Social Media Analytics on News Articles

Hammer!

INFO 7390: Advances in Data Science & Architecture, Spring’18

Under , Prof. Srikanth Krishnamurthy, and TA Tushar Goel

Team Members:

1. Jai Soni
2. Pramod Nagare
3. Saurabh Kulkarni

Overview:

With so many different news from around the world, it takes effort to keep track of a person, a place, about particular sport etc. and or anything that a user wants to follow.

Goals:

To provide user with a fast and easy way to know what news are going on about a search query and people's reaction to the news on social media (Twitter).

1. Build a Sentiment analysis model that can classify the sentiment of a sentence into Positive, Negative and Neutral. Use various classification models, and deploy the model in AWS S3 for further process.

2. Build a user friendly web application, where a user can enter a query and get the summary of sentiments of various news trending on social media.

3. Get location-wise sentiments for the query by the user.

Data:

Data for training the classification model is taken from Kaggle datasets :
https://www.kaggle.com/kazanova/sentiment140/data

Process Outline:

Part 1: Deploying the model

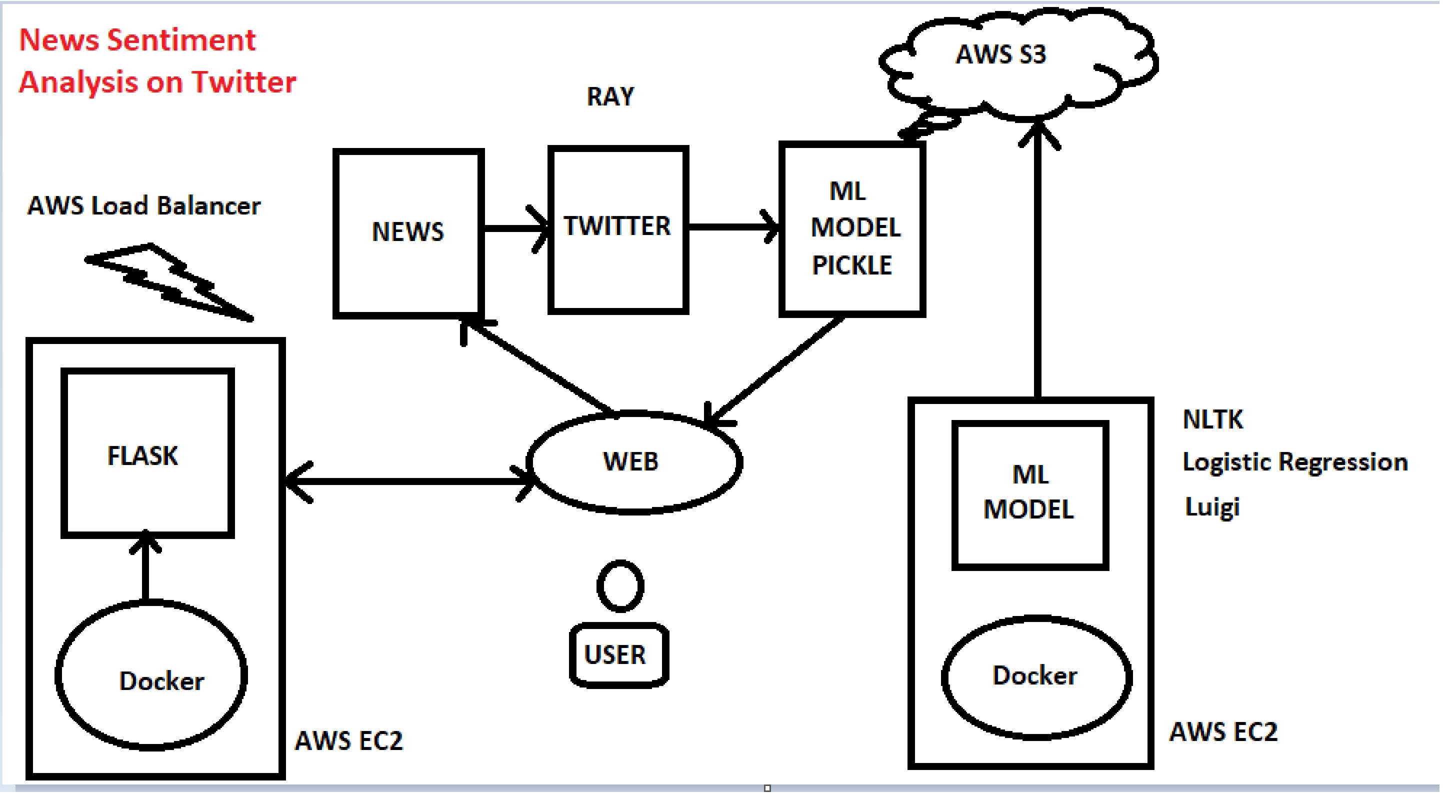
1. Data cleaning
2. Exploratory Data Analysis
3. Study of Supervised approaches and select the best model for prediction
4. Data pre-processing
5. Model training and summarizing accuracy
6. Building data pipeline using Luigi
7. Uploading pickled model to AWS S3
8. Dockerizing entire pipeline

Part 2: Deployment Architecture

1. Get pickled models from AWS S3
2. Get user input from flask web app
3. Scrape news articles related to the query
4. Get tweets for the news articles
5. Compute sentiment analysis through the pickled models and Vader Sentiment analyser
6. Summarize and display various news articles along with there sentiments
7. Get coordinates for the tweets locations and Plot these locations on geographical maps
8. Dockerize entire application
9. Deploy the application on AWS EC2

Part 3: Scale the application for running multiple requests simultaneously using AWS Auto scaling and Load Balancer

Deployment Architecture



Explanation on Implementation

PART 1:

1. We used a dataset from Kaggle
2. We have implemented a docker container for building a classification model using NLTK and Logistic regression. We got AUC score of 87% when using our Logisitc Regression model
3. This model implementation is supported with the Luigi pipelining.
4. This docker container corresponds to the part 1 of the application development.
5. After training and testing the above model for sentiment analysis on twitter feeds we have uploaded the pickle files to the AWS S3 bucket.
6. As we have dockerized the part 1 implementation, we can run the script from any machine which supports docker.
7. For a full stack application development we will leverage the cloud infrastructure with AWS EC2 instance to run the docker container and get the model setup for twitter feeds sentiment analysis.

PART 2:

1. Now we have our machine learning model setup on cloud which can be accessible for predicting real time twitter sentiment analysis.
2. We have another docker container for running Flask application which will be used as a user interface to make the sentiment analysis on the context given by user.
3. This flask application get the input from the user and query to the news api to get the relevant and recent news.
4. We have used ALYEIN news api for getting news headlines.
5. The list of news headline will then forwarded and queried against the Twitter appi, Tweepy.
6. This will list out the tweets related to the each news article.
7. Then the complete list of tweet feeds will be executed against the machine learning model accessible from AWS S3 bucket to get the sentiment for the tweets.
8. In accordance with our ML model we have also used the Vader api for Twitter sentiment analysis.
9. The sentiment analysis of the above tweets then plotted in bar and pie charts for better visualization of the impact of that particular news article which is in context with the user input.
10. At the web interface level user will be able to analyse the overall impact of the news on the people to gather the make further insights and planning.
11. Also, as we are running the docker container for this application, we can use any machine which supports the docker.
12. This docker image is made execute on the AWS EC2 instance to make available through generalize web url for all the user, irrespective of docker support is available or not for that user.
13. As, many users can now access this url/application from anywhere, we need to implement the load balancing for the proper hosting of the application.
14. To take of the above scenario we have implemented the AWS Load balancer to auto scale our infrastructure in real time.

Deployment Details:

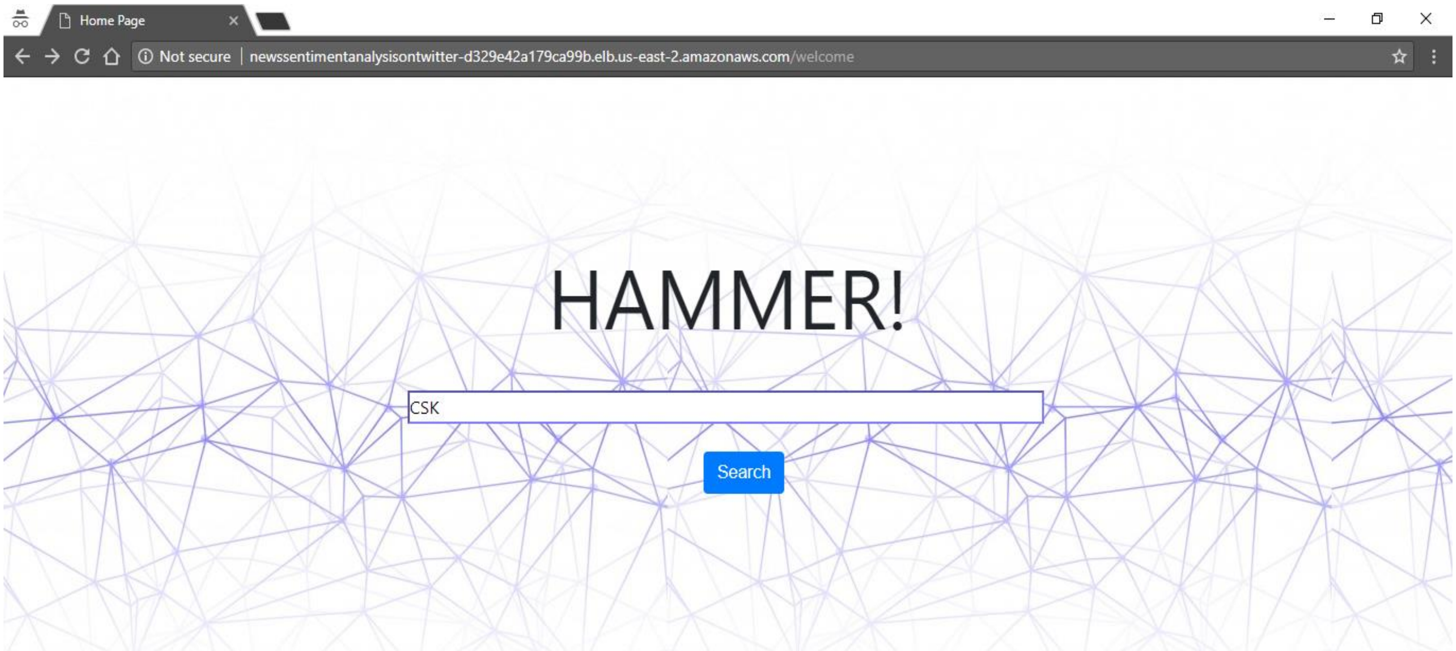
1) Language: Python, HTML, JS, CSS
2) Pipeline: Luigi
3) Container: Docker
4) Cloud Tools/Platforms:AWS (Amazon WEb Services) EC2,S3
5) Tools for Analysis: Jupyter, Bokeh ,HighChart.js, WordCloud
6) API's used: Aylien-news-api, Tweepy, Vader sentiment analyzer, Bokeh, GeoPy
7) Considered Ray for multi processing and fast processing

Known Exceptions occuring on application:

1. On multiple simultaneous hits, Unpickleing error occours sometime.
2. Tweepy Error for Exceded Rate Limits
3. **Homepage:**

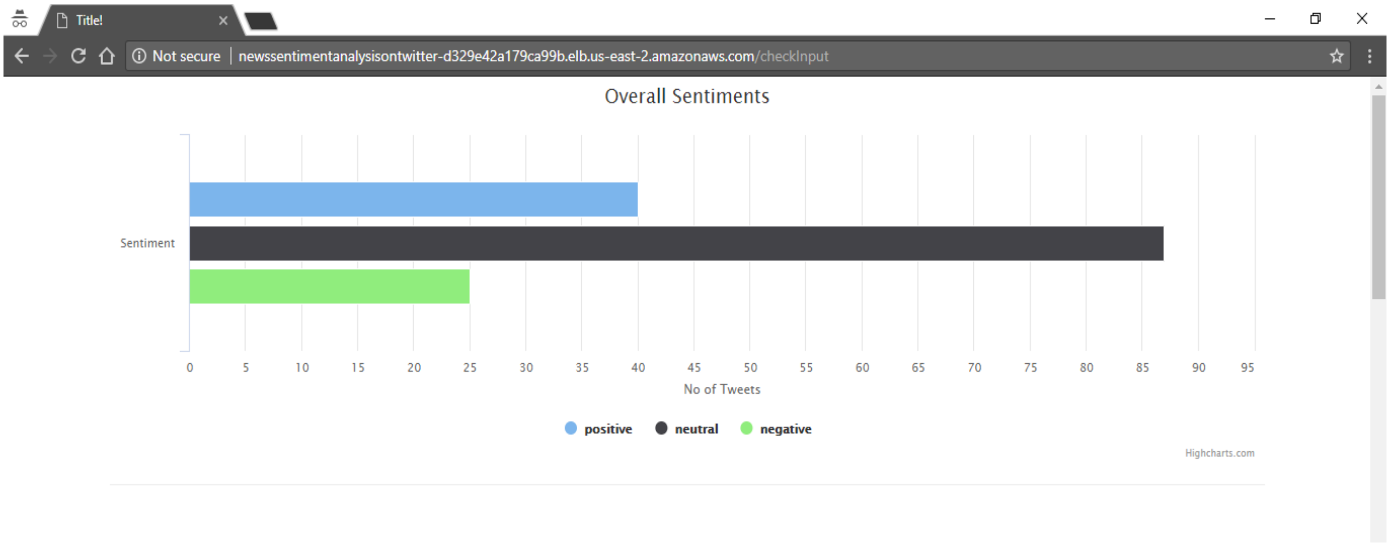
APPLICATION Details:

User Input


Home Page
Not secure
newssentimentanalysisontwitter-d329e42a179ca99b.elb.us-east-2.amazonaws.com/welcome
HAMMER!
CSK
Search

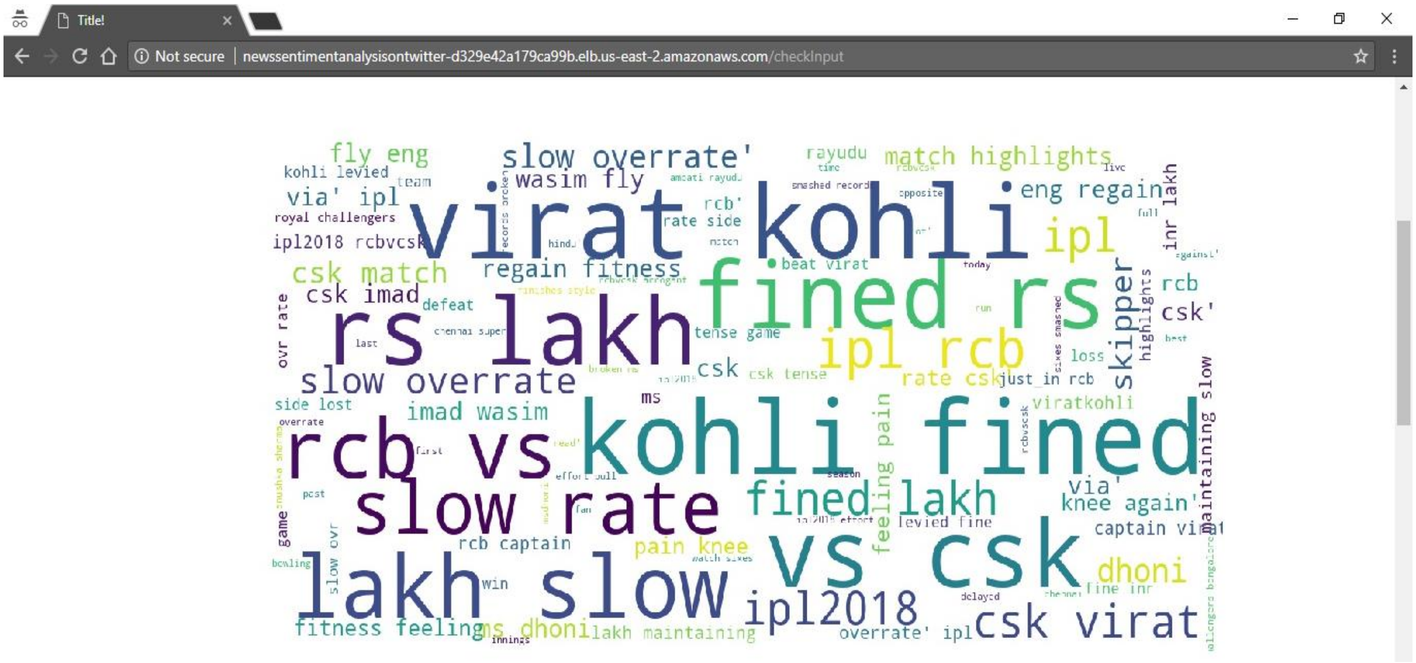
Result 1:

Bar plot for overall sentiment analysis



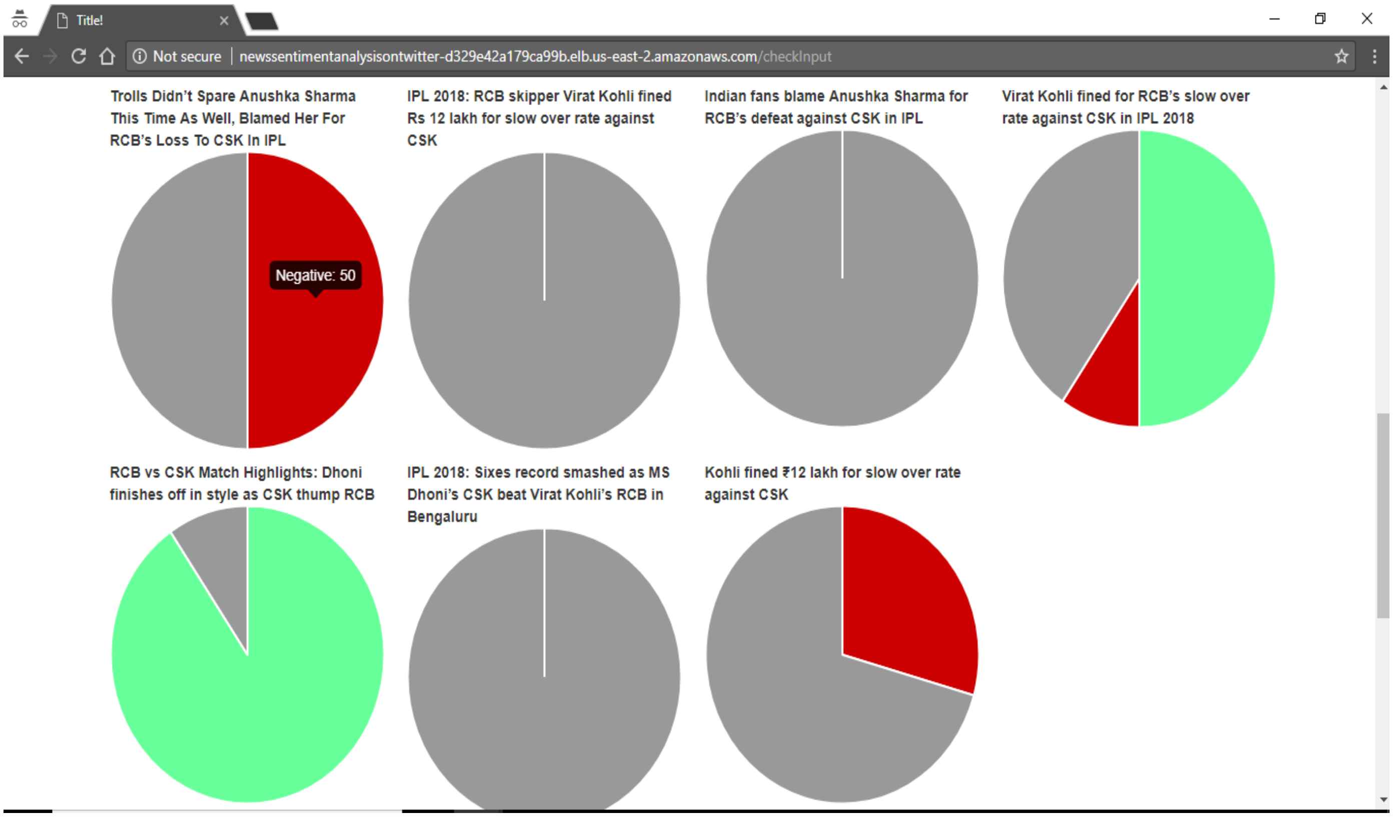
Result 2:

WordColud



Result 3:

Pie Chart for individual news article


Title!
Not secure
newssentimentanalysisontwitter-d329e42a179ca99b.elb.us-east-2.amazonaws.com/checkInput
Negative: 50

Trolls Didn't Spare Anushka Sharma This Time As Well, Blamed Her For RCB's Loss To CSK In IPL

IPL 2018: RCB skipper Virat Kohli fined Rs 12 lakh for slow over rate against CSK

Indian fans blame Anushka Sharma for RCB's defeat against CSK in IPL

Virat Kohli fined for RCB's slow over rate against CSK in IPL 2018

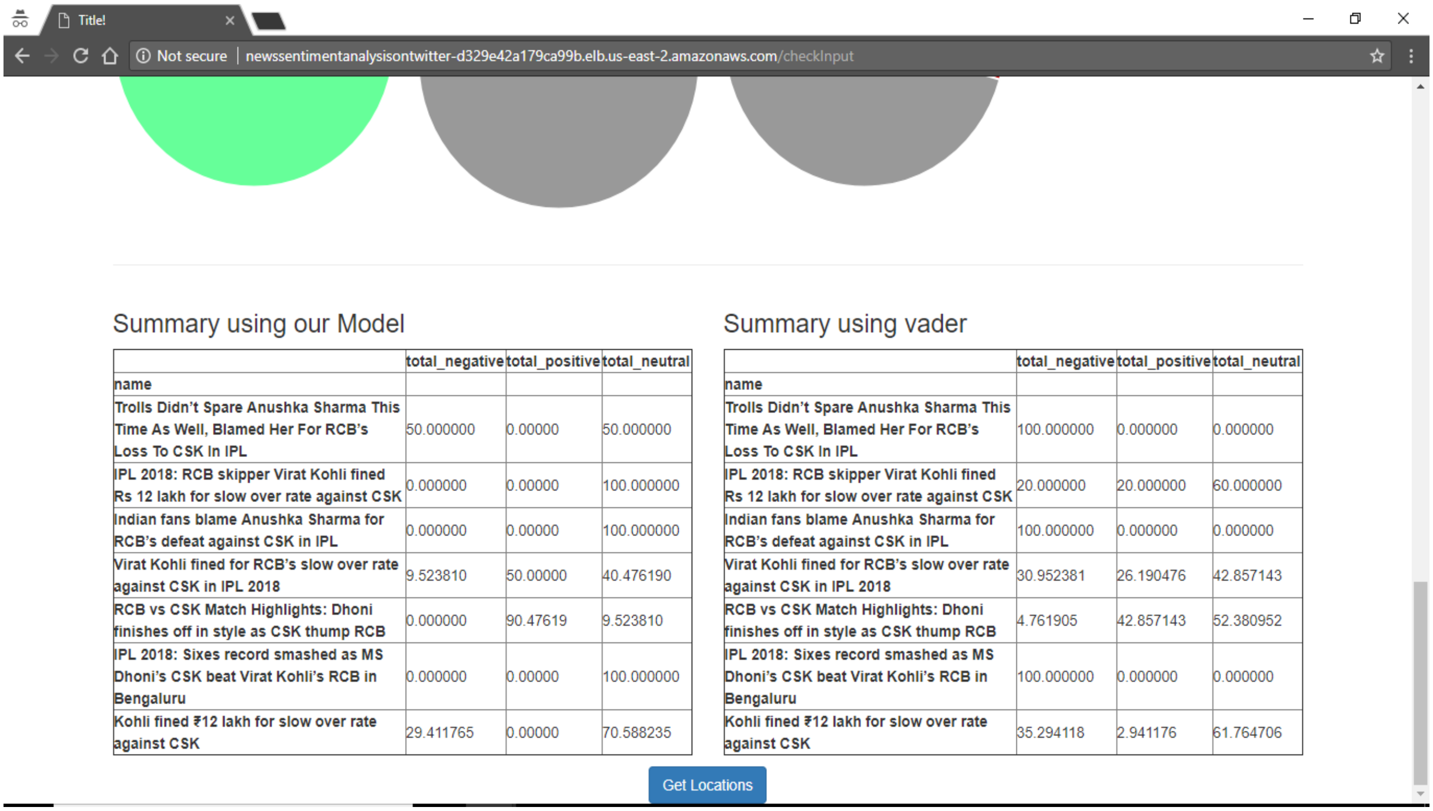
RCB vs CSK Match Highlights: Dhoni finishes off in style as CSK thump RCB

IPL 2018: Sixes record smashed as MS Dhoni's CSK beat Virat Kohli's RCB in Bengaluru

Kohli fined ₹12 lakh for slow over rate against CSK

Result 4:

Overall new article sentiment summary for our model and Vader api


Title!
Not secure
newssentimentanalysisontwitter-d329e42a179ca99b.elb.us-east-2.amazonaws.com/checkInput

Summary using our Model

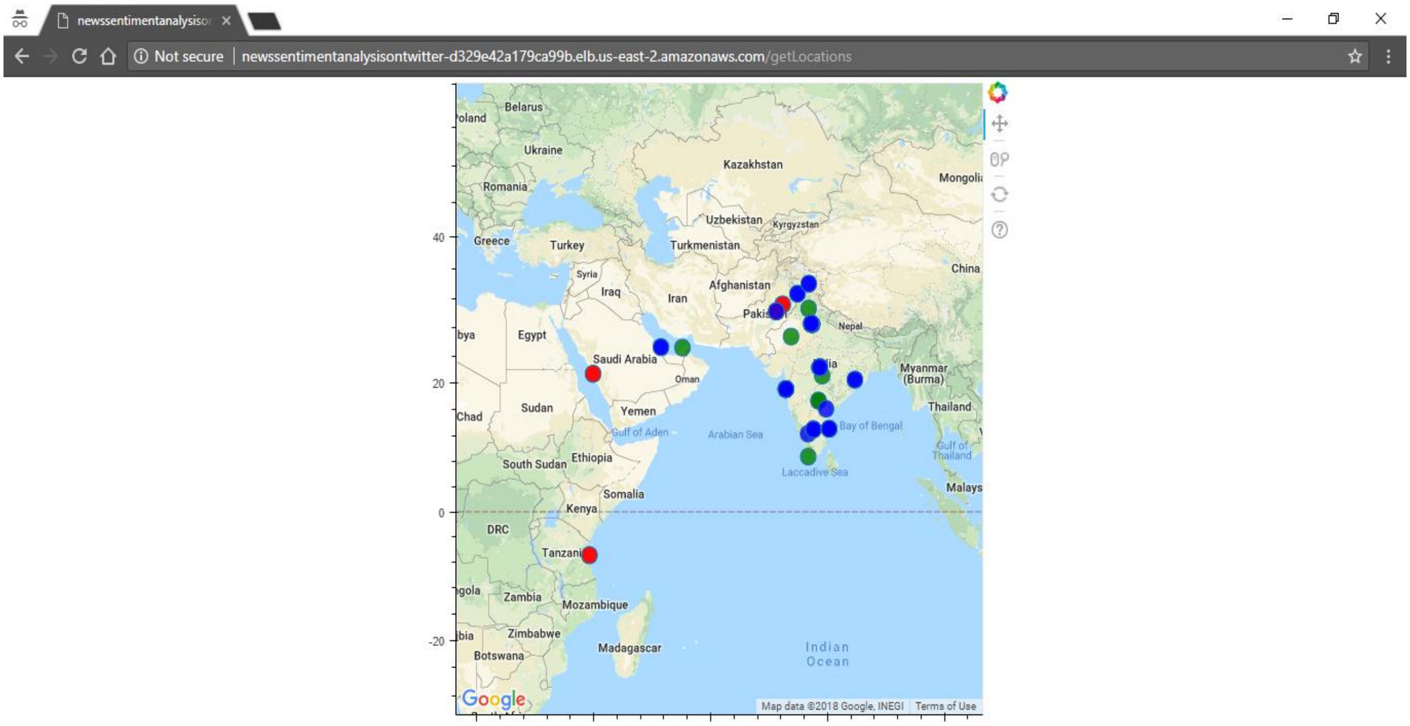
	total_negative	total_positive	total_neutral
name			
Trolls Didn't Spare Anushka Sharma This Time As Well, Blamed Her For RCB's Loss To CSK In IPL	50.000000	0.00000	50.000000
IPL 2018: RCB skipper Virat Kohli fined Rs 12 lakh for slow over rate against CSK	0.000000	0.00000	100.000000
Indian fans blame Anushka Sharma for RCB's defeat against CSK in IPL	0.000000	0.00000	100.000000
Virat Kohli fined for RCB's slow over rate against CSK in IPL 2018	9.523810	50.00000	40.476190
RCB vs CSK Match Highlights: Dhoni finishes off in style as CSK thump RCB	0.000000	90.47619	9.523810
IPL 2018: Sixes record smashed as MS Dhoni's CSK beat Virat Kohli's RCB in Bengaluru	0.000000	0.00000	100.000000
Kohli fined ₹12 lakh for slow over rate against CSK	29.411765	0.00000	70.588235

Summary using vader

	total_negative	total_positive	total_neutral
name			
Trolls Didn't Spare Anushka Sharma This Time As Well, Blamed Her For RCB's Loss To CSK In IPL	100.000000	0.000000	0.000000
IPL 2018: RCB skipper Virat Kohli fined Rs 12 lakh for slow over rate against CSK	20.000000	20.000000	60.000000
Indian fans blame Anushka Sharma for RCB's defeat against CSK in IPL	100.000000	0.000000	0.000000
Virat Kohli fined for RCB's slow over rate against CSK in IPL 2018	30.952381	26.190476	42.857143
RCB vs CSK Match Highlights: Dhoni finishes off in style as CSK thump RCB	4.761905	42.857143	52.380952
IPL 2018: Sixes record smashed as MS Dhoni's CSK beat Virat Kohli's RCB in Bengaluru	100.000000	0.000000	0.000000
Kohli fined ₹12 lakh for slow over rate against CSK	35.294118	2.941176	61.764706

Result 5:

Sentiment analysis plotting on google map.


newssentimentanalysisor
Not secure
newssentimentanalysisontwitter-d329e42a179ca99b.elb.us-east-2.amazonaws.com/getLocations
Belarus
oland
Ukraine
Kazakhstan
Mongolia
Romania
Uzbekistan
Kyrgyzstan
40
Greece
Turkey
Turkmenistan
China
Syria
Iraq
Iran
Afghanistan
Pakistan
Nepal
bya
Egypt
Saudi Arabia
India
Myanmar (Burma)
Oman
20
Chad
Sudan
Yemen
Thailand
Gulf of Aden
Arabian Sea
Bay of Bengal
Gulf of Thailand
South Sudan
Ethiopia
Laccadive Sea
Somalia
Malays
0
Kenya
DRC
Tanzania
gola
Zambia
Mozambique
-20
bia
Zimbabwe
Madagascar
Indian Ocean
Botswana
Google
Map data ©2018 Google, INEGI
Terms of Use