Real Time Analysis of Social Media's Data in Multi-Cluster Environment

Team: **DATA ROOS**

- Venkata Varun Nelakuditi
  (vnh7t@umkc.edu)
  Masters in computer science
  University of Missouri Kansas City

- Vijaybhaskar reddy Agavinti
  (Vactp@umkc.edu)
  Masters in computer science
  University of Missouri Kansas City

- Chandu Chandra
  (vcypr@umkc.edu)
  Masters in computer science
  University of Missouri Kansas City

- Ranga koushik golla
  rg9kt@umkc.edu
  Masters in computer science
  University of Missouri Kansas City

- Prof.Dr. shah, syed jawad H.
  (shs6g7@mail.umkc.edu)
  Big Data and analytics
  University of Missouri Kansas City
  School of Computing and Engineering

ABSTRACT:

These days, people's opinions and product reviews are quite important. These are disseminated via media outlets such as YouTube, personal blogs, and social media sites. Websites like Facebook and Twitter play an important role. Twitter is one such site, which combines microblogging with social networking. Tweets allow us to communicate our thoughts and opinions on Twitter. The major goal of our project is to capture live data and analyze it in real time. Sentiment analysis on live social media's mass data on targeted events, allowing the user to immediately see the impact of his statement on many industries such as the financial market, equities, and tech industry., so that the user can use this report to make last-minute changes to the event in order to change people's minds about it. This application is also utilized to get the most out of day stock trading.

INTRODUCTION:

Project purpose is to understand people's opinions and the impact of those opinions on various domains. This whole process is to be in a real time with low latency and user able to see live analytics/monitor through provided dashboard. The project's purpose is to extract features from a cleaned dataset before applying analytical operations to live tweets in certain categories. As incoming tweets will not contain a categorization, the incoming tweet's category is determined using predictions using machine learning models. Following the categorization of tweets, just the necessary tweets are extracted in order to do sentiment analysis for positive and negative technology analytics. Sentiment analysis can also be predicted using machine learning models that have been developed. This analysis can be useful for a variety of people, including those who are interested in current technological trends, those who want to learn new technologies, and bloggers and reviewers.

Plotly lib for visualization, jupyter IDE for compilation, Dash for generating real-time dashboards, tweepy for collecting live tweets, and pyspark structured stream to perform transformation on real time data, spark MLlib for building models in machine learning, virtual machine, and Ubuntu OSfor virtual simulation on three platforms were the technologies we used.

METHODOLOGY:

Stage 1: preparing machine learning models

Before beginning the project we built machine learning models using spark MLlib to classify live tweets according to their sentiment and categories. For this we used twitter’s 16 million tweets which are pre trained with the sentiment and the other data set is BBC news article dataset where we can get categories of particular topics.

To proceed with sentiment model we built a staging pipeline with the sequence of tokenizing, stop words removal followed by feature vectorization technique , word2 vector and final stage of the pipeline is logical regression after training the above pipeline we got 93% accuracy which met our requirement so we saved this model for later use for category prediction machine learning model we followed same process for the feature extraction for it we used

TF-IDF model and naive bayes as our machine learning algorithm for this model we achieved 98% accuracy so we saved this model into our local, for later use

Stage 2: collecting live tweets

With the help of tweepy api(developers access to be needed) we pulled live tweets based upon search streams we passed in script and we started a local server which helps to push this tweest into port that later accessed by spark in later stage

We used tweepy api 4.8.0 library

Stage 3: running spark job

To perform live analytics we used spark structured streams that are buildup on spark stream library where batch RDDS where processed amd clubbed their results we used stream transactions to preprocess semi structure json data into structured form, machine learning algorithms to run and predict tweet's category and sentiment on live from there we allowed our strem job to run background and write updated data to in memory data frame for every 20 seconds we used that in memory data frame to perform SQL quires and plot data on dashboard which is our stage 4

Structured spark streams has been used

Stage 4: in this stage 4 we visualized analyzed data in dashboard for this we used python dash library and it's call back methods to automatic refresh of data and we used plotly express to visualize interactive graphs

Created a virtual simulation of multi cluster environment where spark stand alone acted as a cluster manager and ubuntu is a chosen os for both master and slave nodes, installed spark environments in all 3 virtual systems, and make them talk through the secure SSH, Our whole setup was running on this system.

Dataset description

Kaggle.com provided the Twitter sentiment dataset and news article datasets (from BBC news). The sentiment of individual tweets is included in the Twitter dataset, which can be used to forecast the sentiment of live streaming tweets. The category of each article is included in the news article dataset, so the categories included in the dataset can be utilized to forecast categories. The News article dataset has over 35 thousand articles, whereas the Twitter dataset has around 1.6 million tweets.

DESIGN:

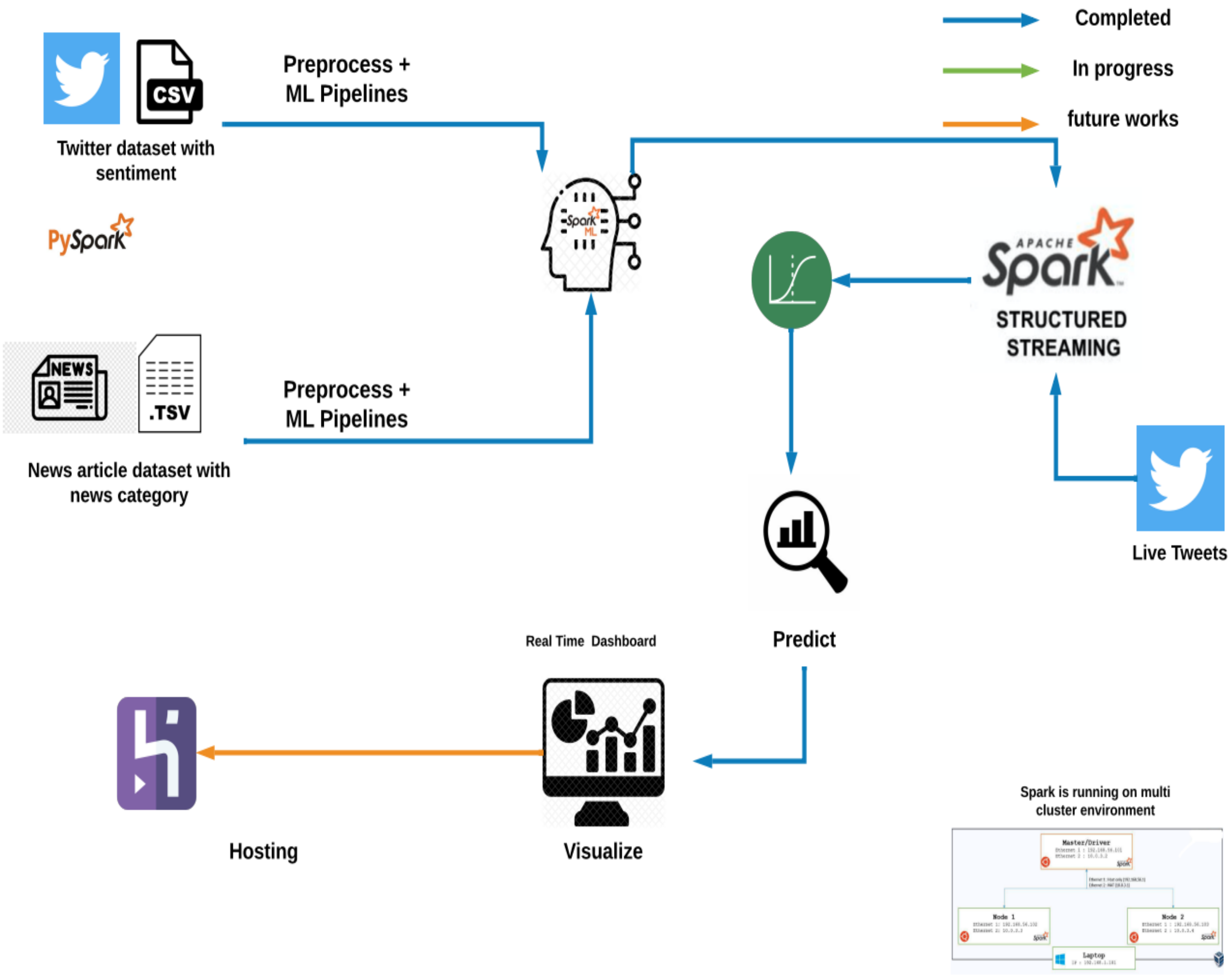

Completed
In progress
future works
Twitter dataset with sentiment
CSV
PySpark
Preprocess + ML Pipelines
Spark ML
APACHE Spark
STRUCTURED STREAMING
NEWS
.TSV
News article dataset with news category
Preprocess + ML Pipelines
Live Tweets
Predict
Real Time Dashboard
Hosting
Visualize
Spark is running on multi cluster environment
Master/Driver
Ethernet 1 : 192.168.56.101
Ethernet 2 : 10.0.3.2
Ethernet 1: Host only (192.168.56.1)
Ethernet 2: NAT (10.0.3.1)
Node 1
Ethernet 1: 192.168.56.102
Ethernet 2: 10.0.3.3
Node 2
Ethernet 1 : 192.168.56.103
Ethernet 2 : 10.0.3.4
Laptop
IP : 192.168.1.101

Fig 1

The flow diagram for the model, Two datasets have been used to train the model, First one, Twitter dataset with sentiment and News article dataset with news category, They have been preprocessed using ML pipeline architecture in spark and spark structured streaming has been used to collect the live data which will be used to visualize the real time dashboard.

Spark is running on multi cluster environment

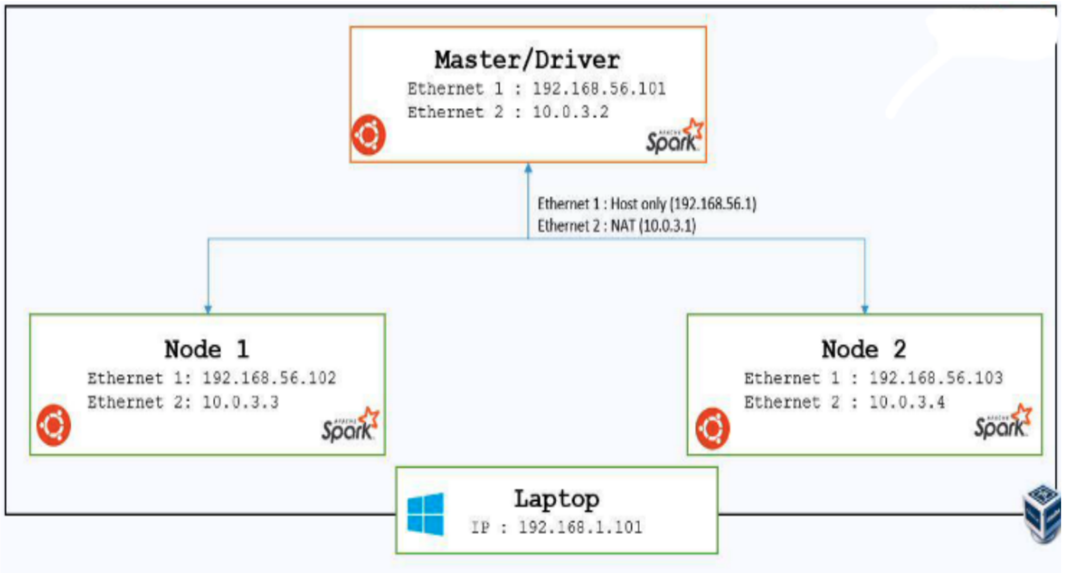
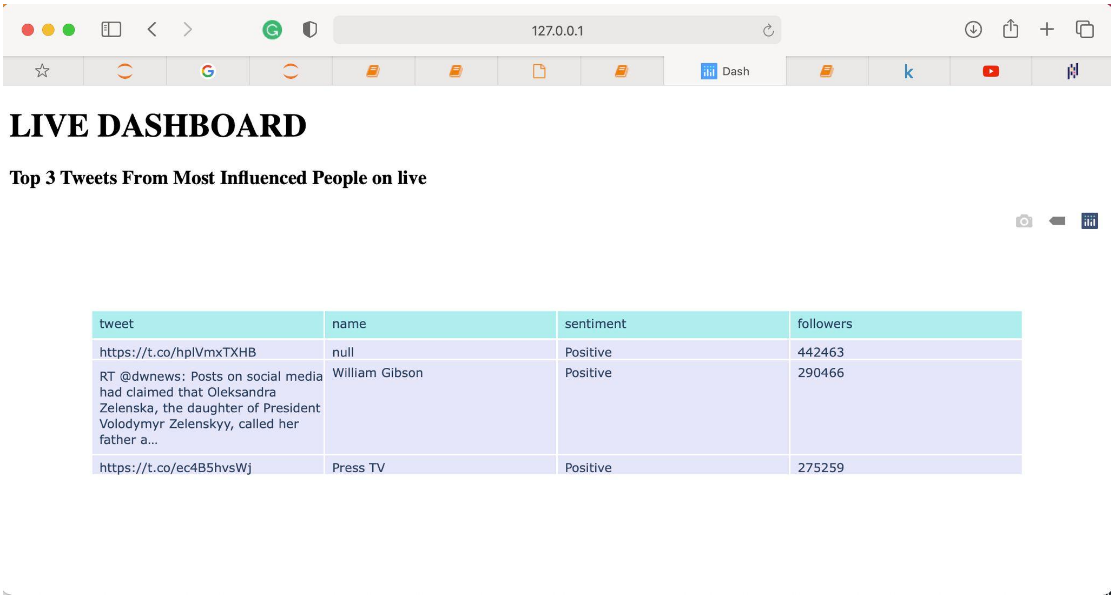

Master/Driver
Ethernet 1 : 192.168.56.101
Ethernet 2 : 10.0.3.2
Spark
Ethernet 1 : Host only (192.168.56.1)
Ethernet 2 : NAT (10.0.3.1)
Node 1
Ethernet 1: 192.168.56.102
Ethernet 2: 10.0.3.3
Spark
Node 2
Ethernet 1 : 192.168.56.103
Ethernet 2 : 10.0.3.4
Spark
Laptop
IP : 192.168.1.101

Fig 1.1
Spark is running on multi clustering environment

RESULTS:

In the Real time data visualization, the following are the results that have been obtained from the model, with the real time twitter tweet datasets.

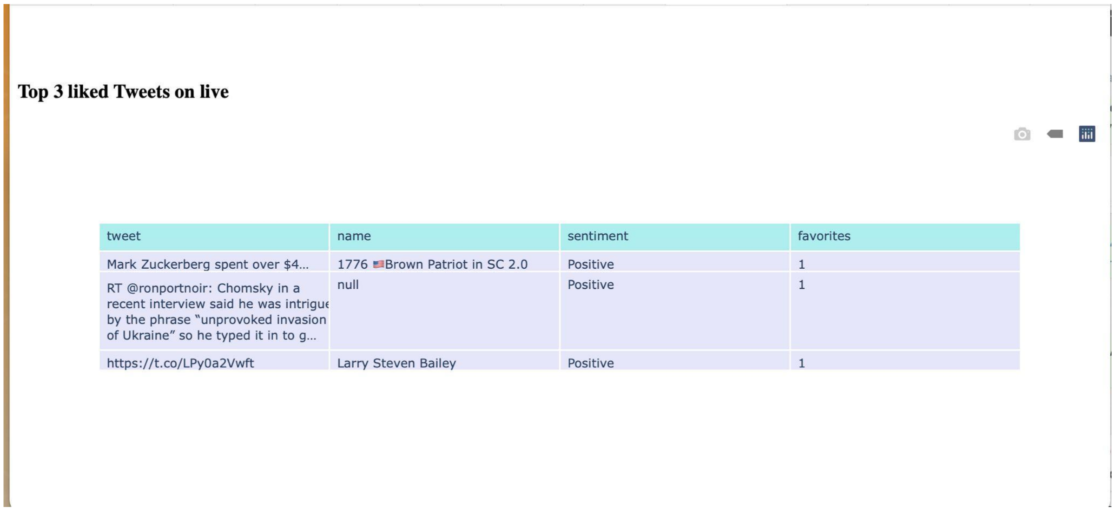


LIVE DASHBOARD

Top 3 Tweets From Most Influenced People on live

tweet	name	sentiment	followers
https://t.co/hplVmxTXHB	null	Positive	442463
RT @dwnews: Posts on social media had claimed that Oleksandra Zelenska, the daughter of President Volodymyr Zelenskyy, called her father a...	William Gibson	Positive	290466
https://t.co/ec4B5hvsWj	Press TV	Positive	275259

Fig 2
Figure 2 represents the Name, Tweet, Sentiment and Number of follower by the top three influencers.

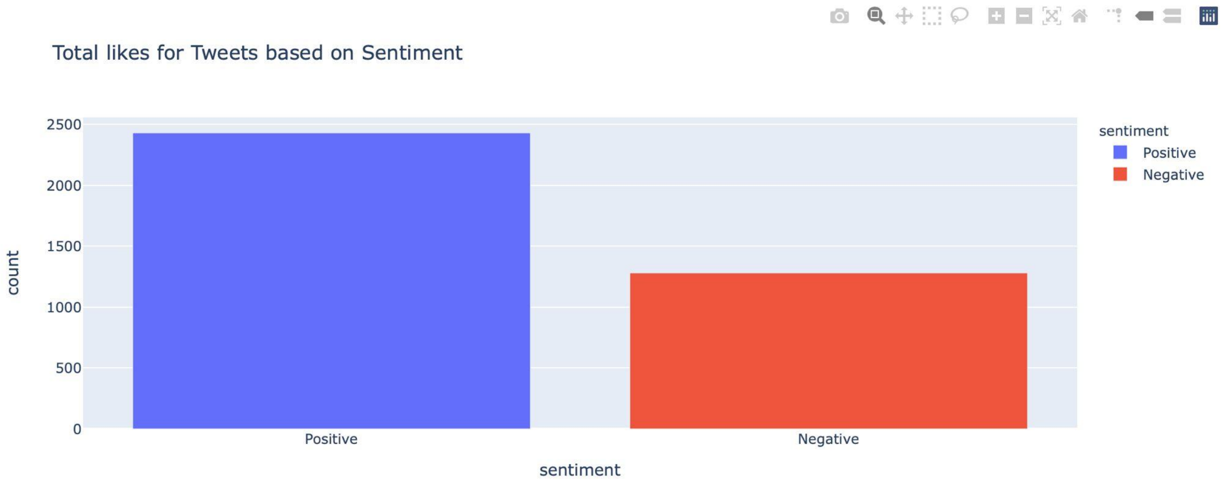


Top 3 liked Tweets on live

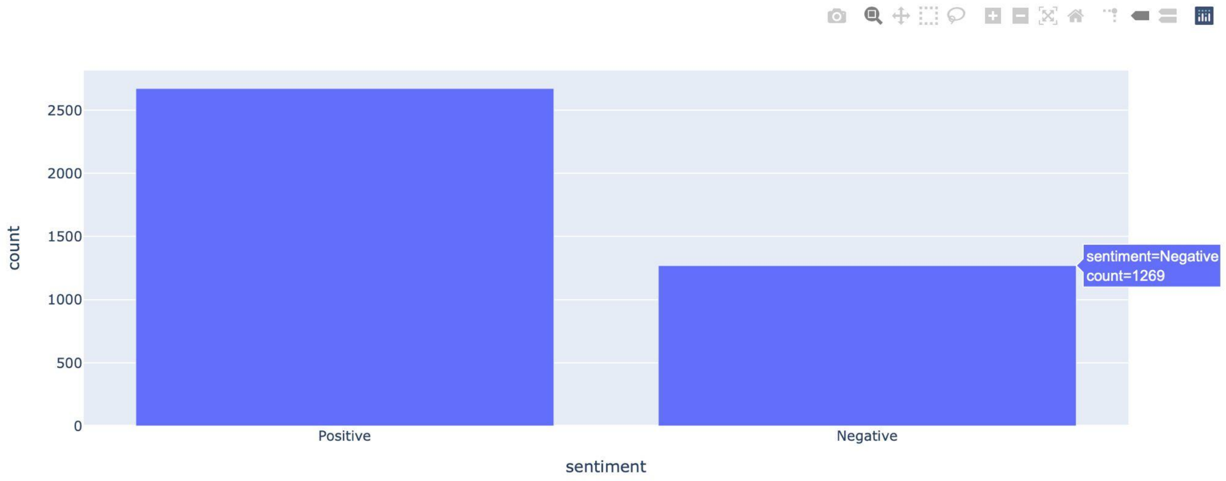
tweet	name	sentiment	favorites
Mark Zuckerberg spent over $4...	1776 🇺🇸Brown Patriot in SC 2.0	Positive	1
RT @ronportnoir: Chomsky in a recent interview said he was intrigue by the phrase "unprovoked invasion of Ukraine" so he typed it in to g...	null	Positive	1
https://t.co/LPy0a2Vwft	Larry Steven Bailey	Positive	1

Figure 2.1 represents the Name, Tweet, Sentiment and top three in number of likes for the tweet

Total likes for Tweets based on Sentiment

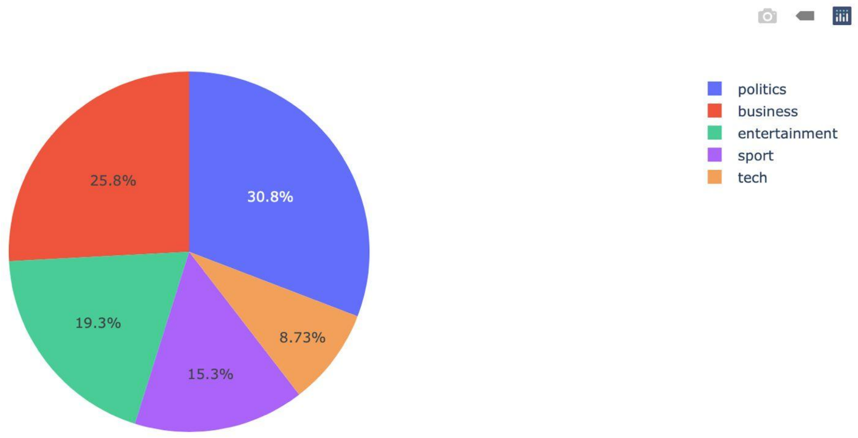


Overall Positive and Negative tweets classification

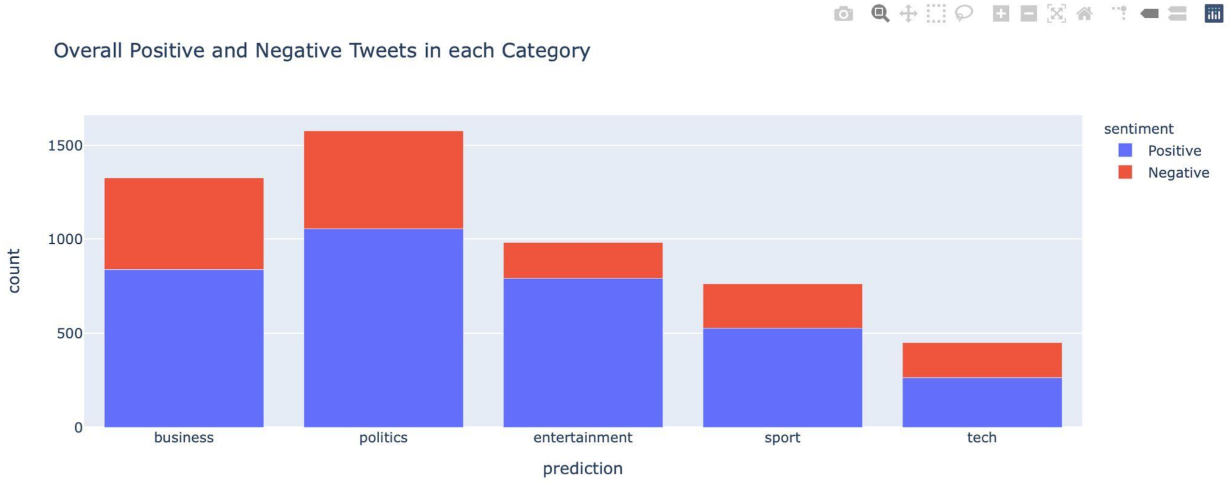


The first image represents the total likes for the tweets in positive sentiment and negative sentiment, From the first figure we can see that the total likes for the positive sentiment is 2400 and total likes for the negative sentiment is1250. As time passes, the number of tweets increases and likes as well.The second image represents the overall positive and negative tweets from the live data collected.

Overall Tweets in each Category



Overall Positive and Negative Tweets in each Category



The first image represents the overall tweets in each category. This model is trained for 5 categories, they are namely politics, business, entertainment, sports and tech. From the collected data 25.8% belongs to business and 30.8% belongs to politics, This data will vary if we change the search straining in the model.

Second image represents the overall positive and negative tweets in each of those 5 categories and they are color coded, blue represents the positive tweets and red represents the negative tweets in that particular category.

APPLICATIONS:

- This model will be useful for day traders, since it is able to analyze the present positive and negative impact on a particular stock, so the user can predict whether the stock price will rise or fall in the coming hours.
- Since it has the ability to analyze with the live social media data, It can also be used to alter the event's outcome based on whether the speech given in the event impacts positively or negatively.

Project extension :

We fearter used nltk library to dig deep into data and analyze sentiment linking with person entitis and opinions for Knowledge discovery management project, here are few results which will indicate users who are associated with the most positive sentiment and negative sentiment persons and entities that are mentioned most frequently in positive and negative tweets and finally we did topic modeling to understand pro and cons in reduced dimensionality space using LDA

REFERENCES:

1. https://www.innovationmerge.com/2021/06/26/Setting-up-a-multi-node-Apache-Spark-Cluster-on-a-Laptop/#toc-heading-8
2. https://www.analyticsvidhya.com/blog/2019/12/streaming-data-pyspark-machine-learning-model/
3. https://medium.com/analytics-vidhya/apache-spark-structured-streaming-with-pyspark-b4a054a7947d
4. https://stackoverflow.com/questions/72046312/how-to-resize-or-set-the-size-of-plotly-express-facet-subplotshttps://stackoverflow.com/questions/72046312/how-to-resize-or-set-the-size-of-plotly-express-facet-subplots