TWITTER SENTIMENT ANALYSIS USING SPARK STREAMING

BY
Manthiramoorthy Cheranthian
Akshaya Sivakumar Karunambika

IDS 561, Analytics for Big Data
University of Illinois at Chicago
May 8,2020

PROBLEM STATEMENT

Businesses are always looking for ways to get immediate feedback from customers to understand the user satisfaction level for their products and services. Initially, customers would be asked to fill out a feedback form on paper or online for a set of questions to help the providers understand how happy they are with their products/services. However, this process of feedback collection may not always prove fruitful as many customers may miss to fill the forms or the standard questions in the form may not fully capture the sentiment of the client.

With the whole world looking to Social media to voice their opinion on anything and everything, Companies have started looking into these platforms to extract insights about customer sentiment about their services. Twitter is one of the most popular social media networks where on an average 6,000 tweets are published per second which corresponds to over 500 million tweets per day.

Thus, we decided to use real-time live twitter data to track and compare the performance to two commercial giants Walmart and Costco. There is a constant debate online about which of the two is better and in which aspects. This analysis can give us insights to the better performer in each product and service. This can be used for understanding the sections where they are performing well and the ones where they need to improve. We have also compared the performance of the two companies in different regions namely New York and San Francisco.

DATA DESCRIPTION

The Data used for this project is from Twitter API. Real time tweets were received through TCP Sockets using Spark Streaming. Tweets are received for processing as and when they are published.

We used three types of filtering for the live Twitter data in our project

- Tweets about Walmart/Costco in the United States of America
- Tweets about Walmart/Costco in San Francisco
- Tweets about Walmart/Costco in New York

TECHNIQUES USED:

There are three parts in our project:

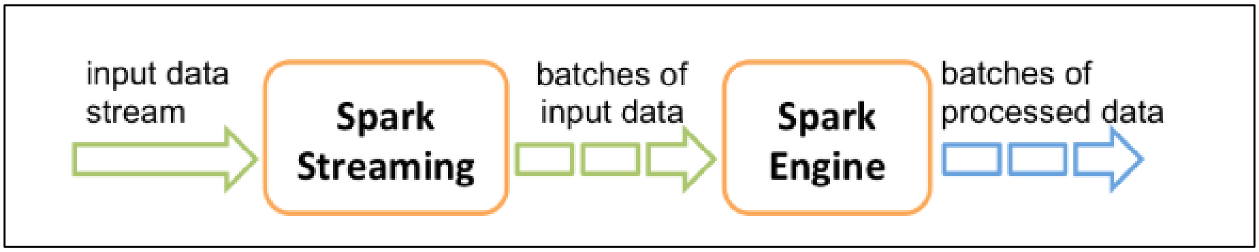
1) Compare the performance of Walmart and Costco based on tweets made all over the United States of America by mapping the sentiment score of every tweet received for each store on a real-time graph.

2) Build a model to classify a tweet as positive, negative, neutral and calculate the aggregate percentage of positive, negative and neutral tweets in each batch received which are in turn loaded into a live graph plot.

3) Region-wise comparison of tweets for Walmart and Costco in two areas, San Francisco and New York and insights for the customer sentiment in each region for the latest news.

Streaming the data:

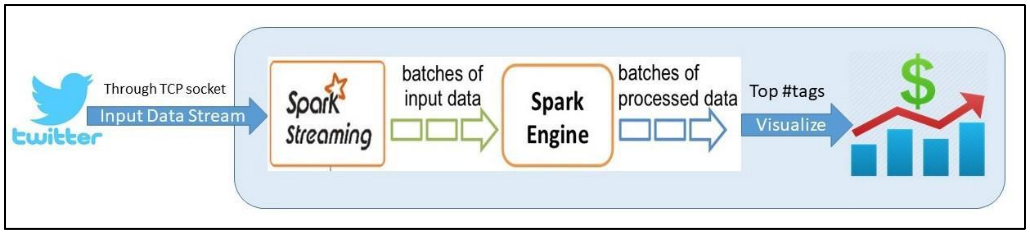
Spark Streaming is an extension of core Spark API, which allows processing of live data streaming from sources like TCP Sockets, Kafka, Flume. It receives real-time input data streams and divides them into batches. These batches are given as input to the Spark Engine to produce the resultant batches of processed data.



In our project, we used *TCP socket method* for streaming the twitter data.

Steps followed to setup streaming data:

1. Create credentials as developer in twitter to access Twitter API
2. App host socket connection is set up to transfer the live tweets
3. Streaming client is set up to receive the tweets every 3 seconds



Part 1: Comparing Walmart and Costco Overall Sentiment

Step I:

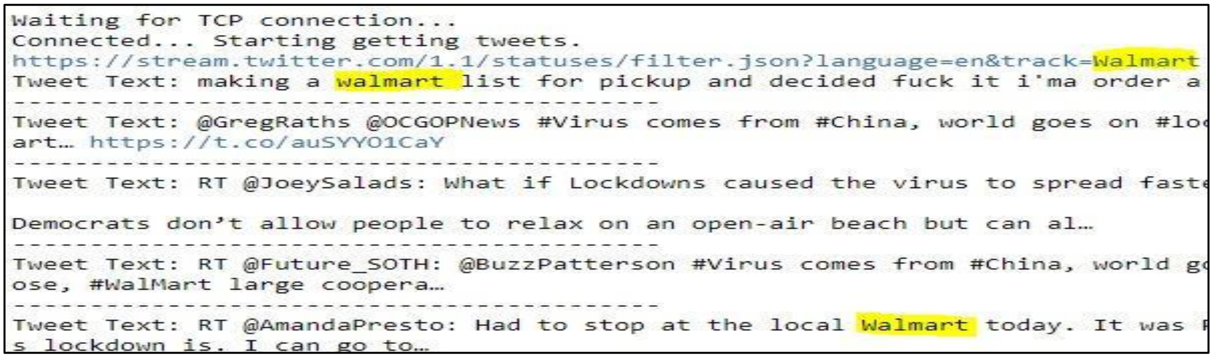
Building the HTTP Server:

- A HTTP Server is built which will connect to Twitter API and get the tweets. Once the tweets are received, they are passed to the Spark Streaming instance.

Filtering tweets referring to Walmart and Costco

Dedicated port numbers were allocated for the two keywords, "**9099**" for Walmart and "**9009**" for Costco. Once our port numbers are set, they actively listen to the inputs from HTTP server outputs (Tweets) for Walmart and Costco separately.

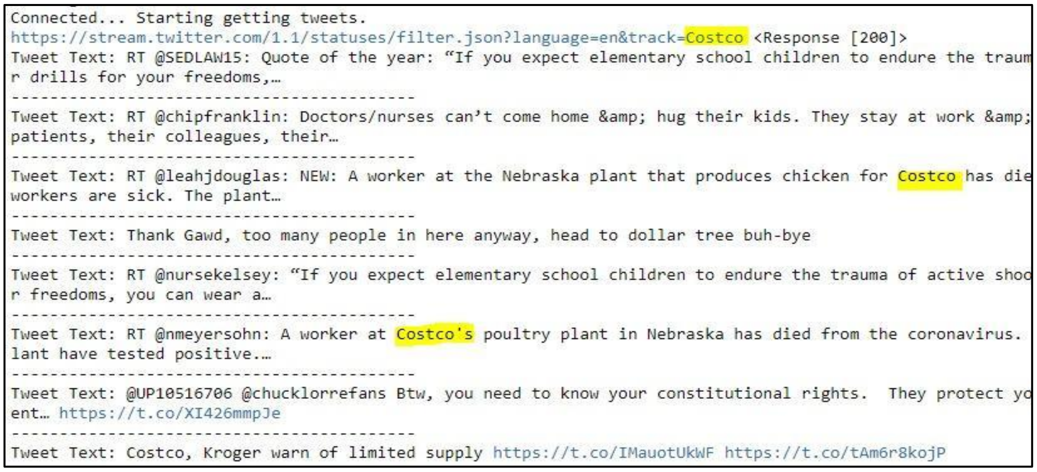
Walmart related tweets



```
Waiting for TCP connection...
Connected... Starting getting tweets.
https://stream.twitter.com/1.1/statuses/filter.json?language=en&track=Walmart
Tweet Text: making a walmart list for pickup and decided fuck it i'ma order a
----------------------------------------
Tweet Text: @GregRaths @OCGOPNews #Virus comes from #China, world goes on #lo
art… https://t.co/auSYYO1CaY
----------------------------------------
Tweet Text: RT @JoeySalads: What if Lockdowns caused the virus to spread faste

Democrats don't allow people to relax on an open-air beach but can al…
----------------------------------------
Tweet Text: RT @Future_SOTH: @BuzzPatterson #Virus comes from #China, world g
ose, #WalMart large coopera…
----------------------------------------
Tweet Text: RT @AmandaPresto: Had to stop at the local Walmart today. It was
s lockdown is. I can go to…
```

Costco related tweets:



```
Connected... Starting getting tweets.
https://stream.twitter.com/1.1/statuses/filter.json?language=en&track=Costco <Response [200]>
Tweet Text: RT @SEDLAW15: Quote of the year: “If you expect elementary school children to endure the traum
r drills for your freedoms,…
------------------------------------------
Tweet Text: RT @chipfranklin: Doctors/nurses can’t come home &amp; hug their kids. They stay at work &amp;
patients, their colleagues, their…
------------------------------------------
Tweet Text: RT @leahjdouglas: NEW: A worker at the Nebraska plant that produces chicken for Costco has die
workers are sick. The plant…
------------------------------------------
Tweet Text: Thank Gawd, too many people in here anyway, head to dollar tree buh-bye
------------------------------------------
Tweet Text: RT @nursekelsey: “If you expect elementary school children to endure the trauma of active shoo
r freedoms, you can wear a…
------------------------------------------
Tweet Text: RT @nmeyersohn: A worker at Costco's poultry plant in Nebraska has died from the coronavirus.
lant have tested positive.…
------------------------------------------
Tweet Text: @UP10516706 @chucklorrefans Btw, you need to know your constitutional rights.  They protect yo
ent… https://t.co/XI426mmpJe
------------------------------------------
Tweet Text: Costco, Kroger warn of limited supply https://t.co/IMauotUkWF https://t.co/tAm6r8kojP
```

Step II:

Building the Spark streaming application.

- Initially we create an instance of Spark Context and then a streaming context with a batch interval of 3 seconds has been created using SparkContext.
- Once the streaming context has been created, we create a DStream that will connect to our Localhost and port.

Step III:

Sentiment Score for each tweet

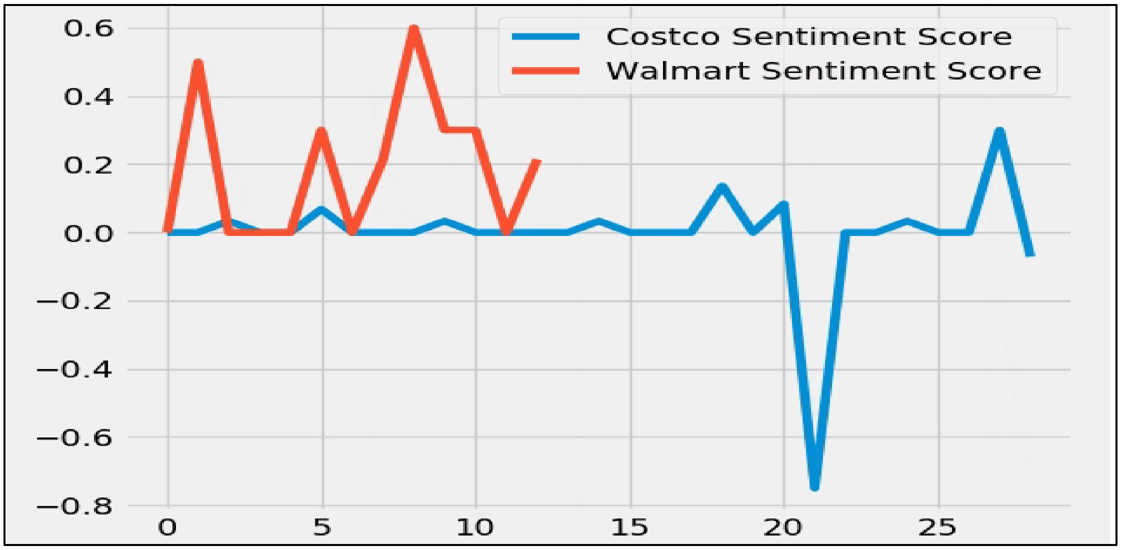
- The Sentiment score for each tweet received is calculated using NLTK sentiment VADER
- TextBlob is a sentiment lexicon which it leverages to give both polarity and subjectivity scores. For Polarity, the values lie between -1 and 1.
- If polarity >0, the sentiment is said to be ‘positive’
  Else If polarity < 0, the sentiment is said to be ‘negative’ else ‘neutral’
- This polarity value is written into two different csv files for each store’s related tweets

Step IV:

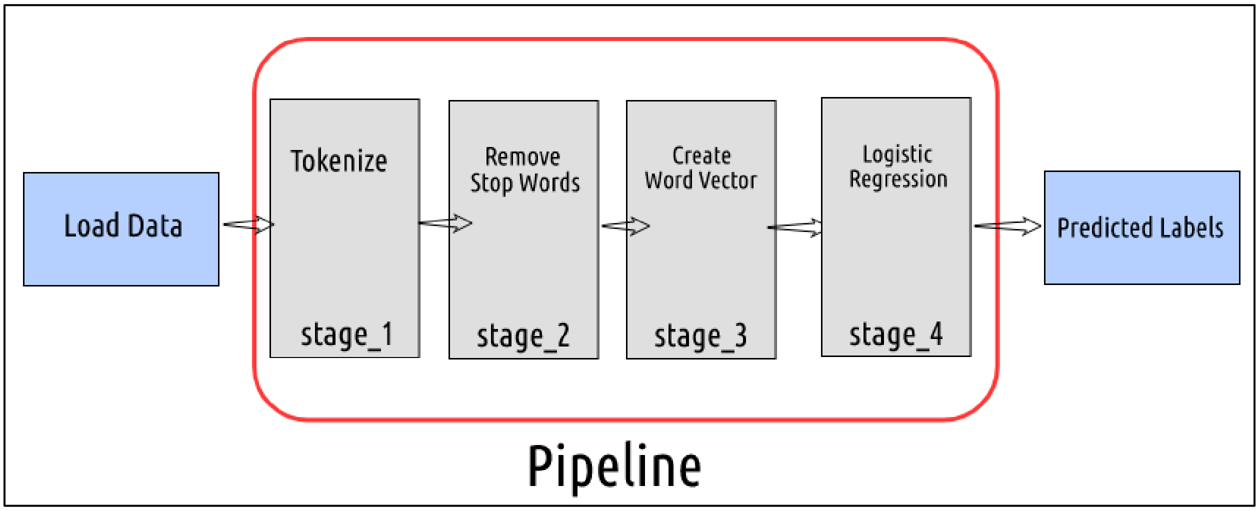
Real time graph

- A live graph shows the real-time sentiment of customers for each batch of tweets received every 3 seconds.
- Two plots are displayed with the same graph, one each for Walmart and Costco.
- Matlibplot library was used for the plot creation.
- FuncAnimation function checks for new data in the respective csv file every 1000ms for making live updates in the graph.

Sentiment Score Plot



Part 2: Model Building



Step I:

Data pre-processing:

Stage1: Tokenization

- First step in data preprocessing is to convert the tweets into a list of words.
- *Regex Tokenizer* is used for this purpose.

Stage 2: Stop words removal

English stop words like a,an,the are filtered from the tweet data using *StopWordsRemover* from pyspark.ml.feature.

Stage 3:

Creating word vector of file size as 100 using *Word2Vec* from pyspark.ml.feature.

Stage 4: Model training and prediction

1) Machine Learning Model:

Building our final Machine learning Model, we use Logistic regression for this part using the vectors as inputs from the preprocessing step. We use **pyspark.ml.classification** library for building our **Logistic regression** model.

Training data:

We obtain the **Sentiment140** training data with **1.6 Million tweets** from Kaggle website from the below link

https://www.kaggle.com/kazanova/sentiment140

Description

Column 1: The polarity of the tweet

0 => negative

2 => neutral

4 => positive

Column 2: ID of the tweet

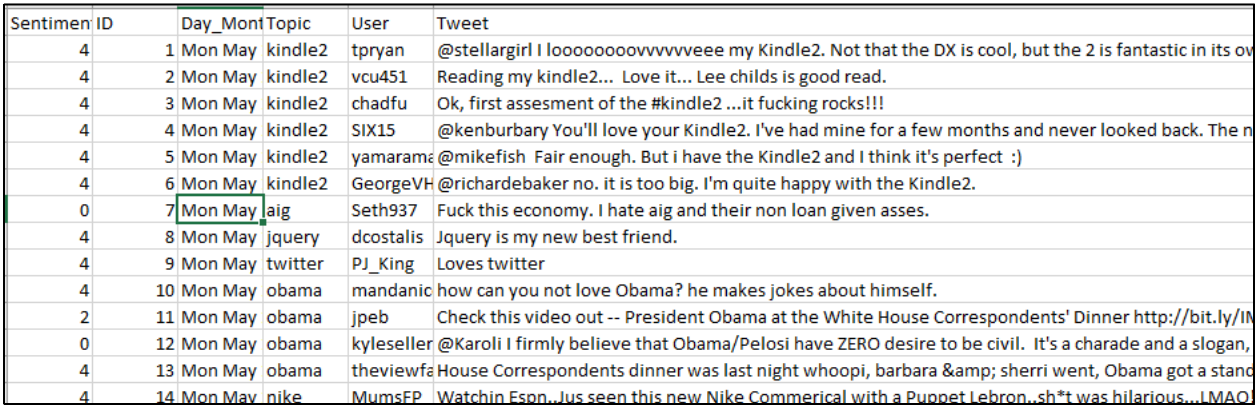
Column 3: Date the tweet was created

Column 4: Query. If there is no query, then this value is NO_QUERY.

Column 5: user that tweeted

Column 6: the text of the tweet

Snapshot of the Training Data:



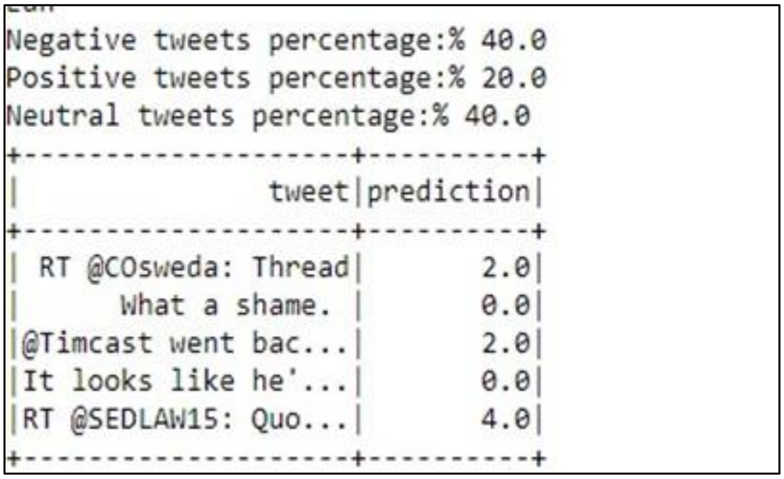
Sentimen	ID	Day_Mont	Topic	User	Tweet
4	1	Mon May	kindle2	tpryan	@stellargirl I looooooooovvvvvveee my Kindle2. Not that the DX is cool, but the 2 is fantastic in its ov
4	2	Mon May	kindle2	vcu451	Reading my kindle2... Love it... Lee childs is good read.
4	3	Mon May	kindle2	chadfu	Ok, first assesment of the #kindle2 ...it fucking rocks!!!
4	4	Mon May	kindle2	SIX15	@kenburbary You'll love your Kindle2. I've had mine for a few months and never looked back. The n
4	5	Mon May	kindle2	yamarama	@mikefish Fair enough. But i have the Kindle2 and I think it's perfect :)
4	6	Mon May	kindle2	GeorgeVH	@richardebaker no. it is too big. I'm quite happy with the Kindle2.
0	7	Mon May	aig	Seth937	Fuck this economy. I hate aig and their non loan given asses.
4	8	Mon May	jquery	dcostalis	Jquery is my new best friend.
4	9	Mon May	twitter	PJ_King	Loves twitter
4	10	Mon May	obama	mandanic	how can you not love Obama? he makes jokes about himself.
2	11	Mon May	obama	jpeb	Check this video out -- President Obama at the White House Correspondents' Dinner http://bit.ly/IN
0	12	Mon May	obama	kyleseller	@Karoli I firmly believe that Obama/Pelosi have ZERO desire to be civil. It's a charade and a slogan,
4	13	Mon May	obama	theviewfa	House Correspondents dinner was last night whoopi, barbara &amp; sherri went, Obama got a stand
4	14	Mon May	nike	MumsFP	Watchin Espn..Jus seen this new Nike Commerical with a Puppet Lebron..sh*t was hilarious...LMAO!

2) Model Prediction:

- Using the trained model mentioned above, we can predict the sentiment values for every incoming streaming tweet at an interval of 3 seconds. Fetched tweets are classified as positive (4.0), neutral (2.0) and negative(0.0) using the trained model.

- Our main objective is to identify the percentage of sentiment distribution among the tweets received in every 3-second interval. This time interval between tweets can be modified depending on the business requirements.

Predicted Sentiment for the Tweets and % share in each fetch:

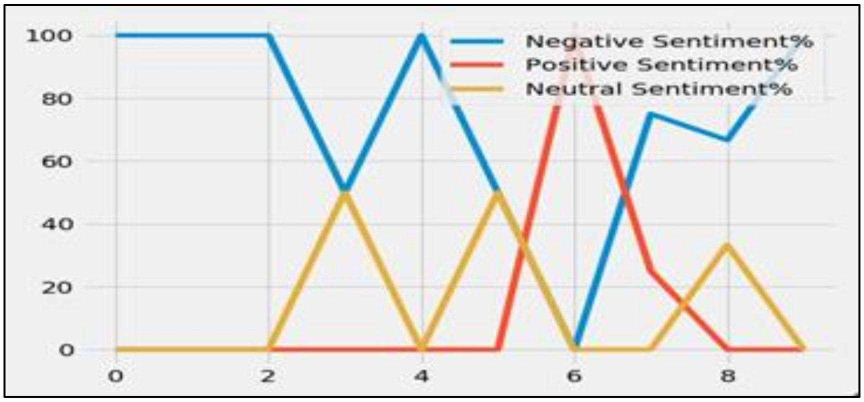


```
Negative tweets percentage:% 40.0
Positive tweets percentage:% 20.0
Neutral tweets percentage:% 40.0
+--------------------+----------+
|               tweet|prediction|
+--------------------+----------+
| RT @COsweda: Thread|       2.0|
|      What a shame. |       0.0|
|@Timcast went bac...|       2.0|
|It looks like he'...|       0.0|
|RT @SEDLAW15: Quo...|       4.0|
+--------------------+----------+
```

3) Data Visualization

- For each set of tweets received, the aggregated Negative, Positive and Neutral sentiment percentage is written into a dataframe.
- The values are then written into a separate csv file for each company, which serve as an input for live graph plot.

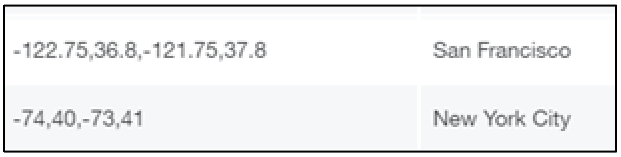
- Once the sentiment scores are written in the file, the FuncAnimation function checks for the updated data in the csv file every 1000ms and plots the graph using Matlibplot library.



Part 3: Region wise Analysis and Comparison

- The tweets were filtered using Location codes in addition to the target keywords like “Walmart” and “Costco” to focus on the tweets specific to the region.
- The Regions compared in this step are New York and San Francisco.

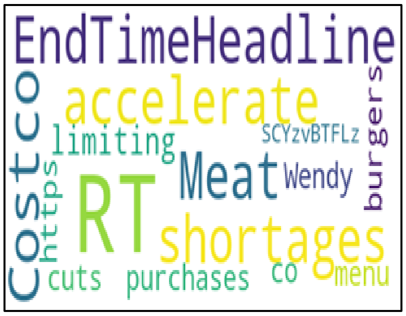
Region codes used to filter the tweets:

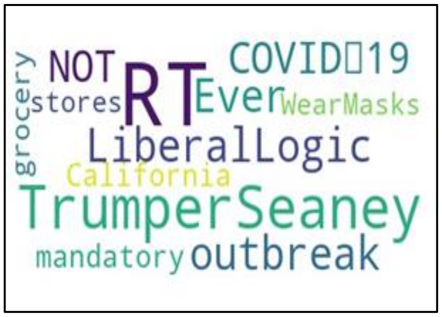


-122.75,36.8,-121.75,37.8	San Francisco
-74,40,-73,41	New York City

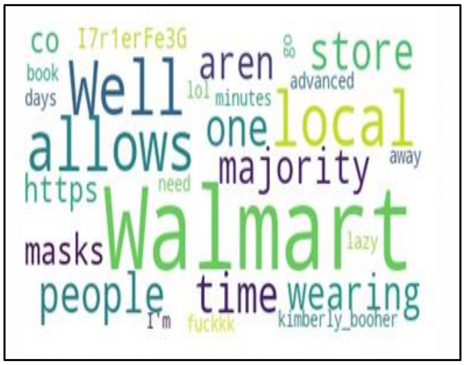
- Model can be used to predict the percentage of positive, negative and neutral tweets in each stream of tweets received for a store in a particular region.
- This allows us to compare the customer sentiment for two different stores in the same region and also the same company in two different regions.
- Once the tweets are fetched from Twitter API, they are plotted using wordcloud library in Python to understand the current trends store wise in each region.

Costco New York **Costco San Francisco**

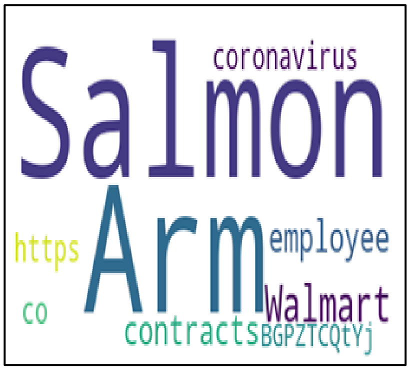




Walmart New York



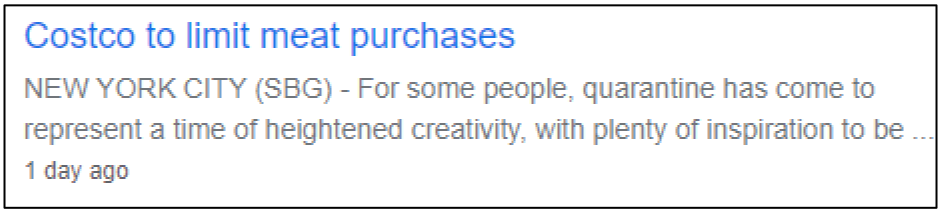
Walmart San Francisco



Insights from Wordcloud:

We were able to correlate the word cloud insights obtained from the tweets with recent news announcements from the retail giants.

- On Monday May 4, The Costco stores in New York announced that Meat purchases were going to be limited to keep with the demand due to the COVID-19 situation.

  Costco to limit meat purchases

  NEW YORK CITY (SBG) - For some people, quarantine has come to represent a time of heightened creativity, with plenty of inspiration to be ...

  1 day ago

  In the word cloud generated the next day (mentioned in the previous page), we found that some of the most common words used in tweets from New York regarding Costco contained the words “Meat”,”Shortages”,”limiting”,”purchases”

- San Francisco Costco stores made it mandatory for shoppers to wear masks in their stores.
  In the word cloud we generated, we found that some of the most used words included “mask”, “store”.

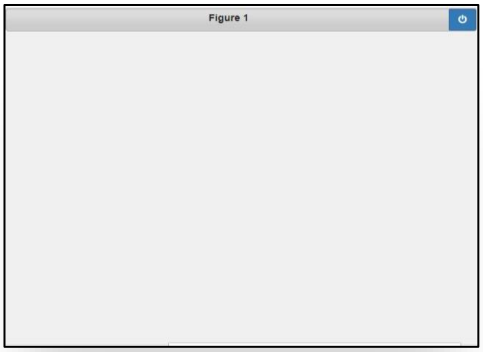
Thus, the word clouds give us an idea about the recent events and announcements associated with a store.

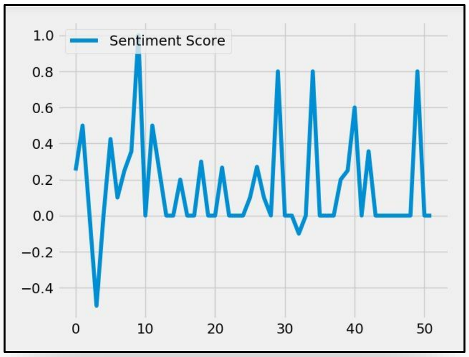
RESULTS:

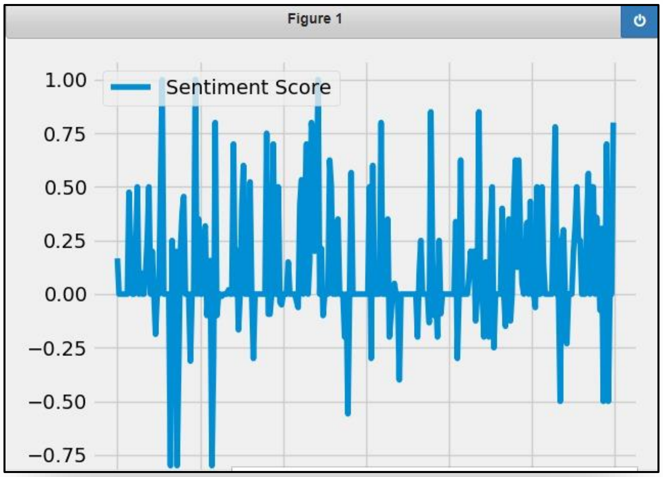
- Live sentiment scores are indicators of customer's reaction for the products and services provided by the enterprises.

- Our project helps in real time tracking of the customer's sentiment based on the tweets received every 3 second interval.

Real time graph of sentiment score for one store







<u>Video showing the real time updation of the compassion plot</u>

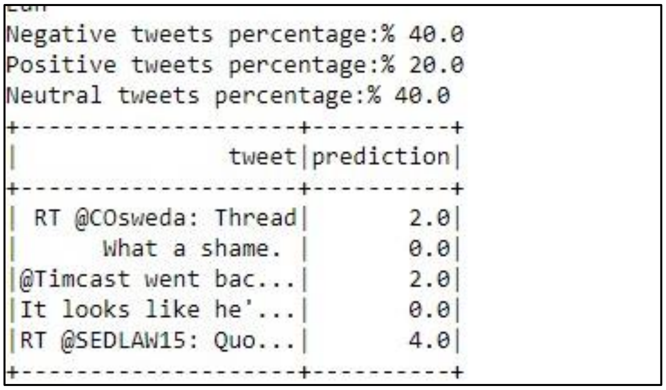
Media1.mp4

- A threshold value for negative sentiments could be maintained if the negative sentiment % and Sentiment score falls below a certain value instant intimation alarms could be set.

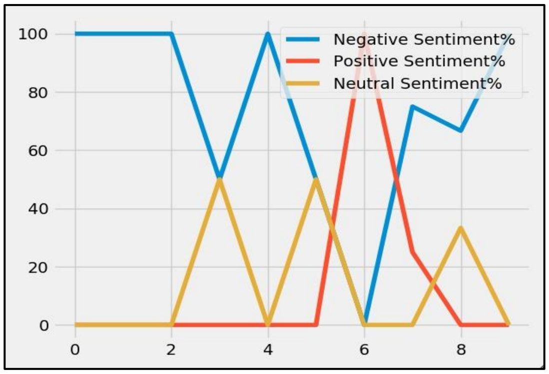
- Tracking this score will help identify the areas that need to be targeted to achieve higher customer satisfaction.

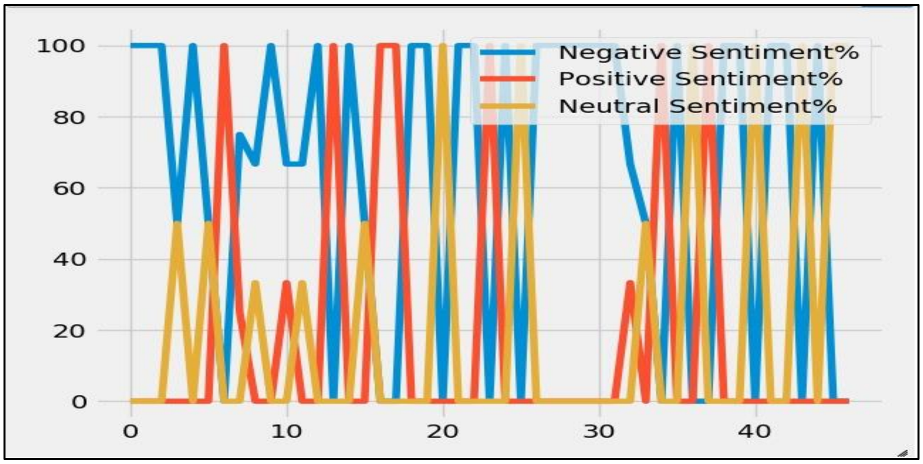
- Aggregated percentage of count of each sentiment type for a store can help track the type of sentiment prevalent among customers at a given point of time.

- This can help to track the customer reactions to a new product/service offered.



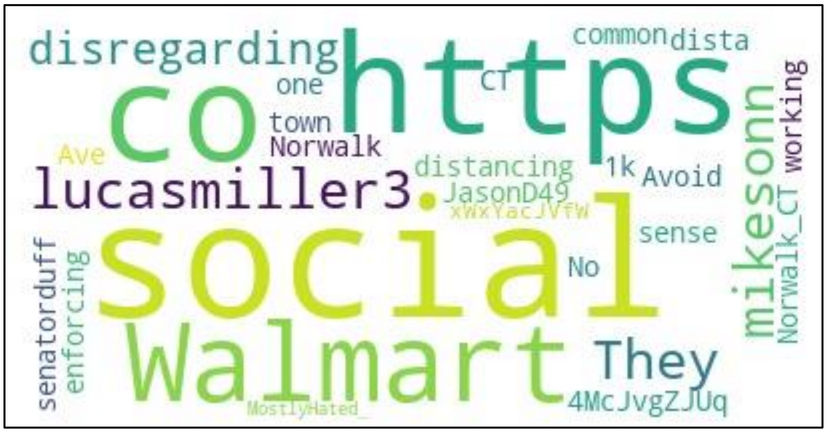
```
Negative tweets percentage:% 40.0
Positive tweets percentage:% 20.0
Neutral tweets percentage:% 40.0
+--------------------+----------+
|               tweet|prediction|
+--------------------+----------+
| RT @COsweda: Thread|       2.0|
|      What a shame. |       0.0|
|@Timcast went bac...|       2.0|
|It looks like he'...|       0.0|
|RT @SEDLAW15: Quo...|       4.0|
+--------------------+----------+
```





- Area wise tracking would help in identifying the regions where the business is performing well and where it needs to be improved.

- Word clouds are one way to understand the current trending topics in a particular store in a specific region.



- Competitors can also be tracked by analyzing the customer sentiment towards their products and services as well.

FUTURE ENCHANCEMENTS:

- Any two brands from any industry can be compared using this method.
- More than two regions can be considered for each store.

ROLE OF TEAM MEMBERS

- Akshaya Sivakumar Karunambika
  Spark Streaming, Sentiment analysis and Documentation

- Manthiramoorthy Cheranthian
  Model building, Word cloud analysis and Documentation

- Rithin Vashishtha
  Data Visualization and Documentation

REFERENCES

- https://spark.apache.org/docs/latest/streaming-programming-guide.html
- https://www.toptal.com/apache/apache-spark-streaming-twitter

- https://www.analyticsvidhya.com/blog/2019/12/streaming-data-pyspark-machine-learning-model/