

**CSE 535 Information Retrieval  
Fall 2020  
University at Buffalo**

**PROJECT 4  
DISSECTING TWITTER DATA TO ANALYZE  
GOVERNMENT AND PUBLIC ATTITUDE  
TOWARDS COVID GOVERNANCE - HOOGL**




**SUBMITTED BY:  
THOMAS ROSSETTI(50140597 )  
MANASA RAO CHAKUNTA(50338196)  
PREETHI THOTA(50336834)**

# 1.OVERVIEW:

Our project is about “DISSECTING TWITTER DATA TO ANALYZE GOVERNMENT AND PUBLIC ATTITUDE TOWARDS COVID GOVERNANCE”. We have built a solution in the form of a website using AWS Solr, Flask, Html, Css, JavaScript, and BootStrap to analyze the government and public attitude towards COVID governance. We have calculated and used the influence score to improve tweets ranking. The data was crawled from twitter using twitter search API. We have collected data for three languages English, Hindi and Italy, three countries USA,INDIA and ITALY and 9 POI’s. We have written queries to collect the data on government and people’s reaction towards government policies on covid19. The crawled data was then processed using python script to extract the required information such as text, hashtags, user names , language, URL etc.

The processed tweet collection is then indexed in Solr. We developed a UI which we named as Twoogle(derived from the names of Twitter and Google) such that Solr acts as the backend of it. We also integrated the analysis of search results to make it more interesting for the user. We also added the advanced search feature to allow the user to filter the tweet, news articles and graph analysis results based on the language. We have taken reference from Google and named our project as “**HOOGLE**”. We have followed the UI policy “**The Simpler The Better**”. Our home page is below:

[Hybrid](#) [Home](#) [Insights](#)



## 2. TECHNOLOGIES USED

For the website of our project we used a combination of flask and bootstrap. Flask is used as the framework for our backend which handles the routing and requests between our individual pages as well as speaking to the different APIs. We tried out a few different APIs throughout the development of this project. The two main ones being the Solr API which lets us access our Solr database. To make using the Solr API easier we employed the help of a package called pysolr. We used the Google News API as well which we accessed using a package called GoogleNews. For the front end we used bootstrap. Bootstrap is a combination of HTML, CSS, and JavaScript. This allowed for us to add cool and different things such as dynamic charts very easily. Our charts for our statistics and insights were generated using the help of Google Charts. To deploy our site to the web we used AWS.

## 3. IMPLEMENTATION STEPS:

### Requirement 1:

#### Influencer Score:

- We have generated influencer scores for each tweet, based on their potential of influence.
- We could say a tweet has more influence if it has more number of likes or more number of retweets.
- Also, we could say a tweet has more influence if the person tweeted it has more number of followers.
- So we have taken followers count also in consideration and we have taken formula as below:

$$\text{max}((\text{retweet\_count} + \text{followers\_count}) / \text{max\_followers\_count} \text{ and } (\text{likes} + \text{followers\_count}) / \text{max\_followers\_count})$$

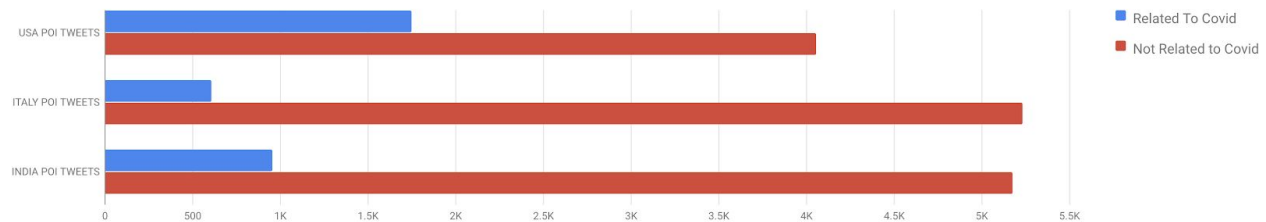
- We normalized the value by dividing with maximum followers count among all poi's.
- We have added a checkbox sort by influencer score in the UI where if you check it, the tweet results which will be getting are based on the influencer score ranking. The one with the highest influencer score is displayed at the top.

## Requirement 2:

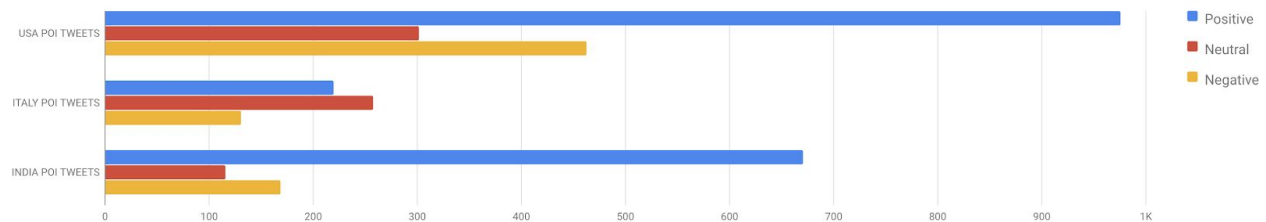
- **Covid and Non-Covid Curve:** For requirement 2 we wanted to see if there was a correlation between the number of POIs coronavirus related tweets per country and the number of active cases for that country. To do this we found the covid related tweets for each POI in the United States, Italy, and India. We then found the sentiments towards these tweets to see how people reacted to the POIs taken towards the situation.

Here we wanted to see if there was a correlation between each countries total covid realated poi tweets and how they were recieved by the viewers and compare it to the covid curve of each country to see if there was an effect

Covid VS Non-Covid Tweets



Reception Towards Covid Related Tweets



- **Topic Analysis:** The topic analysis for each is made from tweets json file by removing stopwords and by stemming each word in the tweets and then we performed topic analysis of all the tweets to extract main topics people are concerned about using **LDA model** and 10 topics are generated. Thus by making the topic as a faceted field to make tweets that included that topic as important and sort them based on topic. The following are the topics we got for the entire search using LDA.



**Requirement 3:** Requirement 3 is all about displaying the information we have collected/created from both requirements one and two. We used google charts to show this data in a visual and meaningful way.

- We have inserted **News Articles** using Google News API for python. We have added a feature when you click on the link button we will be redirected to the webpage which has that article.
- We have also implemented, when we click on the tweets which we get according to the search query we will be able to redirect to the tweet page in twitter.

# News

## Indian hospitals desperate for oxygen as coronavirus cases top 5 million

Reuters

MUMBAI (Reuters) - Coronavirus infections in India surged past 5 million on ... Modi's cabinet, Nitin Gadkari, also tested positive for the coronavirus infection, ...

[Link](#)

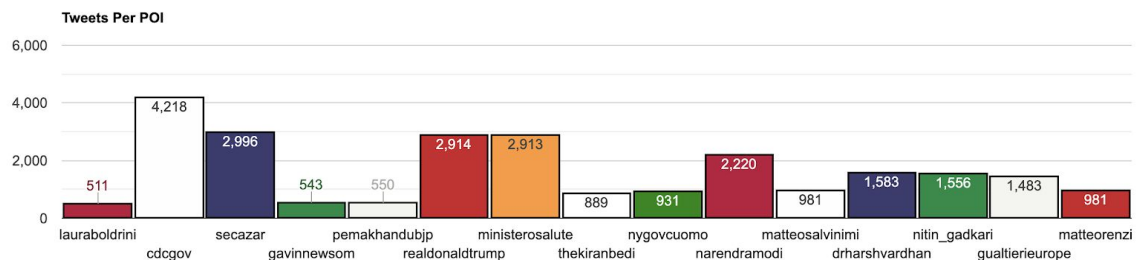
## Coronavirus India Live Updates: India's COVID-19 Tally Cross 51 lakh, Active Cases Cross 1 Million

NDTV

On Wednesday, Union minister Nitin Gadkari has said that he tested positive for coronavirus. In the mandatory COVID-19 tests for parliamentarians conducted ...

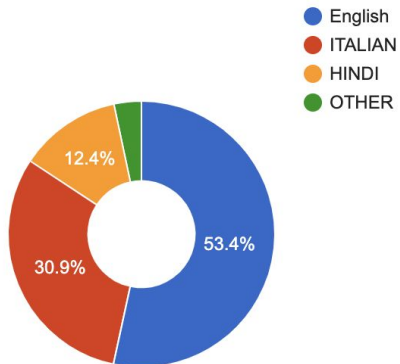
[Link](#)

- The below plot shows the number of tweets posted by each Person of Interest(POI) in the entire collection of tweets.

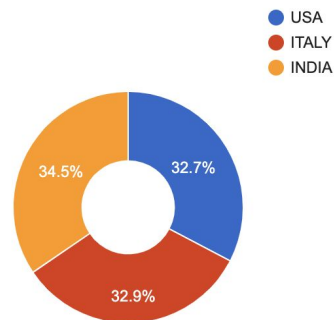


- The following plots show the POI tweets per language in percentage and their tweets from each country.

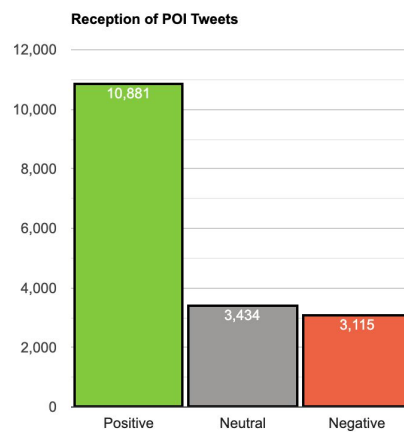
POI Tweets Per Language



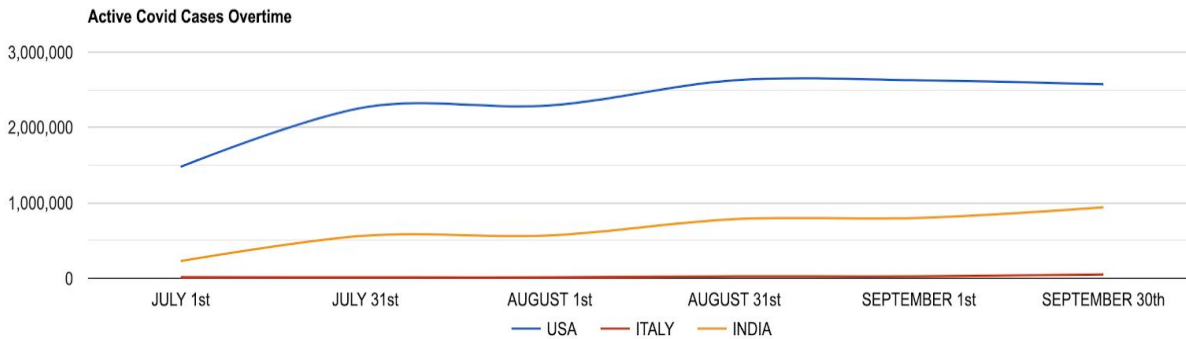
POI Tweets Per Country



- The following plot shows the sentimental analysis of all the poi tweets.



- The following graph shows the covid curve in respective countries.



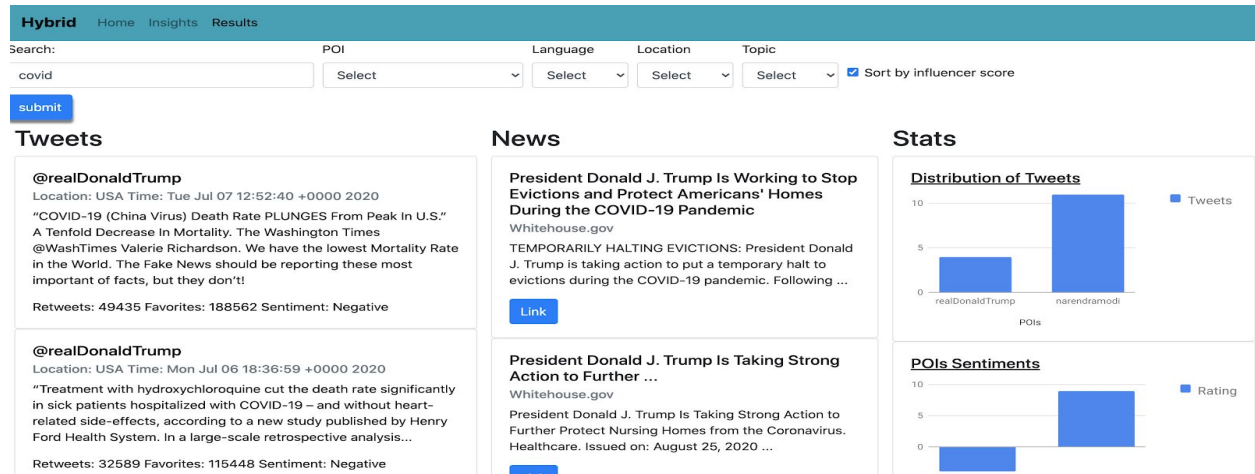
- **Sentiment Analysis:** We have done Sentiment Analysis using NLTK library. We have used SentimentIntensityAnalyzer from nltk.sentiment.vader. We have created the object of SentimentIntensityAnalyzer and passed the each tweet text through polarity\_scores() method using SentimentIntensityAnalyzer object so it gives the polarity Sentiment of the particular tweet i.e either positive, negative or neutral.

We have calculated tweet sentiment for all the tweet text and added sentiment as field in all the tweet files through which we can access Sentiment of a tweet to the front end.

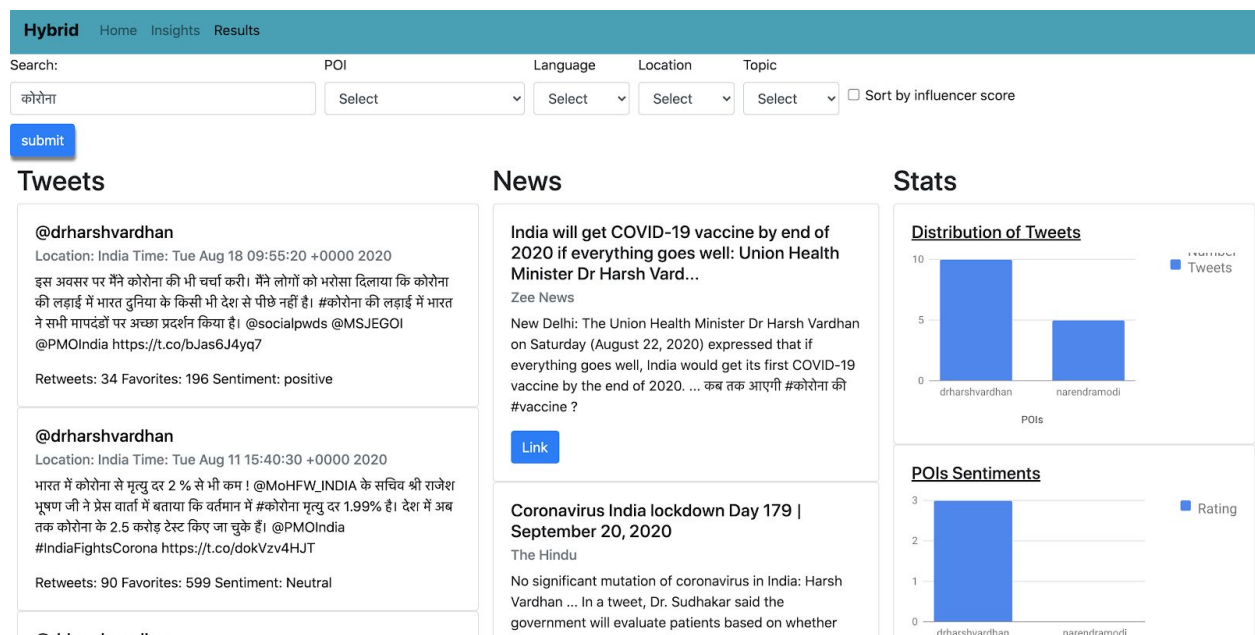
**Requirement 4:** For our faceted search functionality we used your inputs and the information you provide in the drop downs to create a search query. We then send that query to our Solr server (<http://18.234.178.228:8983/solr/#/IRF20P4/core-overview>) using pysolr. The tweet data returned is then parsed and used to create the makeshift tweets that you see on your screen. This tweet data is also used in creating some charts so you can better understand the results of your search query. These charts include the dispersion of tweets between the returned POIs, the sentiment toward the returned POIs, as well as the influencer score for the returned POIs. To get our related news we take in your search query along with the names of the POIs that are to be returned to create a search query of our own that we will send to **Google News** (<https://news.google.com/>) using a google news package for python. The results returned are then parsed and displayed in the news tab.



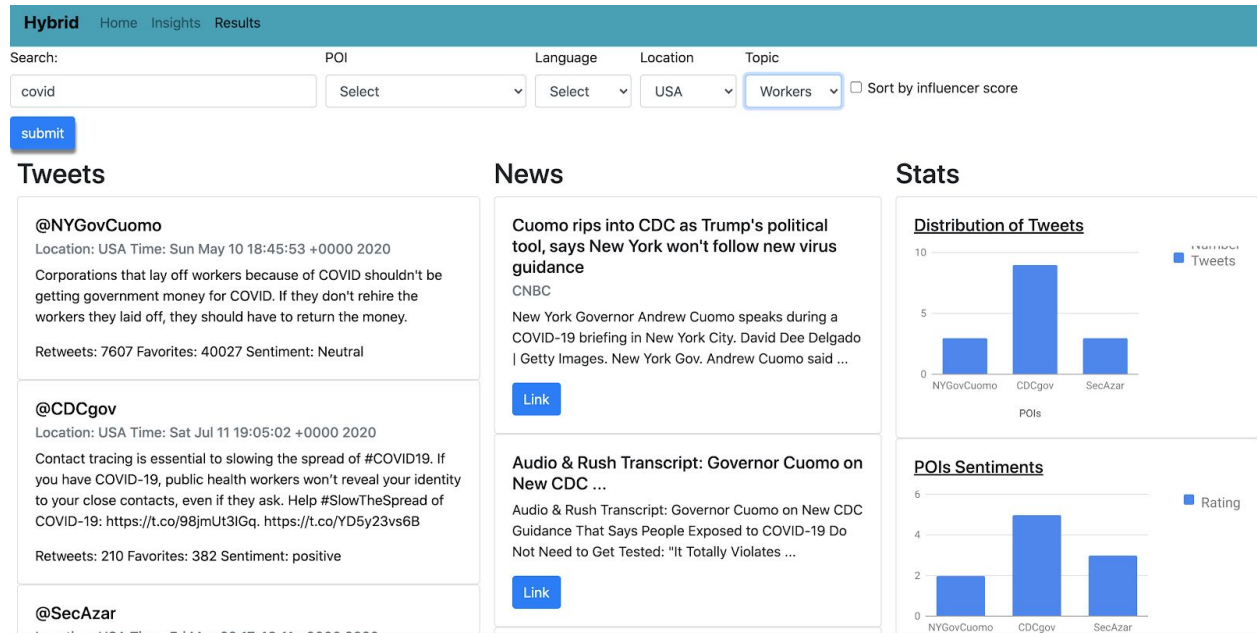
## Ranking based on influencer score:



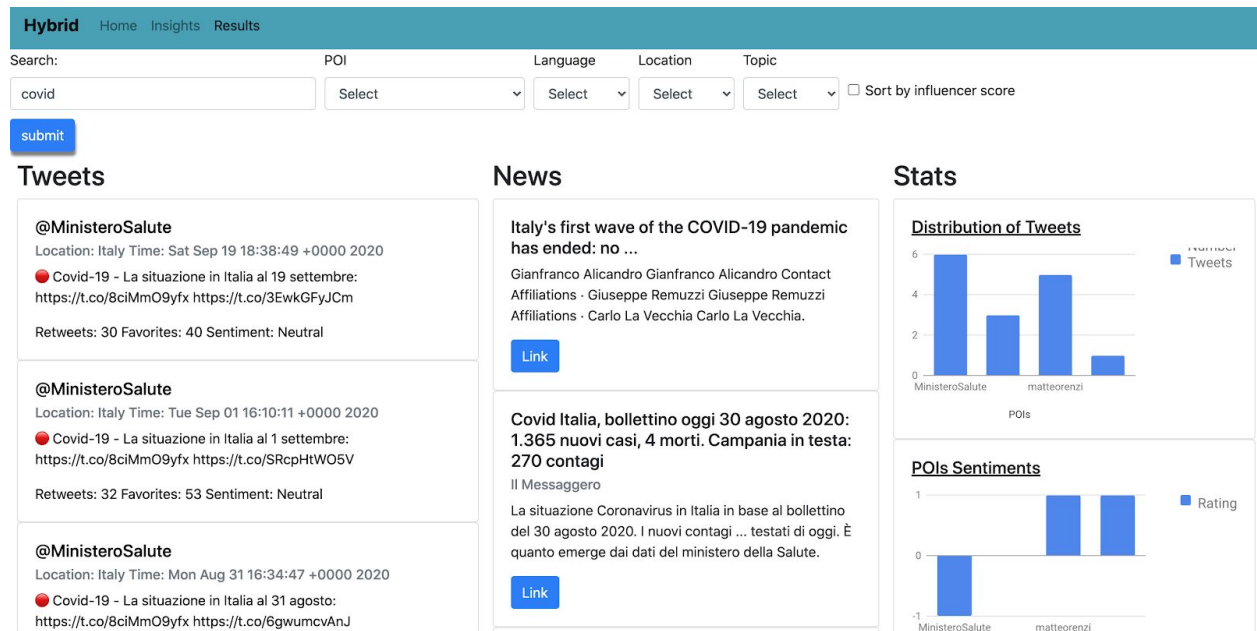
## Hindi:



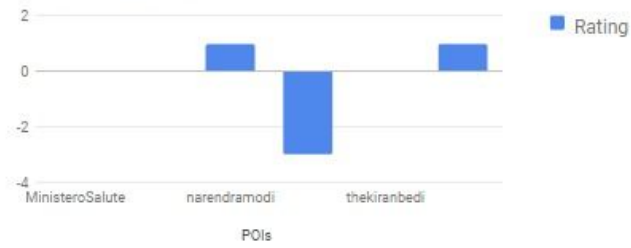
## For query covid, location filer: USA and topic:workers



## Italy:



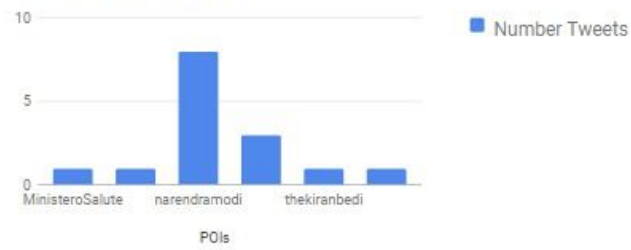
### POIs Sentiments



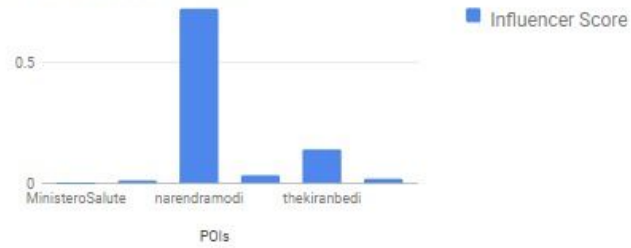
### News Publishers



### Distribution of Tweets



### POI Influencer Scores



## TEAM CONTRIBUTIONS:

Team Member	UBIT Name	UBIT ID	TASKS
Thomas Rossetti	tjrosset	50140597	UI , Graphs and Report
Manasa Rao Chakunta	manasara	50338196	Influencer Score, sentiment Analysis and Report
Preethi Thota	preethit	50336834	Topic Analysis, Crawling Tweets, Preprocessing of Tweets and Report

## DEPLOYED URL:

<http://3.80.46.84:8080/>

## DEMO LINK:

<https://buffalo.box.com/s/kicm7eptjd53obfzpikxy3e9ieoiljt>

## References:

- <https://www.nltk.org/howto/sentiment.html>
- [https://en.wikipedia.org/wiki/Latent\\_Dirichlet\\_allocation](https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation)
- <https://news.google.com/>
- <https://pypi.org/project/googletrans/>
- <https://developers.google.com/chart>
- <https://flask.palletsprojects.com/en/1.1.x/>
- <https://getbootstrap.com/>