to come up with a valid conclusion that will project towards to a positive growth. And, tweets can serve as a powerful weapon in this scenario. The analysing and visualizing of these " sentiments" from the twitter pool can serve as a powerful tool to study and understand the behaviour of the target audience ( geographically /timeline/ community based etc) and know what were the parameters that lead to a specific conclusion.

## 1.2 PURPOSE :

This proposal is for the sentiment analysis of Indians after the announcements pertaining to the extension of lockdown, from relevant #tags on twitter and build a predictive analytics model to understand the behaviour of people if the lockdown is further extended. This is also followed by visualization of the prediction and then a dashboard is also displayed with regards to the people's reaction and emotion towards the government's decision. Followed by the reason why that specific conclusion arrived and identifying the triggers and parameters for the final result.

The speciality of the proposed solution is that we are utilizing up data extracted from the internet and coupling it with NLP techniques supported by Deep learning (Deep NLP). This helps to make our model more robust in functioning, thereby giving it a dynamic behaviour thus giving better outputs as compared to other methods of analysis. Developing the solution using such robust Deep learning techniques makes further room for improvement and future proofs the solution. And this also makes it easier for being applied to other use cases as well. Moreover, we can also easily scale the model if needed.

The sentiment analytics can be used as a useful tool in promoting and targeting the likes and dislikes of the people. This can be used to derive personal interests and the person's preference in matters, so that the company or the organization can strategize towards developing a more profitable & reliable solution and also to serve

the customers better. Twitter sentiment analysis, in particular, can be used to synthesize the opinion of the public, towards a particular product and help make the product better. From a business point of view, this is a stringent requirement, as the customer are the once who should be kept happy. In one of many cases, this model can be used to remove unwanted/spam reviews/opinions that are presented in public forums. There is an ongoing conflict for the removal of " unconstructive" or " baseless " opinions between companies and the customers. Such observations hinder the day to day development of the products presented by the companies and therefore an efficient solution is required.

## LITERATURE SURVEY

### 2.1 EXISISTING PROBLEM:

The challenge here is to detect the right set of emotions conveyed by the public in the matter being investigated. This is very difficult in most cases, as the emotion meant by the users cannot be accurately measures or considered for the purpose of analysis due to lack of robustness of such existing model to predict the emotions the tweet is trying to convey on a whole by looking into the statements/tweets as a whole and also because most models for the matter would only segregate among two or three classes of tweets namely happy, sad and neutral. This can cause invalid statistics and can fiddle with the evaluation process. Moreover, the models based on regular Neural network frameworks need a lot of data to be trained on and requires tons of epochs and fine tuning to give out decent accuracies while predicting the sentiments.

The RNN-CNN based models literally needs hours, if not days to get accurate predictions which can be actually taken into account for the analysis. Accurate classification and the proper approach for such complex problems cannot be handled by the common methods of predictive analysis seen in abundance.

### 2.2 PROPOSED SOLUTION:

The solution is to make use of the State-Of-The-Art machine leaning models, dedicated to tackle problems with ease. We proposed a solution building upon Facebook's own NLP technology, RoBERTa(Robustly Optimized BERT Pretraining Approach ), which used the technology of the BERT model and build upon it to improve the predictability, at the same time, reducing the complexity of a regular neural-net system, along with drasting cut-down on training time.

We are proposing to use the pre-trained distil-roberta model and wrapping the neural-net around our toplevel API of pytorch to build something which can give us one of the most robust configuration that we can possibly get.

With the robust backend running, we have added the functionality of tracking live tweets, covid-19 regional alertness and live case tracking within the user interface by integrating the scripts necessary.

The highlight of the frontend design is the interface for custom tracking of stats and opinions from the twitter community about the current trends and environment. The mainframe also provides dedicated analysis of covid-19 with respect to the lockdown and other important govt. decisions being made accross india.

The analysis of the data is displayed on the page pertaining as the results, with word cloud, bar graphs and other tools of visualization.

## THEORITICAL ANALYSIS

### 3.2 SOFTWARE DESIGNING

Here we have factored our project in 3 phases : **TWITTER DATA EXTRACTON**(including preprocessing) , **SENTIMENTAL ANALYSIS USING AN NLP MODEL**(RoBERTa) and **VISUALIZING IT** . The user will be prompted to enter the parameter ,that he/she is willing to access, in the main UI. The values entered will then be used to extract the pertaining data and is then forwarded to the model to generatre the sentiment analysis . The output is then displayed in the dashboard along with some information about COVID-19's status in India. These 'phases' are handled by seperate independent scripts , wherein a "main.py" is treated as the file where all these scripts are imported and the functons that are defined in them are applied in it.

#### 1. TWITTER DATA EXTRACTION : -

In this phase a python script is written to use "tweepy " library . It basically queries over twitter based on the keyword and number of tweets specified in the search bar. But these extracted tweets contains a lot of unnecessary data henceforth it becomes a major concern to clean and preprocess this data before loading it into the model.Followed this preprocessing , the clean is now ready to be given to the model for analysis. -- ' CLEANER DATA PROJECTS TO CLEANER OUTPUT '

#### 2. SENTIMENT ANALYSIS USING AN NLP MODEL :-

With the recent developments in Natural Language Processing our team concluded to use RoBERTa model for extracting emotions from textual data.The RoBERTa model was proposed in [RoBERTa: A Robustly Optimized BERT Pretraining Approach](#) by Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov. It is based on Google's BERT model released in 2018.It builds on BERT and modifies key hyperparameters, removing the next-sentence pretraining objective and training with much larger mini-batches and learning rates. We managed to train this model with an accuracy of 94.79%. RoBERTa is

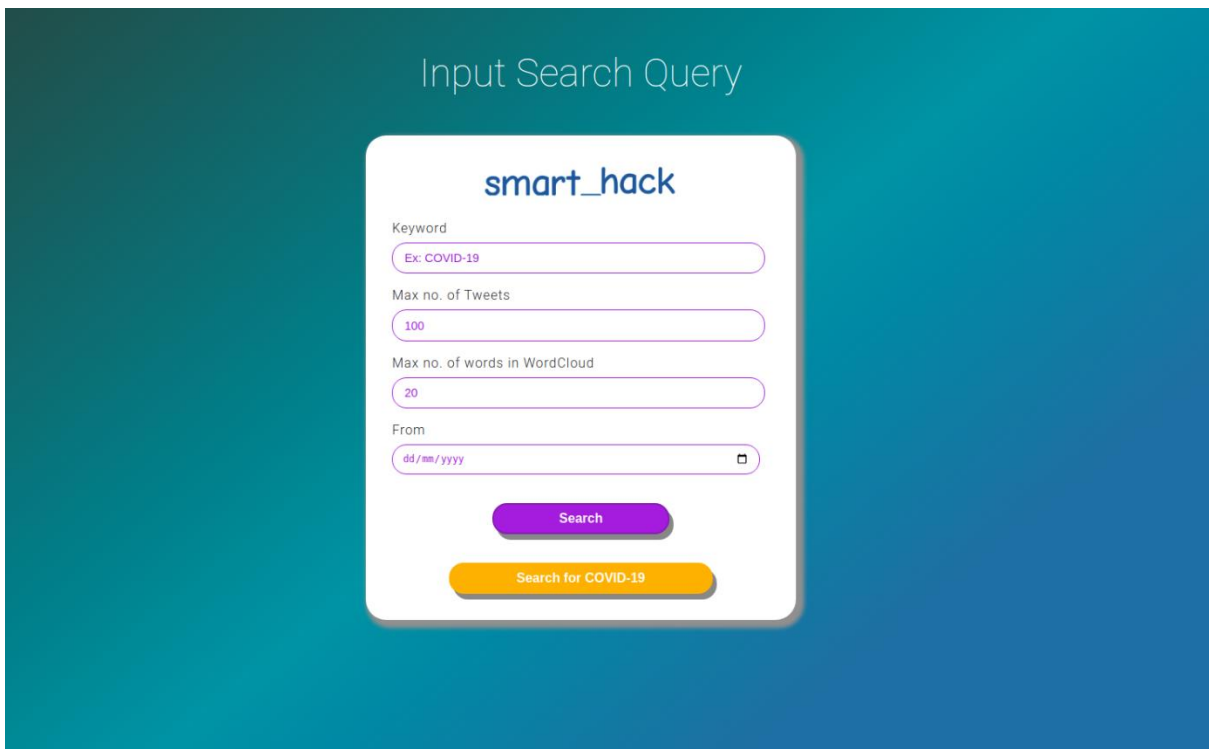
developed based on BERT, they share lots of configs. Still it performs better than BERT because of the following adjustments :

- Bigger training data (16G vs 161G) .
- Using dynamic masking pattern (BERT use static masking pattern).
- Replacing the next sentence prediction training objective.
- Training on longer sequences.

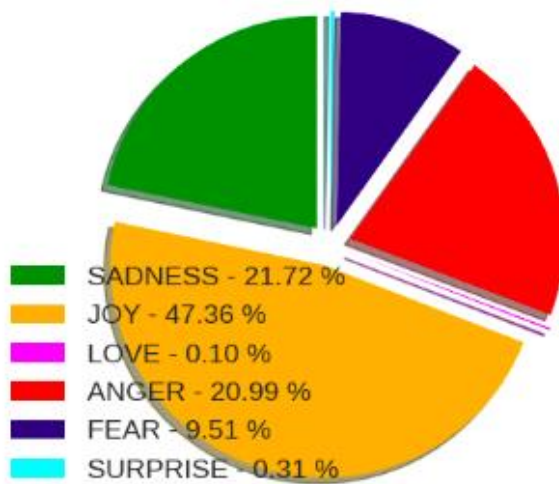
This model has the caliber to classify textual data to SAD , JOY , ANGER , SURPRISE, LOVE and FEAR emotions . The generated analysis is then displayed on a PIE chart as well as on a bar graph using matplotlib.

### 3. VISUALIZING :-

A form is displayed in the main page where the user can input ' Custom Keyword' , ' Maximum Number of Tweets ' , ' Maximum Number of Words in Wordcloud ' and ' Date ' . Followed by this search indices there are 2 search options ' **Search** ' and ' **Search for COVID-19** ' . ' **Search** ' will give the sentiment analysis for the custom keyword that has been provided by the user and specified by other parameters while ' **Search for COVID-19** ' will give sentiment analysis for **COVID in India** based on other parameters ( by default number of tweets is 100).



*A Visual representation of the main page*



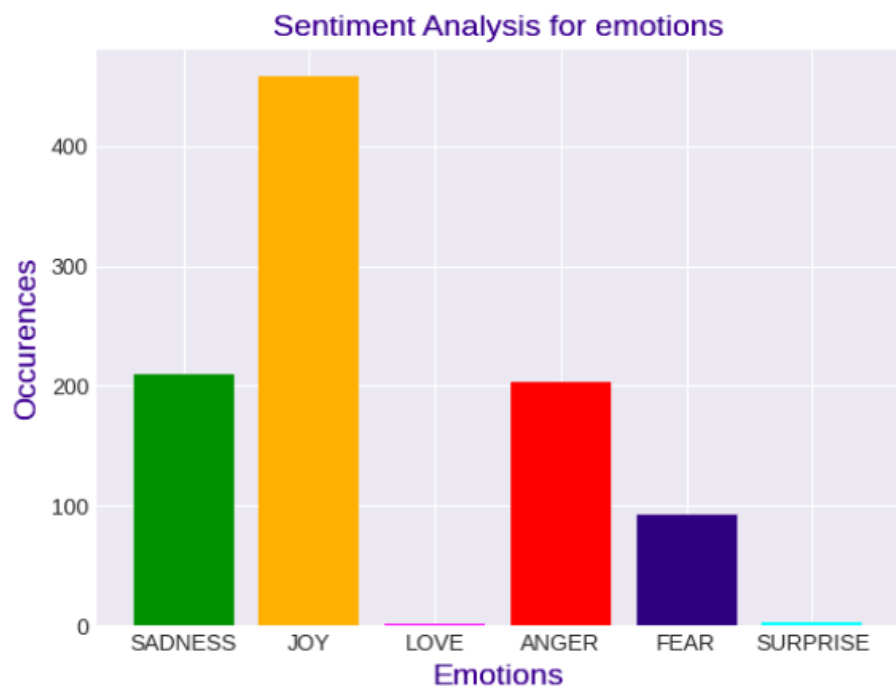
PIE - CHART : Its depicts the percentage of the emotions that has classified by the model from the data

PIECHART WITH PERCENTAGES

BAR CHART : It basically projects the sentiments on a graph.

### Twitter Users Behavior

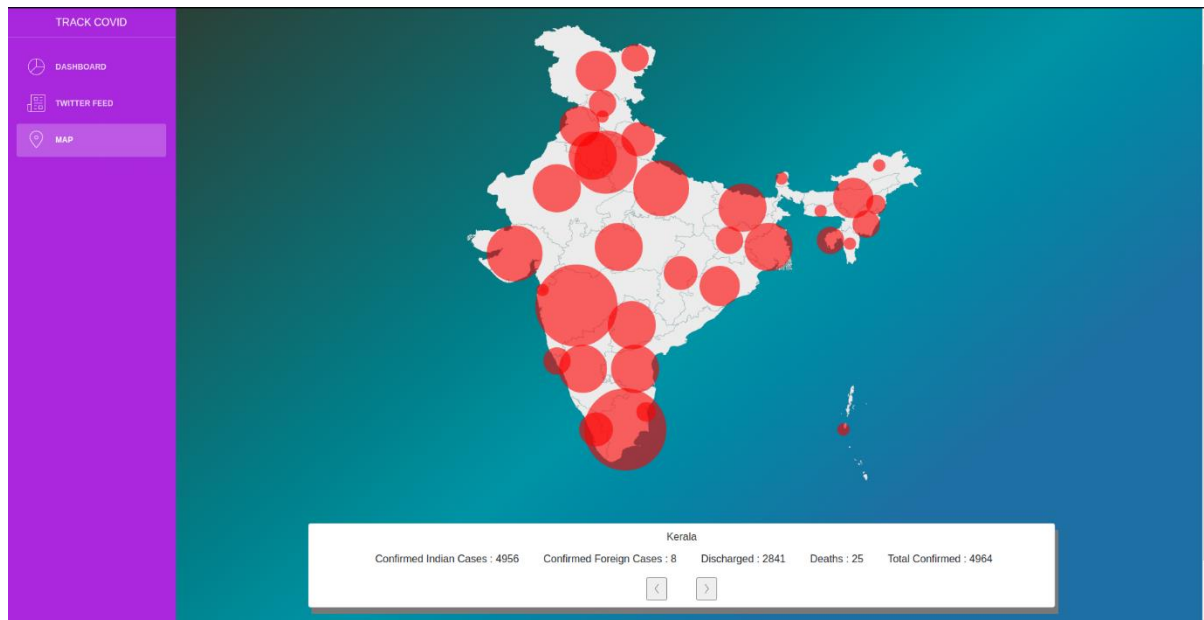
Analysis from last 4 days





---

**MAP :** A map depicting the spread of corona virus across India . The RED circles shows the degree of spread based on confirmed cases , across 35 union-territories/states.



## EXPERIMENTAL INVESTIGATIONS

- After many iterations of testing and inferring the data that we got, most of the "REACTIONS" that we got from Twitter regarding COVID 19 in India is mainly composed of "JOY" and "SAD". It can be inferred that netizens are happy about how our Govt is tackling this epidemic, discussing about their lockdown activities and at the same time a large portion is sad because of the way COVID -19 is affecting their daily lifestyle especially that of daily wage workers.
- The map works efficiently in displaying day to day COVID spread in India.
- Higher the number to tweets extracted, better will be the analysis.

---

## ADVANTAGES OF THIS PROJECT :

- This project not only gives a sentiment analysis of COVID 19 in India , but it can also give sentiment analysis of any feasible word for querying.
- Unlike general sentiment analysis , this gives out analysis based on 6 different emotions .
- Our visualization dashboard also gives a WordCloud for getting to know trending words amongst the tweets we scrapped .
- There also a section which gives us day to day details about newly confirmed cases , recent deaths , cumulative cases , total recovered and active cases .
- A twitter feed is there which gives us latest tweets made by @COVIDNewsByMIB for letting us know what all steps are being taken to combat the virus .\
- The map feature shows a general view about how the virus is branched among states and union territories . Thers also a small form below this map which can be used to access information regarding confirmed indian cases ,confirmed foreign cases , total recovered and total death of corresponding states and union territories .

## DISADVANTAGES OF THIS PROJECT :

- The data which is being extracted from twitter is in its raw form , so for that it is essential to preprocess it which takes some time .
- The data which is being displayed for daily COVID status is from an API, so if any problem occurs in accessing the internal server of the database or value updation in the database , then the displayed data on the corresponding section becomes unreliable.

## RESULT :

The project is able to scrape , preprocess and analyze the tweets made by the users in a reasonable period of time. The output is displayed successfully on the main page along with the update COVID status in India. Overall this project helps the user who are wanting to know more about people's views about COVID , status about COVID and news about COVID.

---

## APPLICATIONS:

This project as a prototype perfoms well enough in :-

1. Generating sentiment analysis of COVID 19 in India ( apparently the location constraint can also be changed by adjusting the keyword ) .



2. As mentioned earlier the scope of this project not only limits to 'COVID 19' but can be expanded to any special keyword for scraping data.
  3. The data which is being displayed in dashboard can be easily accessed and understood by any Entrepreneur / Researcher/ Enthusiast / Student or anyone who requires sentiment analysis for their Project / Knowledge / Research / Documentation .
  4. If anyone wishes to know latest COVID status in India, for any state or union territory then he/she can access it at ease along with a dependable representation of the spread , projected on a map of India .
  5. Recent progress by the government can be tracked by visiting the the twitter feed widget from dashboard which linked with the official twitter account of @COVIDNewsByMIB.
  6. Based on the sentiment analysis the user can strategize and can come up with a tweak of an application , a scheme for an organisation , a conclusion of a decision , an insight about a personality etc. from a target community .
-