Comparative Study between Apple IOS and Android by using Big Data Framework

[1]Dipto Halder, [2]Debadrita Panda
[1]Executive Trainee, [2]Assistant Professor
[1]mjunction Services Ltd, [2]Brainware University

ABSTRACT: In today's dynamic world, acquiring an extra edge over competitors user generated components are getting much more attention. So qualitative approach is very effective in this regards. And needless to say, competition can be also on heterogeneous platform. In this tech-friendly world people are too much fascinated towards digital platform and smartphones are biggest contributor to this. This study aims to compare two biggest preferred operating systems used in the smartphones: IOS and Android by using Big Data framework with the help of Sentiment Analysis.

Keywords: Flume, Hive Query Language, Comparative analysis, Twitter, API, Sentiment Analysis.

INTRODUCTION:-

Comparative analysis is a study that compares and contrasts two things: two life insurance policies, two sports figures, two presidents, etc. The study can be done to find the crucial differences between two very similar things or the similarities between two things that appear to be different on the surface.

Today's mobile devices are multi-functional devices capable of hosting a broad range of applications for both business and consumer use. Like a computer operating system, a mobile operating system is the software platform that determines the functions and features available on your device, such as thumbwheel, keyboards, wireless security, and synchronization, messaging and more. Some of the more common and well-known Mobile operating systems are Google's Android and Apple's IOS because of their Popularity we are going to talk about these two operating systems which are commonly seen in market and then compare them with each other.

Here I have tried to obtain world-wide unstructured data from twitter and tried to analyse those on Big-data platform to compare between them (Apple iOS and Google Android). Here I have taken into consideration of all the features and factors which motivate a buyer to choose one between them and also their pre-purchase or post-purchase behaviour which they had expressed over twitter and also how much satisfied are they?

RESEARCH METHODOLOGY:-

Big Data Analytics: It is the process of analyzing large data sets containing a variety of data formats. we can uncover hidden information, unknown correlations, customers preferences, and other useful business information by this process. The primary goal of Big Data analytics is to help organizations to make more informed business decisions by enabling data scientists and other analytics professionals to analyze a large volume of data and different formats of data that cannot process by the conventional system. These types of data could be generated from the web server log, text from customer emails, machine data captured by sensors connected by Internet of Things.

Big Data Analysis: It is the process to know the present behavior, trends and to make future predictions. Various Big Data solutions are used for extracting useful information from Big Data. Big Data solution is a technology that:

Enables the storage of very large amounts of data.
Stores data in low-cost storage devices.
Keeps data in a row or unstructured format.

Big Data analytics performed by a Big Data solution or technologies helps organizations in the following ways:

- Brings improvement in the tourism industry by analyzing the behavior of customers and their trends.
- Enables improvement in the medical field for detecting disease quickly.
- Helps the defense sector by enabling better monitoring.
- Help the insurance industry by better monitoring.
- Helps the insurance industry by better customer relationship managements.

Data Warehouse: Data Warehousing is a group of methods and software to enable data collection from functional systems, integration, and synchronization of that data into a centralized database. Data Warehouse means something that is achieved by integrating data from multiple sources that support logical reporting, structured decision-making, and random queries. The data warehousing process includes cleaning, integration and consolidation of data.

On comparing a data warehouse to a Big Data solution we find that Big Data solution is a technology and data warehouse is an architecture. They are two distinct things. Big Data Technology is just a medium to store and operate huge amount of data, A warehouse is a way of organizing data so that when someone queries or fetches data from it, that person know about other people having the same data for different purposes. There is a scope of data reconcilability in the case of a data warehouse.

Twitter Data Analysis Using Hadoop Framework:

Today's industries and some survey companies are taking decisions by data obtained from the web. As we know the social media is a rich collection of data that is mainly in the form of unstructured data from which we can do analysis on those real-time data which is collected on some situation or on particular things. Twitter is one of the most widely used social networks, and popularity of twitter is increasing day by day as the number of tweets grow exponentially each day. The twitter data is used widely for business analysis. In this paper, we collected tweets from Movie fans during the month of March 2017 using Flume. Twitter is a web application which contains the rich amount of data that can be structured, semi-structured and unstructured data. These data can be collected by using Flume, which is one of the components of the BIG DATA eco-system. It has been performed sentiment analysis on twitter data to know the emotions of Movie Lovers. Since twitter data is in unstructured form and we cannot store and analysis these type of data by using RDMS and SQL query. For analysis purpose, HIVE technology and its queries HQL (Hive Query Language) have been used. In this paper, sentiments are categorized and grouped into 3 groups that come under positive, neutral and negative tweets.

Methodology:
For Twitter data analysis we are going to follow the following Methods:

- Creating twitter Application programming Interface (API)
- Collecting data from twitter using Flume
- Analyzing *twitter* data using Hive
- Creation of learning sheet from row data (unstructured data)
- Applying Dictionary based machine learning algorithm for sentiment analysis
- Visualizing sentiment using Excel sheet, Power View, and Tablue Desktop

Collecting data from twitter using Flume:
Apache Flume is one of the components in Hadoop ecosystem used in transferring large amount of data from distributed resources to a single centralized repository. It is robust and faults tolerant, and efficiently collect data. Flume is specifically designed to push data from various sources to the various storage systems in Hadoop ecosystem, like HDFS and HBase. The simplest unit of flume is a flume agent. Flume agent can be used to move data from one location to another – specifically, from applications producing data to Hadoop Distributed File System (HDFS) HBase, etc. Each Flume agents has three components: source, channel, and sink. The source is responsible for consuming events delivered by an external source like a web server. The sink is responsible for delivering the data to the destination. The channel is a buffer that stores data from the source i.e. Sources ingest events into the channel and the sinks drain the channel.

The configuration of Flume: Flume agents can be configured using plain text configuration files.For real-time streaming, flume-env.sh, flume.conf, and .bashrc file configured according to our requirements. "flume-env.sh" is configured to set environment variables.Here we have to set JAVA_HOME and FLUME_CLASSPATH."flume.conf" is configured to set source ,sink,and channel properties and also to set consumerKey, consumerSecret, Twitter.accessToken,and Twitter.accessTokenSecret.In the Flume configuration file, we need to configured the following Agents:

- ➢ Name the components of the current agent.
- ➢ Configure the source.
- ➢ Configure the sink.
- ➢ Connect the source and the sink to the channel.
- ➢ Configure the channel.
- ➢ we can have multiple agents in Flume. We can differentiate each agent by using a unique name. And using this unique name , we have to configure each flume agents.

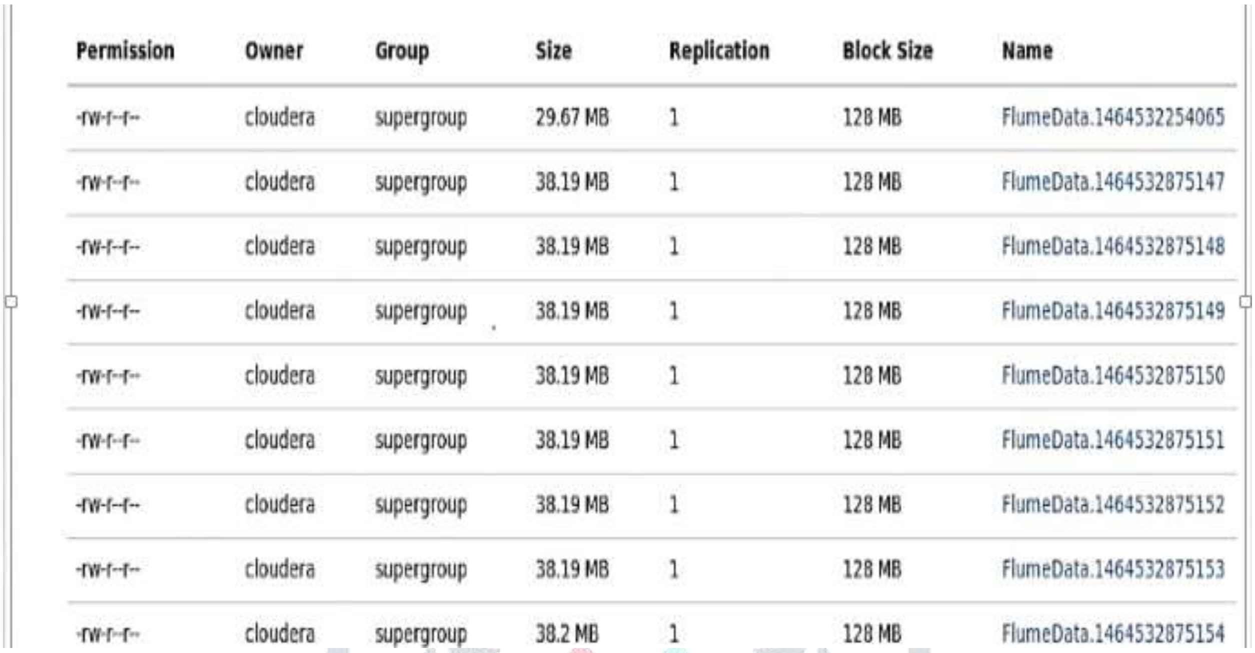
Naming the Flume Agents: First, we need to name the flume agent as shown below:

```
Agent_Name.sources =source_Name
   Agent_Name.sink= Sink_Name
Agent_name.Channels= Channel_name
```

```
 TwitterAgent.sources= Twitter
  TwitterAgent.channels= MemChannel
  TwitterAgent.sinks=HDFS
 TwitterAgent.sources.Twitter.type = com.cloudera.flume.source.TwitterSource
 TwitterAgent.sources.Twitter.channels=MemChannel
 TwitterAgent.sources.Twitter.consumerKey=3649eunRliZPI43s6z8T5LKjA
TwitterAgent.sources.Twitter.consumerSecret=WERNHGD2RTYSAwKKI9h1rnCD33P06IB
 TwitterAgent.sources.Twitter.accessToken=102368189-bjD3VbSMdBOeyoQ93MKJTYUvfS04jmYDoGV3dJelR
TwitterAgent.sources.Twitter.accessTokenSecret=4f3wEp3X1mOZBqavjbVEBLqzcibkU7c0ce1lTUEcr4H0t
TwitterAgent.sources.Twitter.keywords= #iphone
TwitterAgent.sinks.HDFS.channel=MemChannel
TwitterAgent.sinks.HDFS.type=hdfs
TwitterAgent.sinks.HDFS.hdfs.path=hdfs://localhost:8020/flume/tweets
TwitterAgent.sinks.HDFS.hdfs.fileType=DataStream
TwitterAgent.channels.MemChannel.transactionCapacity=100
TwitterAgent.sinks.HDFS.hdfs.writeformat=Text
TwitterAgent.sinks.HDFS.hdfs.batchSize=1000
TwitterAgent.sinks.HDFS.hdfs.rollCount=10000
TwitterAgent.sinks.HDFS.hdfs.rollInterval=600
TwitterAgent.channels.MemChannel.type=memory
TwitterAgent.channels.MemChannel.capacity=10000
```

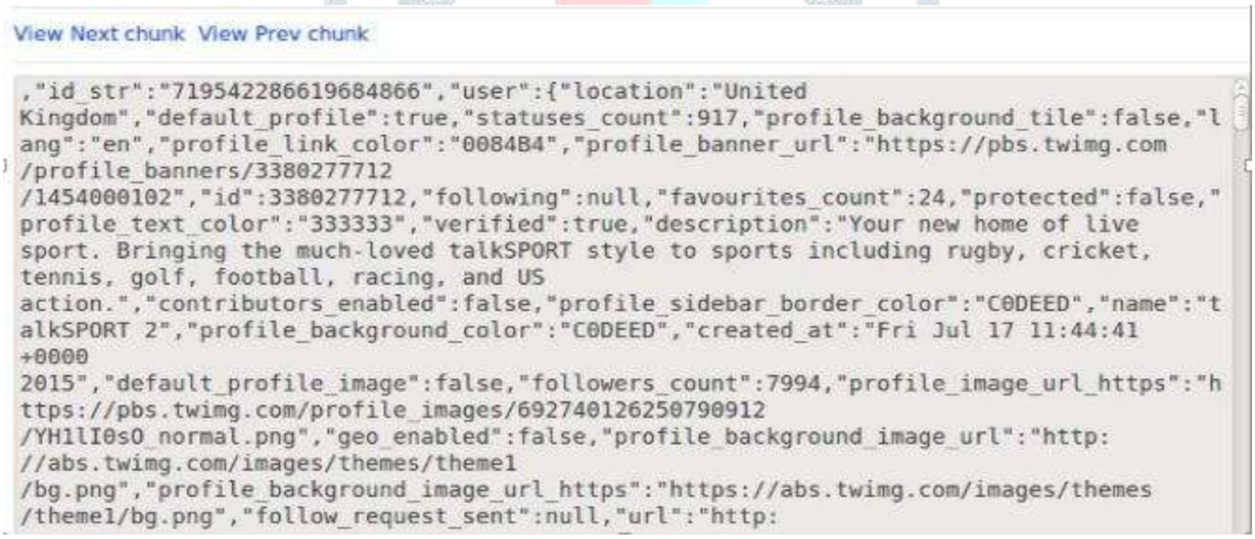
Analyzing Twitter data using Hive:

After running the flume by setting configuration files, twitter data automatically saved into specified location in Hadoop Distributed File System (HDFS). The data that we got from twitter is in JSON format. The following figure shows how data is Stored in the HDFS in a documented format.



Permission	Owner	Group	Size	Replication	Block Size	Name
-rw-r--r--	cloudera	supergroup	29.67 MB	1	128 MB	FlumeData.1464532254065
-rw-r--r--	cloudera	supergroup	38.19 MB	1	128 MB	FlumeData.1464532875147
-rw-r--r--	cloudera	supergroup	38.19 MB	1	128 MB	FlumeData.1464532875148
-rw-r--r--	cloudera	supergroup	38.19 MB	1	128 MB	FlumeData.1464532875149
-rw-r--r--	cloudera	supergroup	38.19 MB	1	128 MB	FlumeData.1464532875150
-rw-r--r--	cloudera	supergroup	38.19 MB	1	128 MB	FlumeData.1464532875151
-rw-r--r--	cloudera	supergroup	38.19 MB	1	128 MB	FlumeData.1464532875152
-rw-r--r--	cloudera	supergroup	38.19 MB	1	128 MB	FlumeData.1464532875153
-rw-r--r--	cloudera	supergroup	38.2 MB	1	128 MB	FlumeData.1464532875154

Fig.1.: Twitter data in Hadoop Distributed File System (HDFS)



```
,"id_str":"719542286619684866","user":{"location":"United
Kingdom","default_profile":true,"statuses_count":917,"profile_background_tile":false,"l
ang":"en","profile_link_color":"0084B4","profile_banner_url":"https://pbs.twimg.com
/profile_banners/3380277712
/1454000102","id":3380277712,"following":null,"favourites_count":24,"protected":false,"
profile_text_color":"333333","verified":true,"description":"Your new home of live
sport. Bringing the much-loved talkSPORT style to sports including rugby, cricket,
tennis, golf, football, racing, and US
action.","contributors_enabled":false,"profile_sidebar_border_color":"C0DEED","name":"t
alkSPORT 2","profile_background_color":"C0DEED","created_at":"Fri Jul 17 11:44:41
+0000
2015","default_profile_image":false,"followers_count":7994,"profile_image_url_https":"h
ttps://pbs.twimg.com/profile_images/692740126250790912
/YH1lI0sO_normal.png","geo_enabled":false,"profile_background_image_url":"http:
//abs.twimg.com/images/themes/theme1
/bg.png","profile_background_image_url_https":"https://abs.twimg.com/images/themes
/theme1/bg.png","follow_request_sent":null,"url":"http:
```

Fig.2: Twitter data format

From these data first, we have created a table in HDFS location to provide schema over twitter data. The Twitter data contains a different type of information about feeds like the text of the tweet, the sender of tweets, timestamp, etc. The scheme is set to provide structure over twitter data which is stored in Hadoop Distributed File System. We can say that we have converted the unstructured data into a structured format. For we use custom serDe concepts. A serDe is a combination of a Serializer and a Deserializer. The Deserializer interface is an interface which takes a string or binary representation of a record and converts it into a Java object that Hive can understand. The Serializer, however, will take a Java object that Hive has been working with, and turn it into something that Hive can write to HDFS or another supported system [5]. The concepts of serDe are that it help to read the data that is in the form of JSON format for that we are using the custom serDe for JSON so that our hive can read the JSON data. And can create a table in our prescribed format.

Hive Query Language (HQL) for creating table on the top of Twitter data is given below:

```
CREATE EXTERNAL TABLE tweets (
  id BIGINT,
  created at STRING,
  source STRING,
  favorited BOOLEAN,
  retweet_count INT,
  retweeted status STRUCT<
    text: STRING,
    user:STRUCT<screen_name:STRING, name:STRING>>,
  entities STRUCT<
    urls:ARRAY<STRUCT<expanded_url:STRING>>,
    user_mentions:ARRAY<STRUCT<screen_name:STRING,name:STRING>>,
    hashtags:ARRAY<STRUCT<text:STRING>>>,
  text STRING,
  user STRUCT<
    screen_name:STRING,
    name:STRING,
    friends_count:INT,
    followers_count:INT,
    statuses_count:INT,
    verified:BOOLEAN,
    utc_offset:INT,
    time_zone: STRING>,
  In_reply_to_screen_name STRING
) ROW FORMAT SERDE 'com.cloudera.hive.serde.JSONSerDe'
LOCATION '/user/dipto /tweets_data';
```

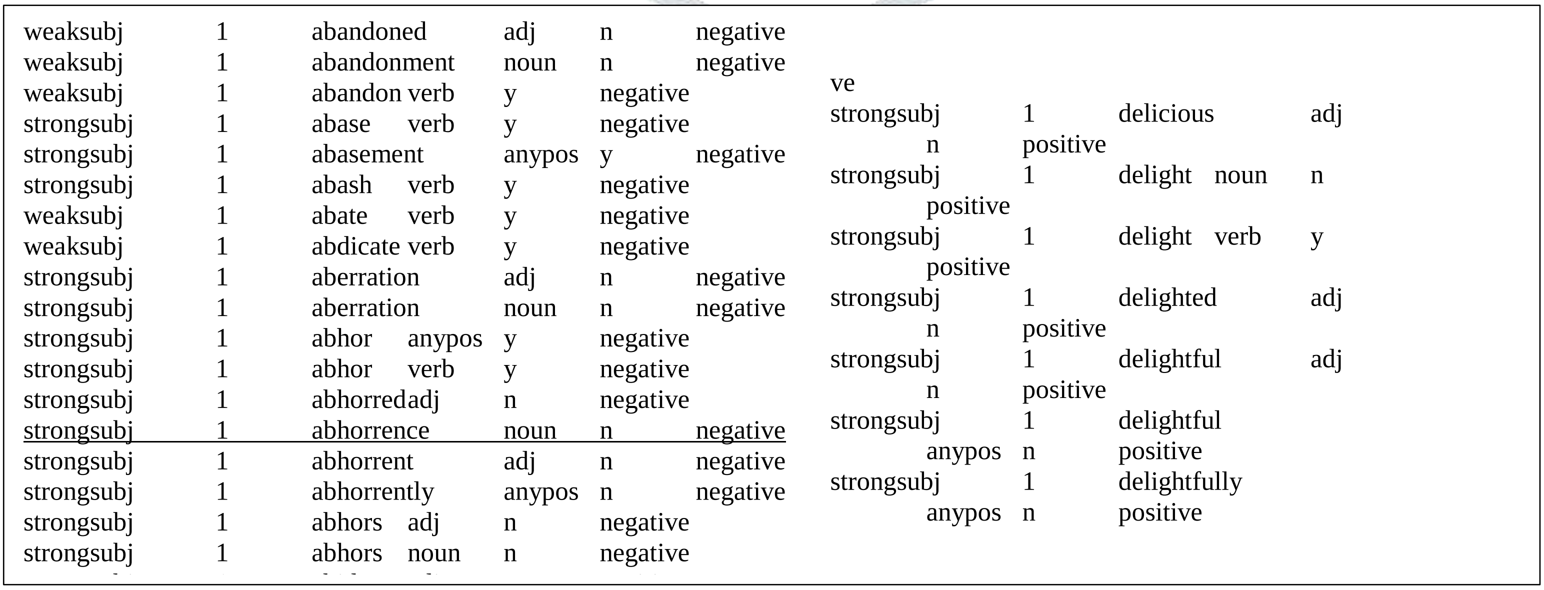
Creation of learning sheet from row data (unstructured data)

For Dictionary based machine learning technique, I have created a data sets that contain the movie reviewers' positive, negative reactions. In this analysis, I have categorized tweets sentiments into three categories positive negative and neutral .And every sentiment is assigned with some weights.

Sample Emotion and text in Tweets.

Sentiments types	Dictionary Sample	Tweets Sample
Positive	incredible	#iphone7 is great Very user-friendly #andriod
Negative	bored	Very costly #iphone
Neutral	common	

The following tables show the sample of learning sheet which has been created to perform machine learning algorithm.



weaksubj	1	abandoned	adj	n	negative
weaksubj	1	abandonment	noun	n	negative
weaksubj	1	abandon	verb	y	negative
strongsubj	1	abase	verb	y	negative
strongsubj	1	abasement	anypos	y	negative
strongsubj	1	abash	verb	y	negative
weaksubj	1	abate	verb	y	negative
weaksubj	1	abdicate	verb	y	negative
strongsubj	1	aberration	adj	n	negative
strongsubj	1	aberration	noun	n	negative
strongsubj	1	abhor	anypos	y	negative
strongsubj	1	abhor	verb	y	negative
strongsubj	1	abhorred	adj	n	negative
strongsubj	1	abhorrence	noun	n	negative
strongsubj	1	abhorrent	adj	n	negative
strongsubj	1	abhorrently	anypos	n	negative
strongsubj	1	abhors	adj	n	negative
strongsubj	1	abhors	noun	n	negative
ve					
strongsubj	1	delicious	adj	n	positive
strongsubj	1	delight	noun	n	positive
strongsubj	1	delight	verb	y	positive
strongsubj	1	delighted	adj	n	positive
strongsubj	1	delightful	adj	n	positive
strongsubj	1	delightful	anypos	n	positive
strongsubj	1	delightfully	anypos	n	positive

Twitter data cleaning process using hive and its queries Hive query Language(HQL)

In the initial phase, the movie fans' tweets are retrieved from twitter using Flume and stored into Hadoop Distributed File System(HDFS) in specified location. After retrieving tweets,we have put schema over Flume data by using hive query.I have added serDe jar so that hive can understand the format of data . In the second phase, the tweets are preprocessed using dictionary based technique. Using dictionary based technique, each word is compared with learning sheet listed as positive , negative and neutral words. From these words, I have given a score to each of the words as the positive "+1" , negative "-1" and neutral "0". Data cleaning process pictorially depicted in Fig--------

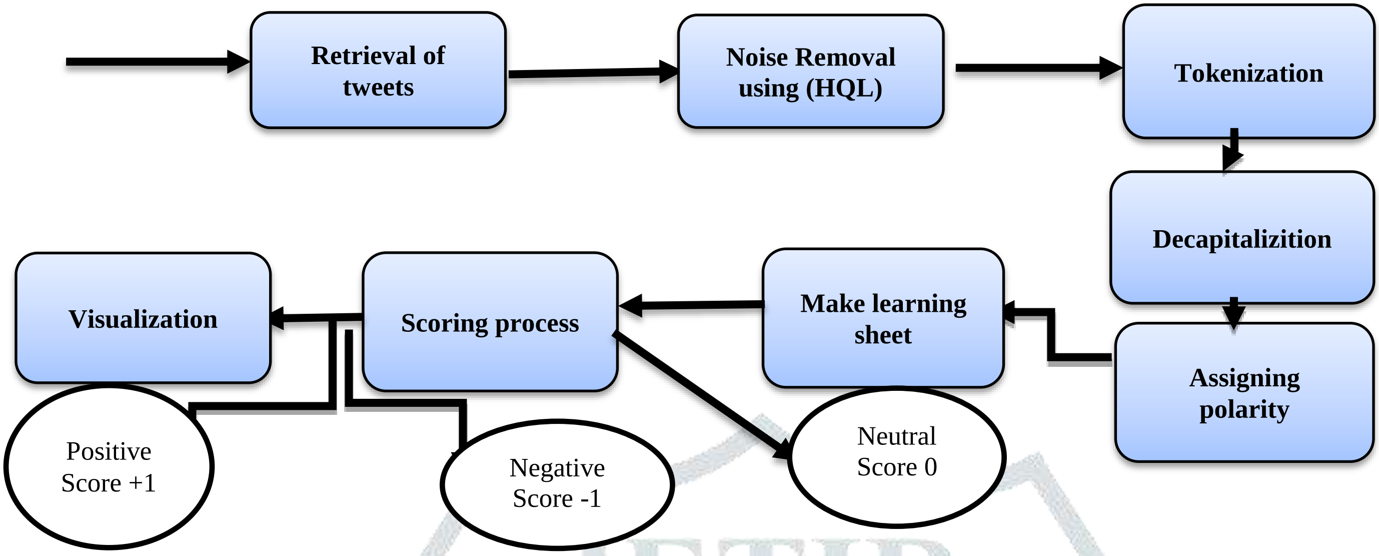


Fig.3: Flow Chart of this Study

I. **Retrieval of twitter data:** Using Twitter Application Programming Interface (API) I have retrieved the tweets of the movie "logan". I have used Flume to retrieve the tweets.

II. **Creation of Learning sheet**:The datasets from the row tweets that contain movie fans' positive, negative reaction.

III. **Schematization of learning sheet:** Then I have schematized the learning sheet using Hive Query Language to give schema according to my requirements.There are two attributes in my learning sheet word and its corresponding polarity.

IV. **Pre-processing of tweets:** Pre-processing starts the text preparation into a more structured representation.Pre-processing includes the following steps:

- ➢ **Data Filtering:** Filter out some information from raw tweets which are required for my analysis such as user Id,Timestamp,Text,and user's time-zone using HQL.
- ➢ **Decapitalization:** After Data Filtering,Tweet text converted into small letters.Here Decapitalization is done by Hive Query language (HQL)
- ➢ **Tokenization:** Tokenization is used to identify all words in a given text.

V. **Scoring:** After processing, I got only accurate and meaningful tweets. Now tweets are compared with learning sheet to assign a polarity to each word.

Codings:

1. -- Clean up tweets

```
CREATE VIEW tweets_simple AS
SELECT
  id,
  cast ( from_unixtime( unix_timestamp(concat( '2014 ', substring(created_at,5,15)), 'yyyy MMM dd hh:mm:ss')) as timestamp) ts,
  text,
  user.time_zone
FROM tweets_raw
;
```

2. -- Get the tweets based on matching time zones

```
CREATE VIEW tweets_clean AS
SELECT
  id,
  ts,
  text,
  m.country
 FROM tweets_simple t LEFT OUTER JOIN time_zone_map m ON t.time_zone = m.time_zone;
```

3. -- Compute sentiment

```
a. create view l1 as select id, words from tweets_raw lateral view explode(sentences(lower(text))) dummy as words;
```

```
b. create view l2 as select id, word from l1 lateral view explode( words ) dummy as word ;

 c.create view l3 as select
     id,
     l2.word,
     case d.polarity
       when  'negative' then -1
       when 'positive' then 1
       else 0 end as polarity
  from l2 left outer join dictionary d on l2.word = d.word;
```

```
6.-- Create the sentiments
create table tweets_sentiment as select
  id,
  case
    when sum( polarity ) > 0 then 'positive'
    when sum( polarity ) < 0 then 'negative'
    else 'neutral' end as sentiment
 from l3 group by id;
```

```
7. -- put everything back together and re-number sentiment
CREATE TABLE tweetsbi
STORED AS ORC
AS
SELECT
  t.id,t.country,
  case s.sentiment
    when 'positive' then 1
    when 'neutral' then 0
    when 'negative' then -1
  end as sentiment
FROM tweets_clean t LEFT OUTER JOIN tweets_sentiment s on t.id = s.id;
```

Coding for analysis:

1. Describe tweets;
2. Select id from tweets limit=100;
3. Select count (*) from tweets;
4. Select created_at from tweets;
5. Select user.time_zone from tweets;
6. Select text from tweets;
7. Select retweet_count from tweets;
8. Select retweeted_status.text from tweets;

Commandline codes in ubuntu:

1. bin/flume-ng agent --conf ./conf/ -f conf/flume.conf - flume.root.logger=DEBUG,console -n TwitterAgent

2.make directories

- hadoop fs -mkdir /iphone_db /tweets_data
- hadoop fs -mkdir /iphone_db /timezonemap
- hadoop fs -mkdir /iphone_db /learningsheet
-

3.Load data to respective directories

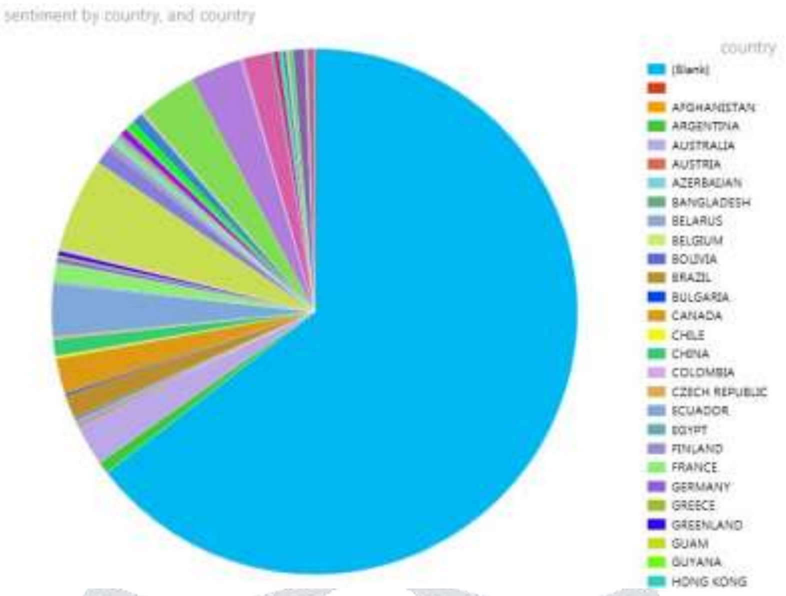
- hadoop fs -copyFromLocal /home/cloudera/ iphone_db / /apple_db /tweets_data
- hadoop fs -copyFromLocal learningtable.txt / iphone_db /learningsheet
- hadoop fs -copyFromLocal time_zone_map.tsv / iphone_db /timezonemap

Other requirements:

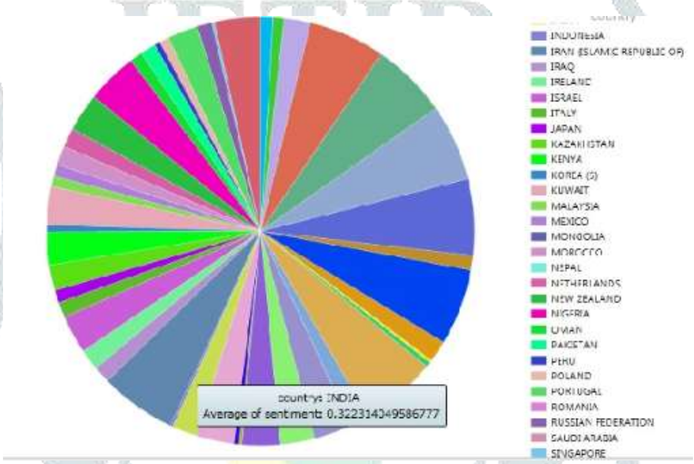
1. silverlight
2. odbc driver

Visualization:

1.

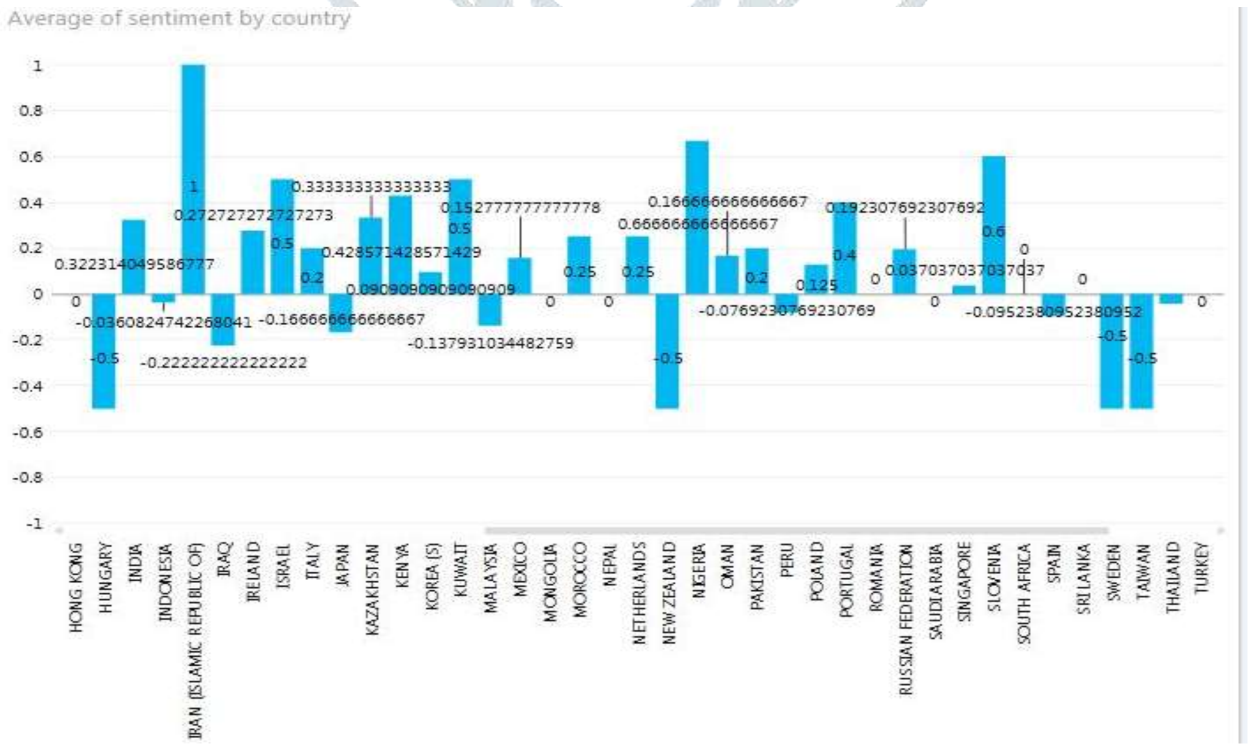


Country-wise Apple ios

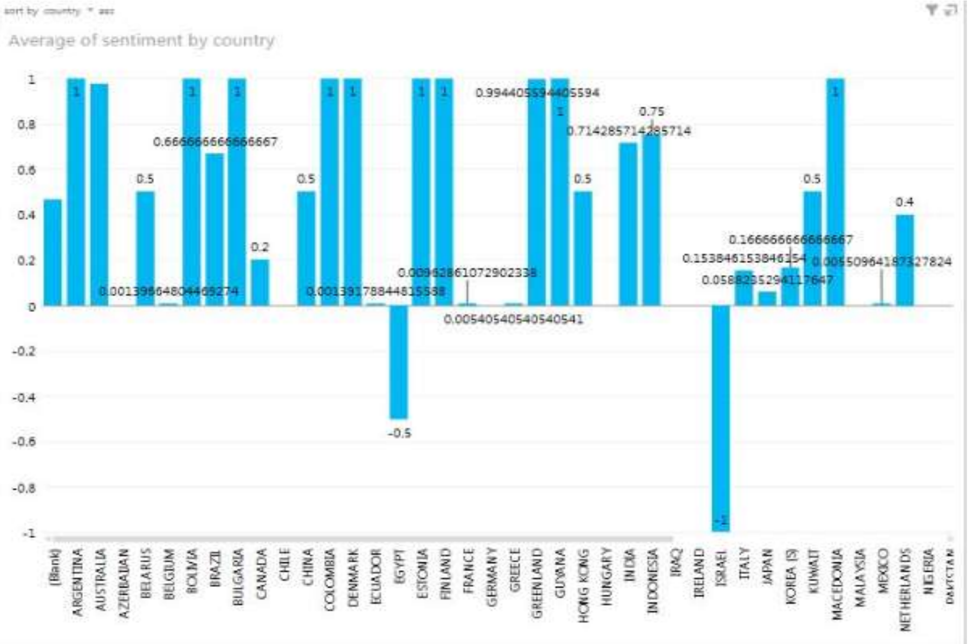


Country-wise Android

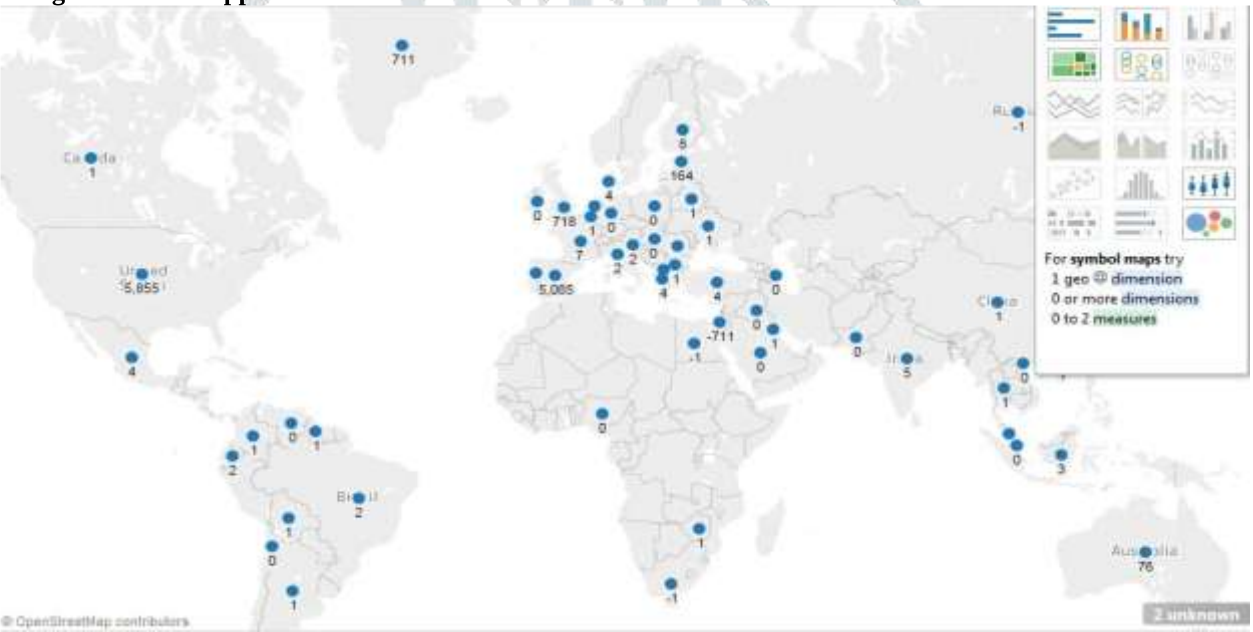
2.



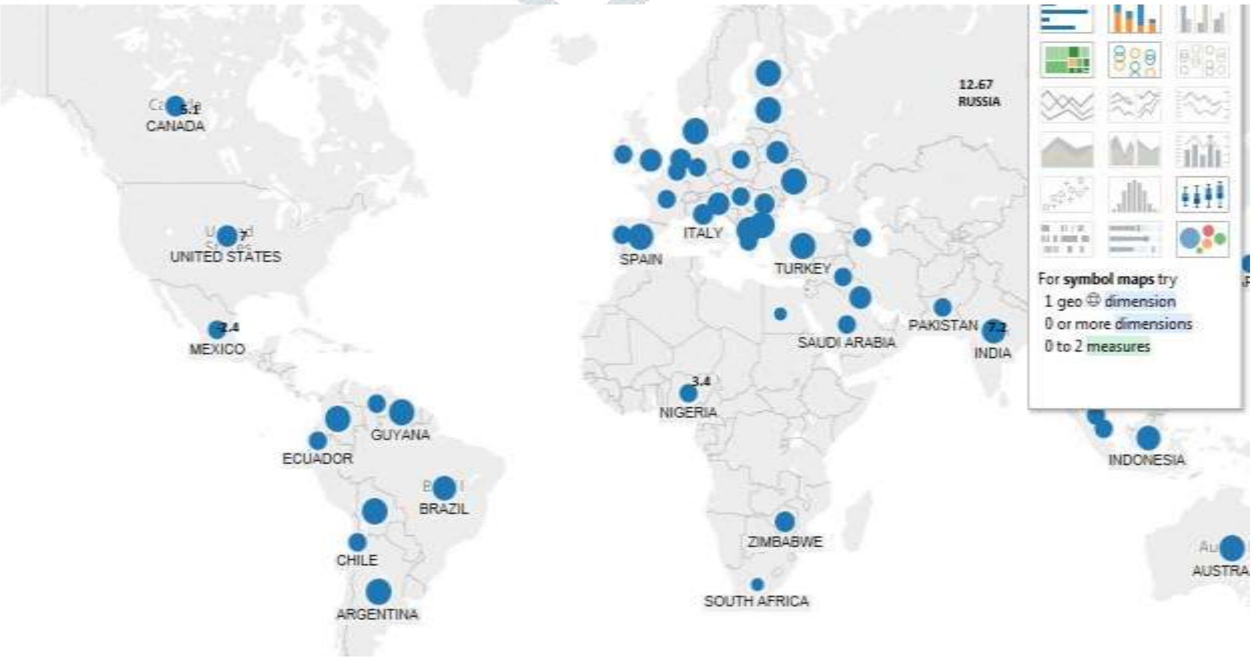
Country-wise Average-tweets for Android



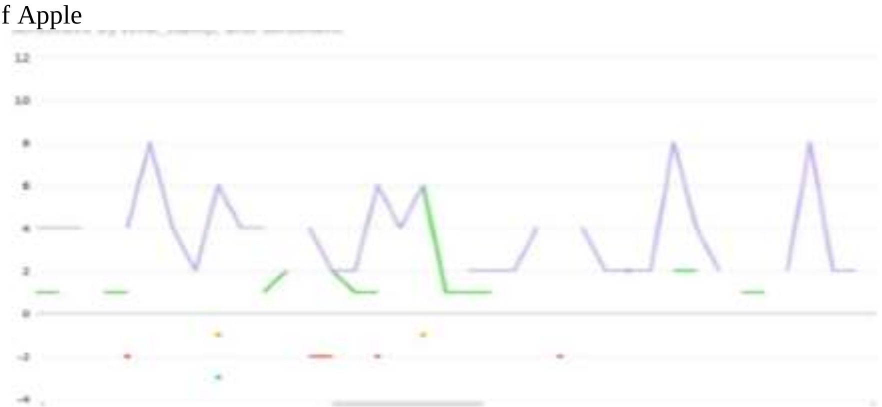
Country-wise Average-tweets for Apple



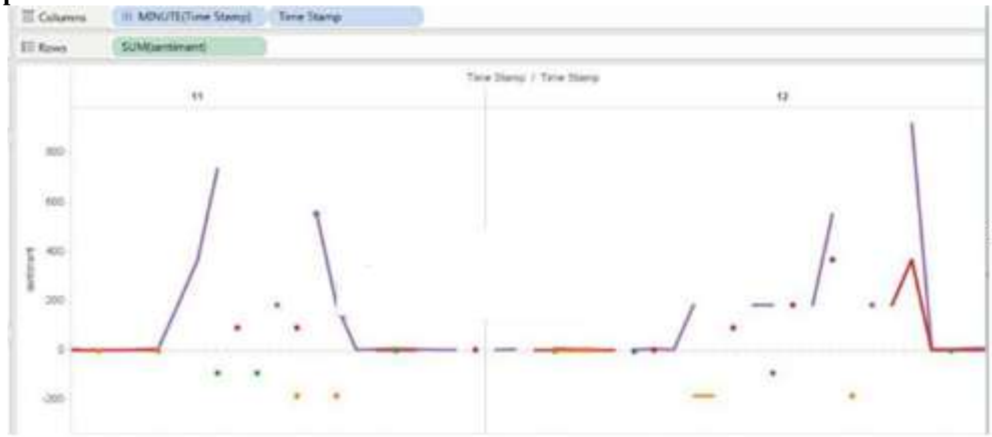
Map for average tweets of android



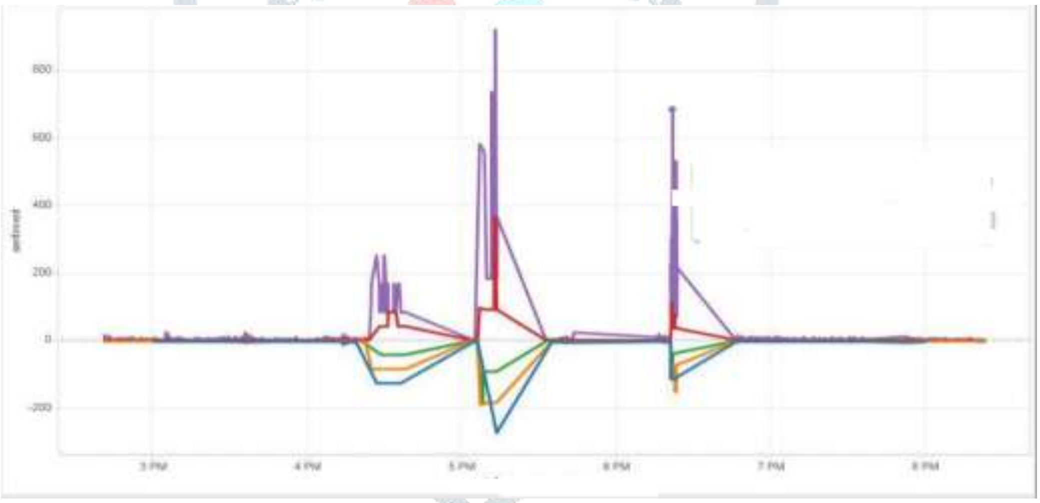
Map for average tweets of Apple



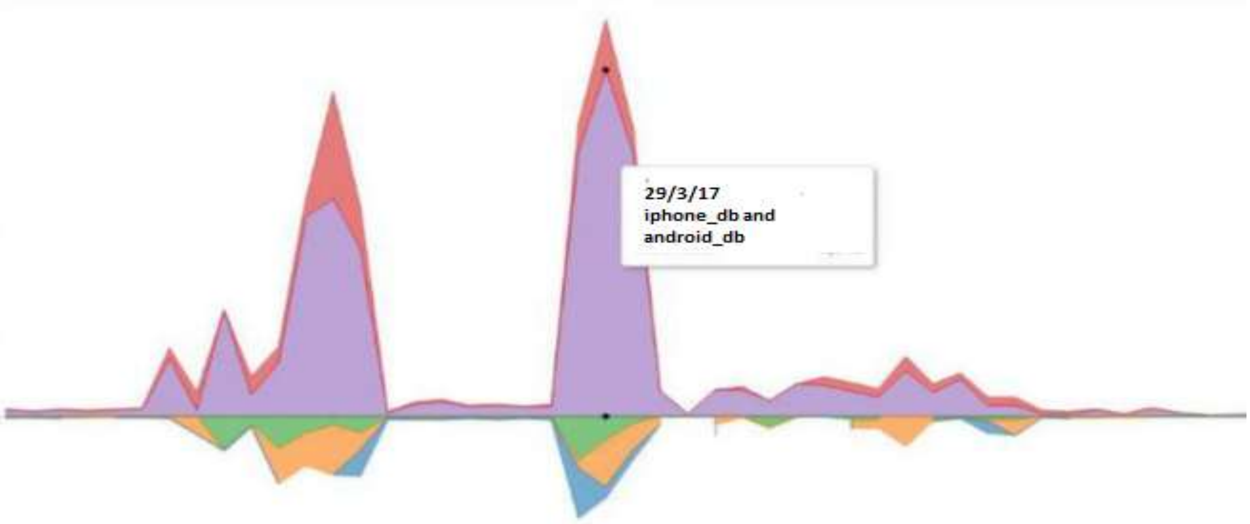
Line chart for Apple-ios



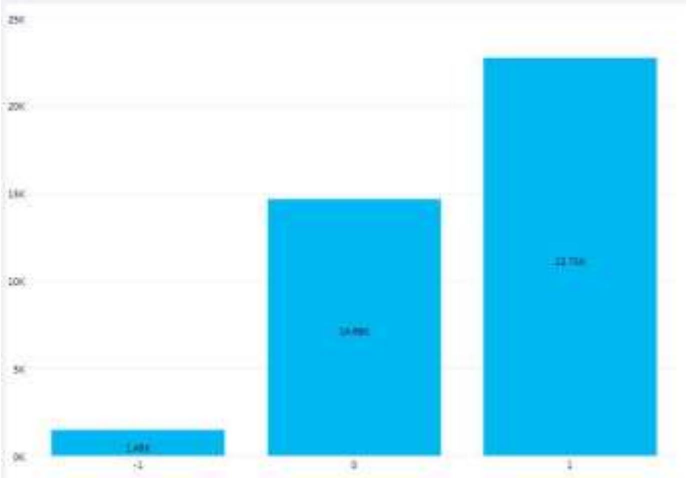
Line chart for Android



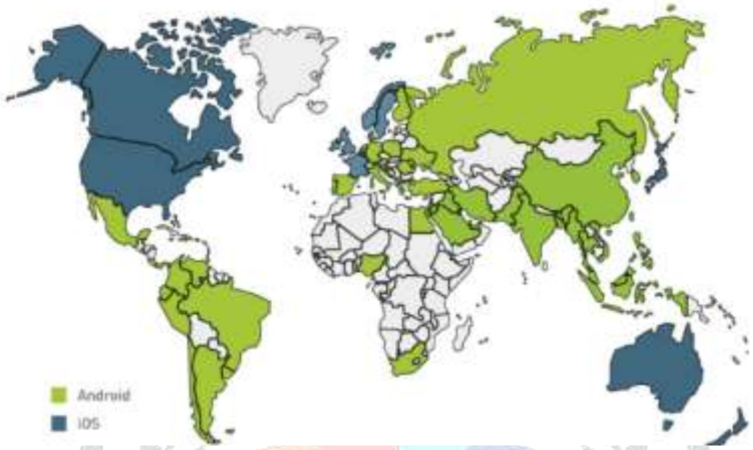
Multi-Line chart for emotion change for Apple and Android



Area diagram for Apple and Android



Overall tweets count for Apple and Android



World-wide dominance of Apple and Android

Interpretation:

Here from the above diagrams, we can see that in USA, Canada and Australia, the iphone users are maximum in number. And in the rest of the world it is dominated by mainly android.

But from the twitter data analysis it is clear that the positive feedback for Apple are also in Greenland, Portugal, Finland and India etc. and negative in Russia, UAE, Nigeria and Mexico etc.

One most interesting finding is that the average tweets of apple is always greater than android for maximum countries. The reason is the retweets count is higher for apple.

The overall satisfaction for apple is positive with a slight negative tone, so I can interpret it as a way that most of the users are happy with apple whereas, android has approx. 3.5:1.2 positive negative tweets that means they have positive as well as negative view for android.

The multi-line chart and area diagram show that the peak of apple is always higher than android that means the apple users are more satisfied with the product and that is why their tweets have a positive tone.

Conclusion: So, from the sentiment analysis and from the comparative study we cant exactly figure out the market share and satisfaction. Some factors like tax, currency fluctuation, geographic premium positions also have a contribution that drive the tone of tweets.

Future Scope: In future I would like to extend my work using real time machine learning and try to map the sentiments with the market-share.

References

[1] Bughin, Jacques, Michael Chui, and James Manyika. "Clouds, big data, and smart assets: Ten tech-enabled business trends to watch." McKinsey Quarterly 56.1 (2010): 75-86.

[2] Das, T. K., D. P. Acharjya, and M. R. Patra. "Opinion mining about a product by analyzing public tweets in Twitter." Computer Communication and Informatics (ICCCI), 2014 International Conference on. IEEE, 2014.

[3] Mehta, Sneha, and Viral Mehta. "Hadoop Ecosystem: An Introduction."

[4] Singh, Neelam, Neha Garg, and Varsha Mittal. "Big Data–insights, motivation and challenges."

[5] Zikopoulos, Paul, and Chris Eaton. Understanding big data: Analytics for enterprise class hadoop and streaming data. McGraw-Hill Osborne Media, 2011.

[6] http://www.datasciencecentral.com/profiles/blog/list?user=0h5qapp2gbuf8

[7] http://www.hindustantimes.com/tech/india-the-most-expensive-place-to-buy-an-iphone-6s-but-why/story-nSDEAr7AynBfI4yBM0hf1O.html