ISSN No. 2349-7165

Analysis of Reviews of an International Movie using Sentiment Analysis: A Qualitative approach.

Ms. Debadrita Panda
(MBA (Business Analytics), IISWBM)
Research Scholar (Burdwan University).
E-mail : debadritapanda1@gmail.com

Mr. DiptoHalder
(MBA (Marketing), IISWBM)
Manager, Mjunction Services Limited.

ABSTRACT

Now a day we are very much reactive about anything. We like to provide feedback. We love to see feedbacks before taking any decision. Even sometimes we make perception by getting influenced by reviews or feedbacks about any product or service.

If we see from the perspective of Marketers reviews can help them in proper targeting. Proper analysis of reviews is no doubt very beneficial for companies to get an extra edge in terms of both competition and profitability.

When it comes to giving feedback or review, it should be expressive means it should be qualitative in terms of data. As per a survey by IBM we create about 2.5 quantillion bytes of data every day. Needless to say it is a Big Data.

Big data is a very popular term that describes the huge growth of data in both structured and unstructured formats. Today unstructured data is larger in volume than structured data. Big data is a collection of data sets both large and complex that is becomes difficult to process using on-hand database management tool.To analyse, the large amount of data and to fetch meaningful information from huge amounts of data for future action is the main challenge for Big data analysis.

Indeed International Movie is now being considered as a popular source of entertainment. Not only this but also people like to give ratings on social medias heavility. With this dynamic and busy world where everyone is only focusing at competitive advantage, this work provide a quick and scientific analysis which can be used by Box offices, this field related Magazines, Blogs to target efficiently to appropriate customer segment.

This study is aims to find the reaction of various customers for an international movie by using Flume for Twitter Streaming and HQL for sentiment analysis and finally Tableau and Excel for visualization.

Keywords: Qualitative research, Big Data, Twitter Streaming, Sentiment Analysis.

Introduction

Big data is a very new technology which is used for processing and storing millions of datasets, which cannot be processed by using traditional database technologies.Processing of Big data using traditional RDMS takes more time,cost and resources.From the 20th century onwards the world wide web has changed the way of expressing their views. People expressing their thought through social media like facebook, twitter, etc. Twitter is a second biggest social network.Twitter users generate million of tweets per hour. Tweets can be categorized by the hashtag for which they are posting their tweets.Many organizations and serve companies are using these tweets for doing some analysis.They can predict the rate of success of their product or they can visualize the data that they have collected for analysis.But, this huge amount of data cannot store and analyzed using traditional database system.

To gain attention of customers and to make them feel customized, businesses are now focusing on user generated content (UGC) analysis. Recreation Industry is playing a very important role in India. As per a report of consulting firm PwC this industry is set to grow at a faster pace of 10.55% CAGR, outshining the global average of 4.2% CAGR. In its annual sector forecast for 2017-2021, PwC said the Indian M&E sector will touch $45.1 billion by 2021, up from $27.3 billion at the end of 2016. To strategize this business area more, Sentiment Analysis has a significant contribution. And needless to say, as the UGC is huge and voluminous sentiment analysis using Big Data is effective in this regards.

This huge and unstructured data can be store and process by using Apache Hadoop Ecosystem,In this paper, Hive and Flume are used for analyzing twitter data to know the emotions of Movie Fans for a hollywood global movie 'Logan'.Here the raw twitter data have been collected by using Apache Flume.It is used for collecting real-time data from Twitter. It is a very popular system for moving a large amount of data into HDFS.Summarization,ad hoc querying and analyzing of large data sets can be done using Hive.

Review of Literature:

According to Chaudhuri et al. 2011; Turban et al. 2008; Watson and Wixom 2007, Big Data Science has its roots in the longstanding database management field. It relies heavily on various data collection, extraction, and analysis technologies. [1] [2] In addition to porting their traditional RDBMS-based product information and business contents online, detailed and IP-specific user search and interaction logs that are collected seamlessly through cookies and server logs have become a new gold mine for understanding customers' needs and identifying new business opportunities. Web intelligence, web analytics, and the user-generated content collected through Web 2.0-based social and crowd-sourcing systems (Doan et al. 2011; O'Reilly 2005) have ushered in a new and exciting era of BI&A 2.0 research in the 2000s, centered on text and web analytics for unstructured web contents. [3] Big Data starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data (HACE theorem).[4] Autonomous data sources with distributed and decentralized controls are a main characteristic of Big Data applications.[5]The main fuel of Big Data is Information. One of the fundamental reasons for Big Data phenomenon to exist is the current extent to which information can be generated and made available.[6] Digitization and datafication have become pervasive phenomena thanks to the broad availability of devices that are both connected and provided with digital sensors. Digital sensors enable digitization while connection lets data be aggregated and, thus, permits datafication. Cisco estimated that between 2008 and 2009 the number of connected devices overtook the number of living people [7] and, according to Gartner [8], by 2020 there will be 26 billion devices on earth, more than 3 devices on average per person. The pervasive presence of a variety of objects (including mobile phones, sensors, Radio-Frequency Identification - RFID - tags, actuators), which are able to interact with each other and cooperate with their neighbors to reach common goals, goes under the name of the Internet of Things, IoT [9,10]. This increasing availability of sensor-enabled, connected devices is equipping companies with extensive information assets from which it is possible to create new business models, improve business processes and reduce costs and risks [11]. In other words, IoT is one of the most promising fuels of Big Data expansion.

Objectives of Paper:

As qualitative data is becoming more and more popular as a Source of User Generated Contents, this paper aims to use these huge pools of data and find out the ultimate sentiment of movie viewers.

It facilitates the way to use reviews, feedbacks on social medias, finally sharpens the targeting and positioning of similar types of movies as well as customization of the taste of movie viewers.

Methodology:

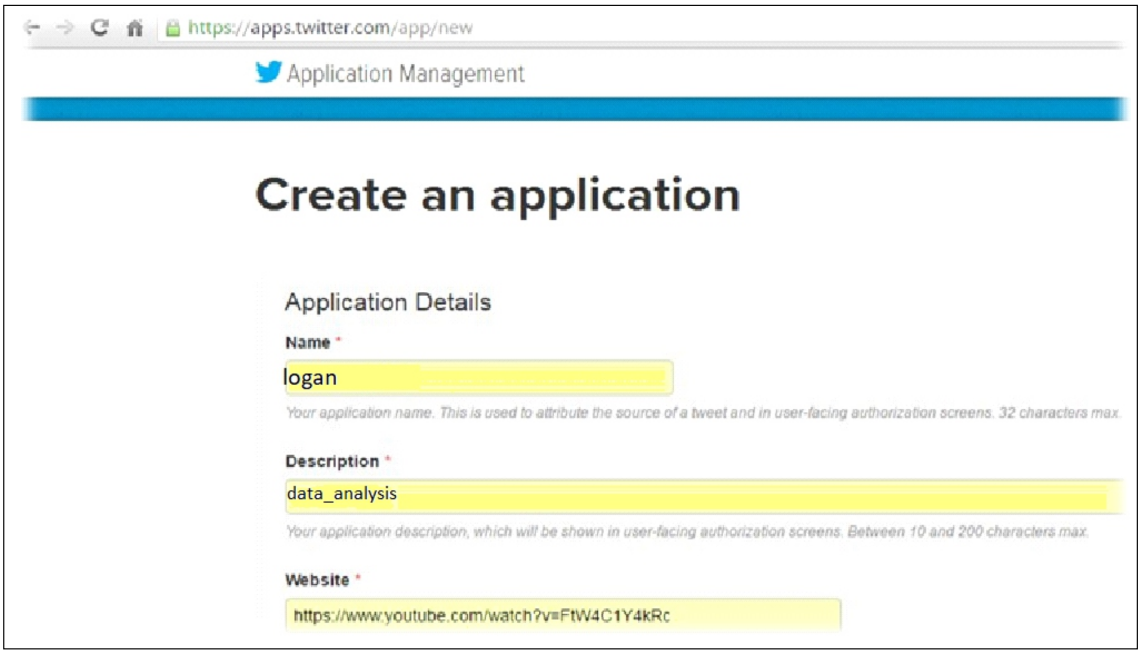
For Twitter data analysis the following Methods have been followed:

- 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

Creating twitter Application programming Interface (API):

To do sentiment analysis on twitter data, first we need twitter data so to get twitter data, we must have an account in twitter developer. Following steps describe how to create Twitter API

- Go to the https://apps.twitter.com/ site.
- Sign in with a Twitter account.
- Create an Application by clicking on new App Button provided by them



https://apps.twitter.com/app/new

Application Management

Create an application

Application Details

Name *

logan

Your application name. This is used to attribute the source of a tweet and in user-facing authorization screens. 32 characters max.

Description *

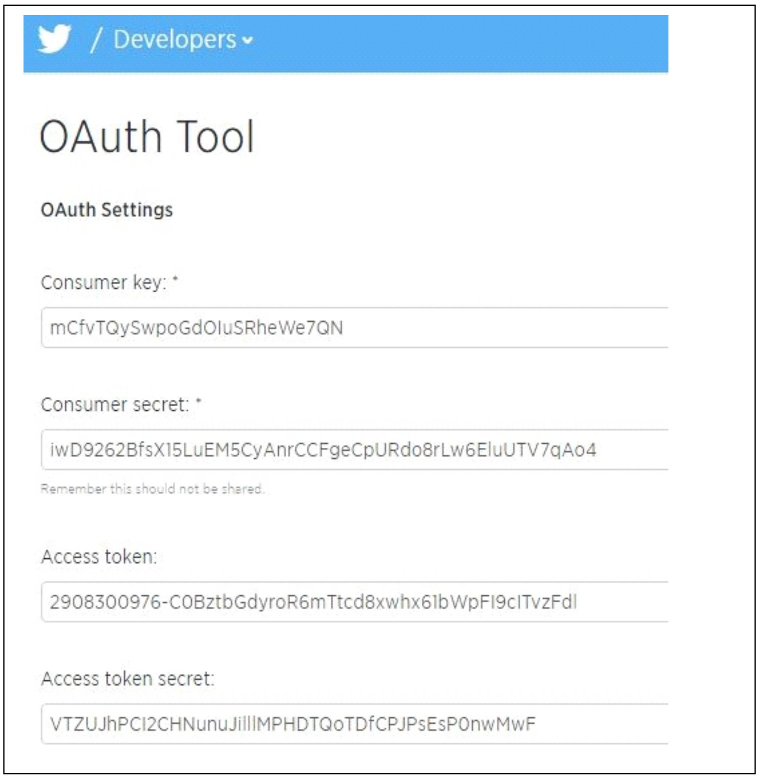
data_analysis

Your application description, which will be shown in user-facing authorization screens. Between 10 and 200 characters max.

Website *

https://www.youtube.com/watch?v=FtW4C1Y4kRc

- Access token secret keys are provided for getting data from twitter. These keys used in Flume configuration file for streaming real-time data from twitter. The following is the figure that shows how the application data looks after creating the application.



OAuth Tool

OAuth Settings

Consumer key: *

mCfvTQySwpoGdOIuSRheWe7QN

Consumer secret: *

iwD9262BfsX15LuEM5CyAnrCCFgeCpURdo8rLw6EluUTV7qAo4

Remember this should not be shared.

Access token:

2908300976-C0BztbGdyroR6mTtcd8xwhx61bWpFI9cITvzFdl

Access token secret:

VTZUJhPCI2CHNunuJilllMPHDTQoTDfCPJPsEsP0nwMwF

Analysis

Collecting data from twitter using Flume:

Apache Flume is one of the components in Hadoop ecosystem used in transferring a 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 architecture is pictorially depicted in Fig:1.

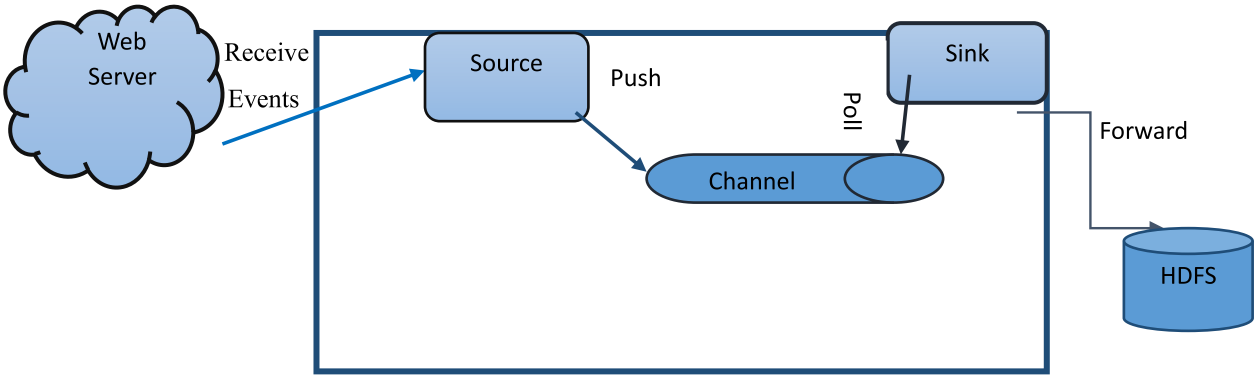


Fig 1: Architecture

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, The configuration is done by 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.
- There can bemultiple agents in Flume. The differentiation of each agent has been done by using a unique name. And using this unique name , the configuration of each flume agents have been completed.
- Naming the Flume Agents: First, the naming is done by the below mentioned way:- flume agent as shown below:

```
Agent_Name.sources =source_Name

 Agent_Name.sink= Sink_Name

 Agent_name.Channels= Channel_name
```

Flume supports various kinds of sources,channels, and sinks. Some of them listed in the following table:

In this project,Data streaming is done using twitter source through a memory channel to an HDFS sink.Here agent name id is TwitterAgent,

TwitterAgent.sources = Twitter
TwitterAgent.channels = MemChannel
TwitterAgent.sinks = HDFS

Configuration of the source:

After listing the components of the agent, sources, sinks, and channels by providing values to their properties are described .Each source has a separate list of properties.The property name "type" used to specify the type of sources we are using.Along with property "type" ,the values of all required properties of a particular source are needed for configuration . In this analysis source is twitter.following are the properties to which values must be provided to configure it.

```
TwitterAgent.sources.Twitter.type = Twitter (type name)
TwitterAgent.sources.Twitter.consumerKey =  "value"
TwitterAgent.sources.Twitter.consumerSecret = "value"
TwitterAgent.sources.Twitter.accessToken =   "value"
TwitterAgent.sources.Twitter.accessTokenSecret ="value"
```

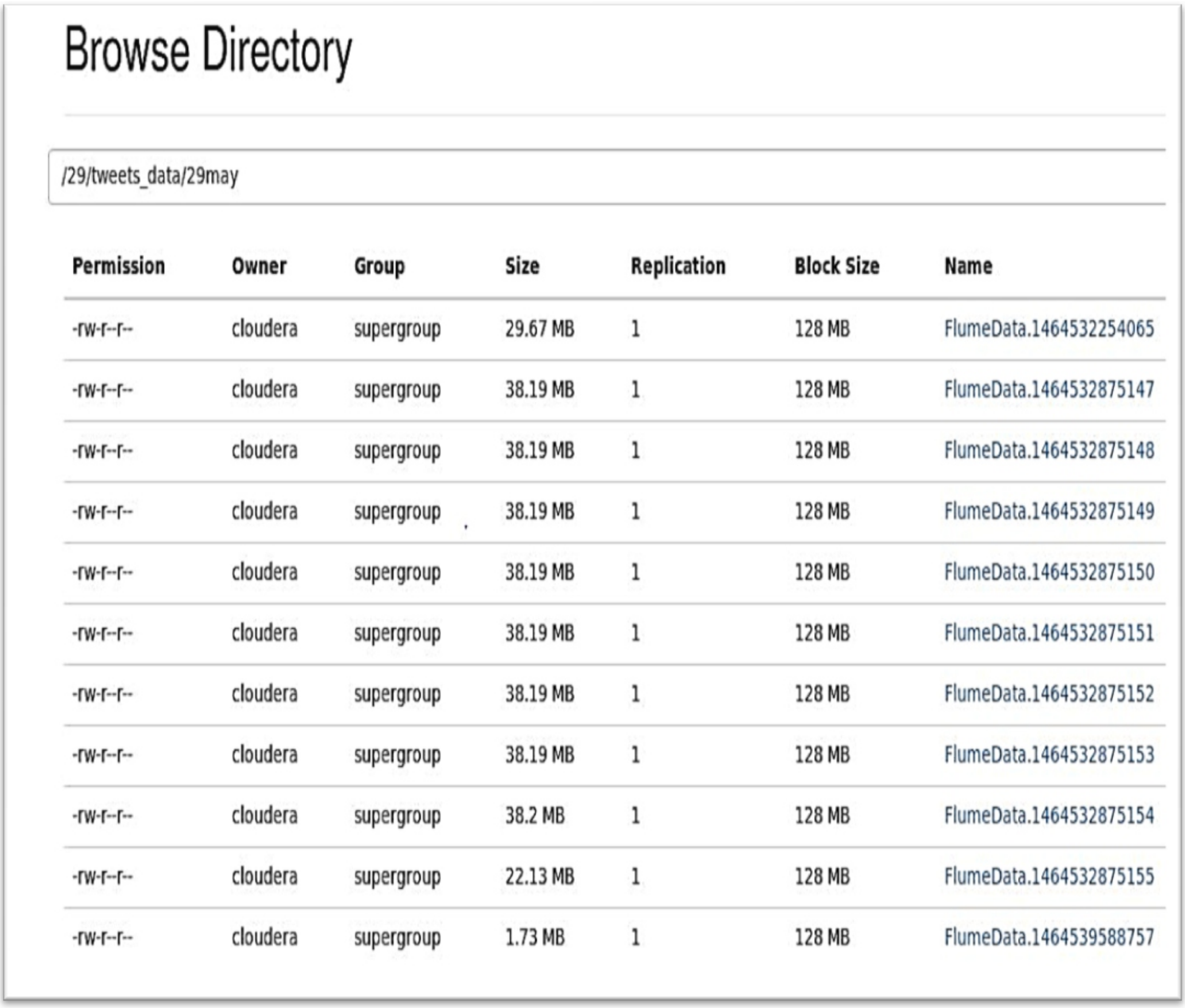
TABLE . flume.conf files for Twitter data streaming on hashtag "Logan"

```
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=DmBLDVPGe0GLO37rSUNeSJK7oVlvaD3RwKKI9h
1rnCD33P06IB

TwitterAgent.sources.Twitter.accessToken=102368189
-bjD3VbSMdBOeyoQ93DGLYZvfS04jmYDoGV3dJelR

TwitterAgent.sources.Twitter.accessTokenSecret=4f3wEp3X1mOZBqavjbVEBLqzcibkU7c0ce1lTUEc
r4H0t

TwitterAgent.sources.Twitter.keywords= #Logan

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
```

Unique Keys

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.



Browse Directory

/29/tweets_data/29may

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
-rw-r--r--	cloudera	supergroup	22.13 MB	1	128 MB	FlumeData.1464532875155
-rw-r--r--	cloudera	supergroup	1.73 MB	1	128 MB	FlumeData.1464539588757

Fig-2: Twitter data in Hadoop Distributed File System (HDFS)

From these data first, a table has been created 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. It can be said that the data have been converted from unstructured to 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.The concepts of serDe arethat 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
```

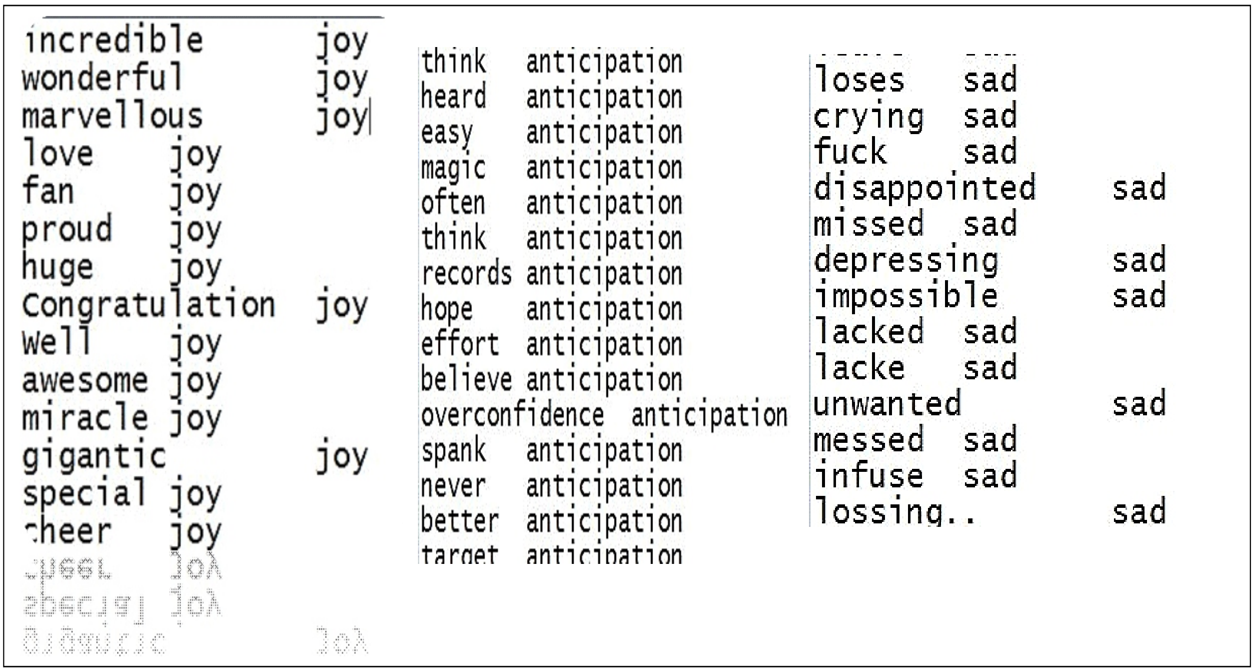
Creation of learning sheet from row data (unstructured data)

For Dictionary based machine learning technique, a data set need to be created that contain the movie reviewers' positive, negative reactions. In this analysis, categorizations of tweets are done into three categories namely positive, negative and neutral sentiments and every sentiment is assigned with some weights.

TABLE: Sample Emotion and text in Tweets.

Sentiments types	Dictionary Sample	Tweets Sample
Positive	Incredible	#nolan is incredible
Negative	Bored	Boring movie #nolan
Neutral	Common	Expectation is very comm on from #nolan like other wolverine series movie.

The following tables show the sample of learning sheet which we have created to perform



incredible	joy
wonderful	joy
marvellous	joy
love	joy
fan	joy
proud	joy
huge	joy
Congratulation	joy
Well	joy
awesome	joy
miracle	joy
gigantic	joy
special	joy
cheer	joy

think	anticipation
heard	anticipation
easy	anticipation
magic	anticipation
often	anticipation
think	anticipation
records	anticipation
hope	anticipation
effort	anticipation
believe	anticipation
overconfidence	anticipation
spank	anticipation
never	anticipation
better	anticipation
target	anticipation

loses	sad
crying	sad
fuck	sad
disappointed	sad
missed	sad
depressing	sad
impossible	sad
lacked	sad
lacke	sad
unwanted	sad
messed	sad
infuse	sad
lossing..	sad

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

After retrieving tweets,we have put schema over Flume data by using hive query.we 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 joy,anticipation,fear,anger, and sad words. From these words, we give a score to each of the words as the joy word "+2" , the anticipation word "+1",the anger word "-3",the fear word "-2" , and the sad word "-1" . The process pictorially depicted in Fig-2

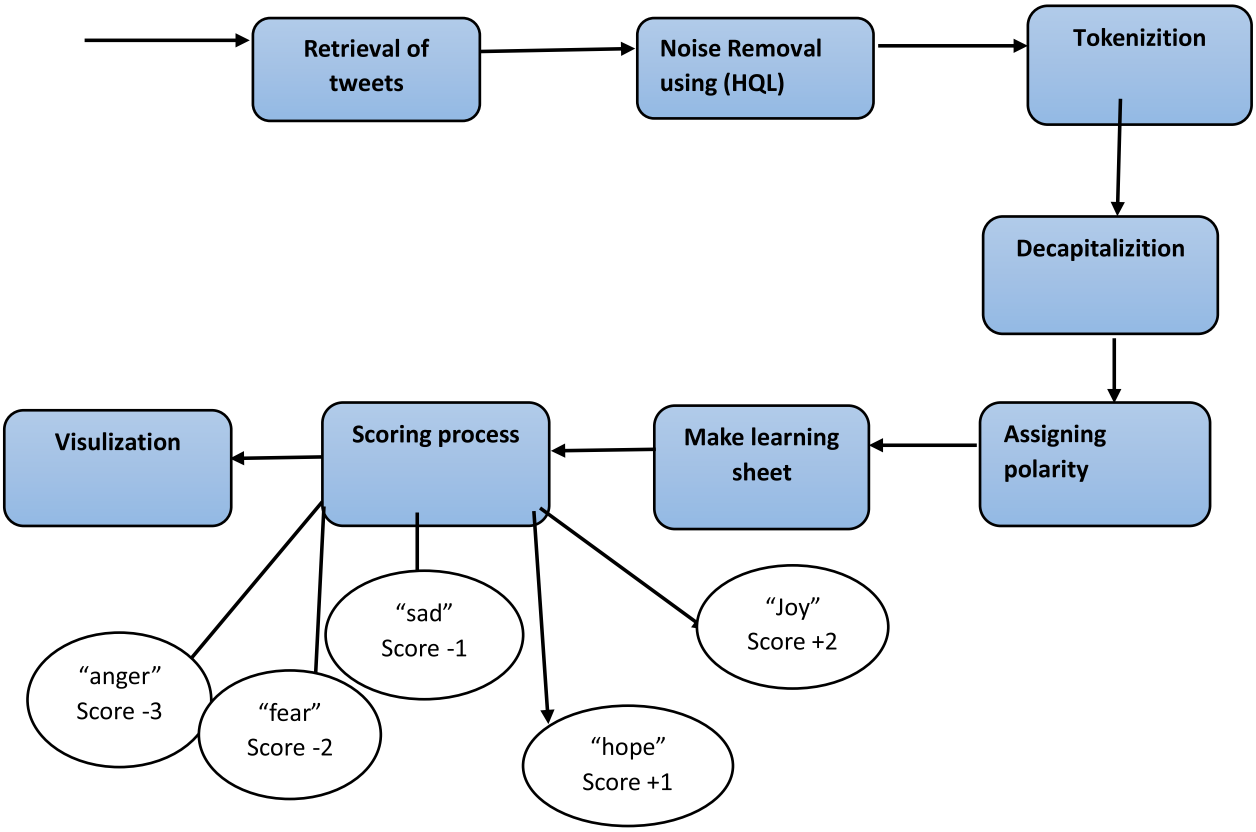


Fig-3: Flowchart of the entire process

We processed all tweets as follow:

I. **Retrieval of twitter data:** Using Twitter Application Programming Interface (API)I have retrieved the tweets during March 2017. I have used Flume to retrieve the tweets.

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

III. **Schematization of learning sheet:** I have schematized the learning sheet using Hive Query Language to give schema according to our requirements.There are two attributes in our 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:** Here we filtered out some information from raw tweets which are required for our analysis such as user Id,Timestamp,Text,and user's time-zone using HQL.This process depicted in snapshot1.
- **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.

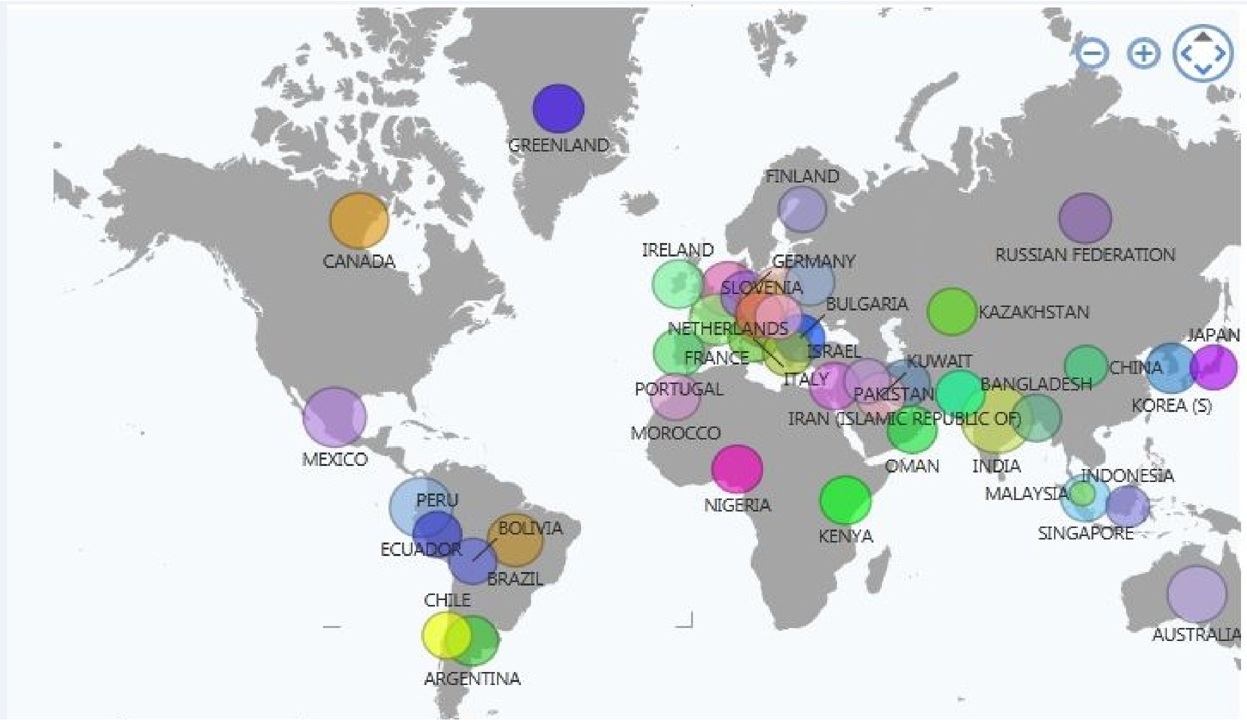
V. **Scoring:** After processing, we get only accurate and meaningful tweets. Now tweets are compared with learning sheet to assign a polarity to each word.

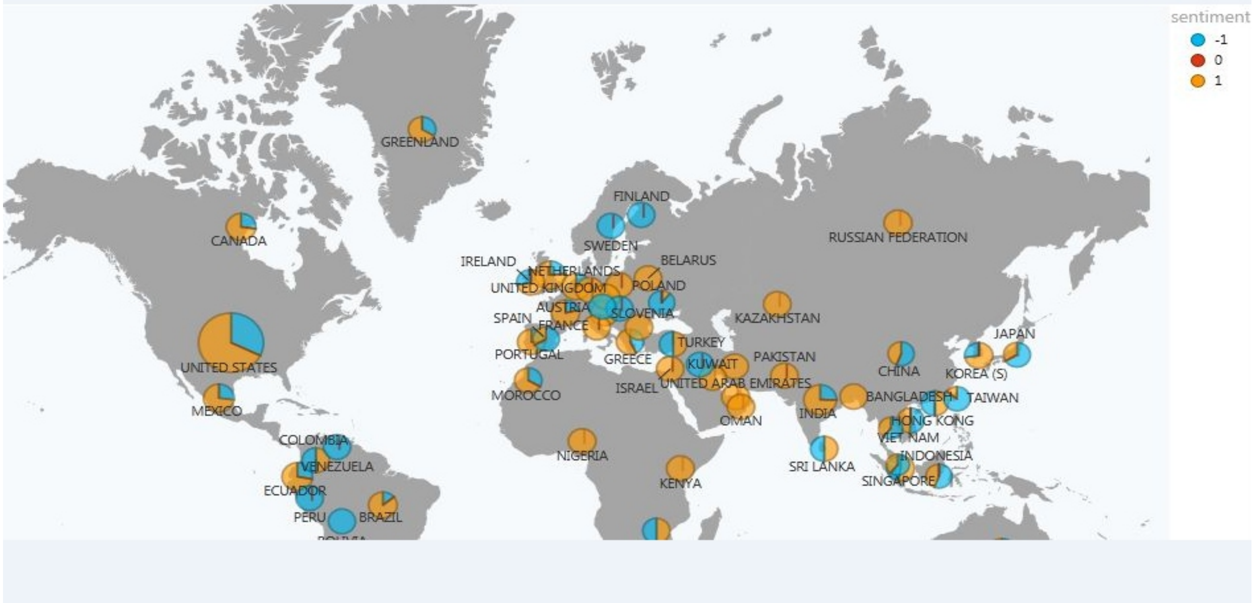
Input: Tweets, Learning_sheet Output: Sentiment_score (joy, anticipation, fear, anger, and sad), timestamp, and id

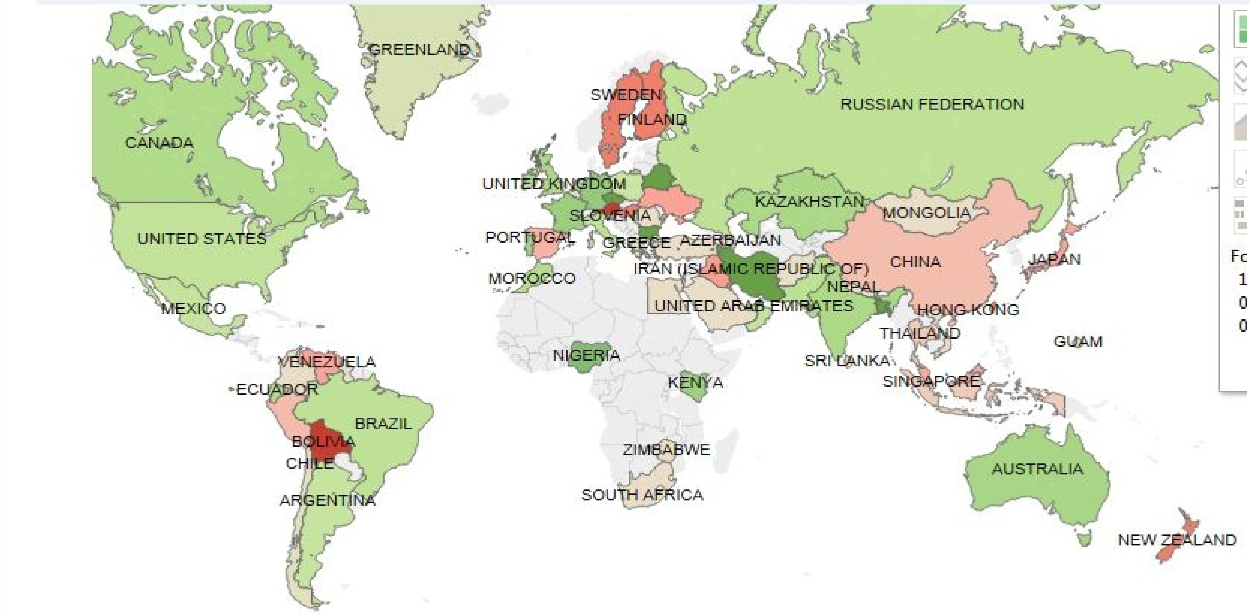
```
Step 1.  Begin
Step 2.  Schematize Twitter Data
Step 3. Retrieve (id, time_stamp, text, user.time_zone)
Step 4. For each tweet Ti
Step 5.  For each Ti .id
Step 6.  Retrieve (id.timestamp, id.text)
Step 7. Converts the id.text to lowercase for standardization Step 8. Array [id.text] = Store id.text into array
Step 9. Array [id.word] = Tokenize (Array [id.text]) End Loop
Step 10. /* assigning polarities */
Step 11.If (Learning_sheet.word ==Ti.word)
                  If (Array [id.word] == Learning_sheet.word)
                       When Learning_sheet.polarity=”joy”
                          THEN Assign id.polarity= 2
             When Learning_sheet.polarity= “anticipation”
                          THEN Assign id.polarity= 1
             When Learning_sheet.polarity =” anger”
                          THEN Assign id.polarity= -3
             When Learning_sheet.polarity= “fear”
                          THEN Assign id.polarity= -2
             When Learning_sheet.polarity= “sad”
                          THEN Assign id.polarity= -1
             End Loop
End Loop
Step 12. For Each id.polarity
Step 13. If (id.polarity==2)
                          THEN Assign id.sentiment = “joy”
             Else if (id.polarity==1)
                          THEN Assign id.sentiment = “anticipation”
             Else if (id.polarity==-3)
                          THEN Assign id.sentiment = “anger”
             Else if (id.polarity==-3)
                          THEN Assign id.sentiment = “anger”
             Else if (id.polarity==-2)
```

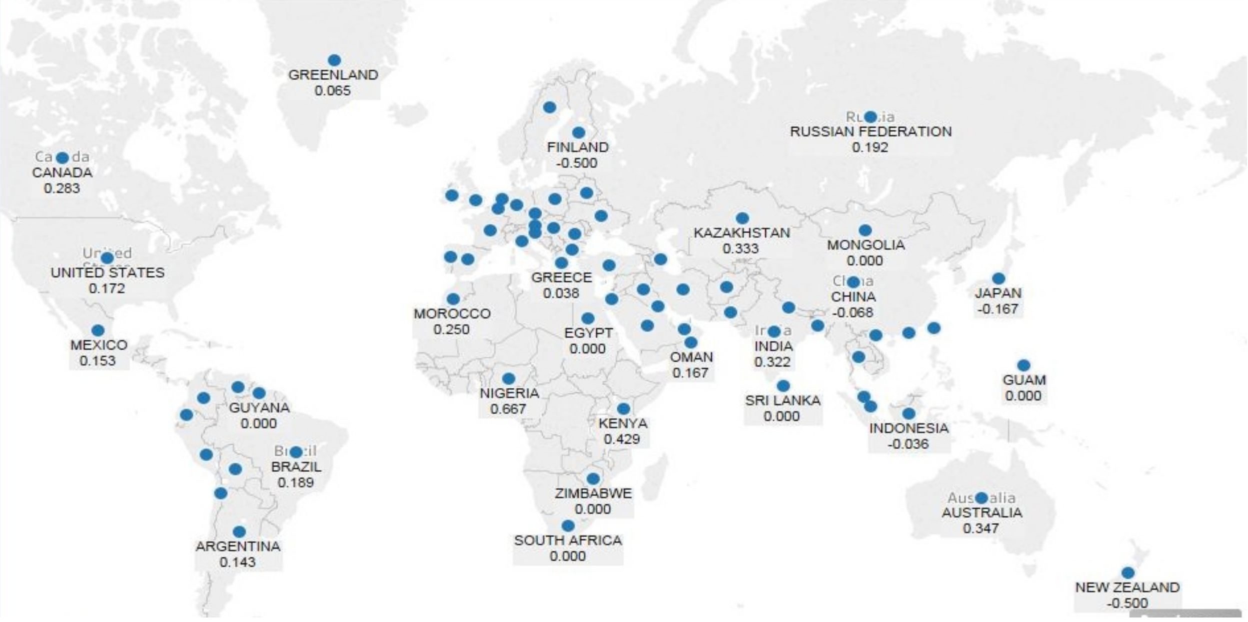
```
                THEN Assign id.sentiment = "fear"
        Else if (id.polarity==-1)
                THEN Assign id.sentiment = "sad"
        End Loop
End Loop
Step 14. For each id.sentiment assign weight Step 15. N=count (id)
Step 16. For 1 to N
Step 17.        When (id.sentiment == "joy")
                THEN Assign id.sentiment.weight=2 id.sentiment.weight = $\sum$1id. sentiment. weight
Step 18.        When (id.sentiment == "anticipation")
                THEN Assign id.sentiment.weight=1 id.sentiment.weight = $\sum$1id. sentiment. weight
Step 19.        When (id.sentiment == "anger")
                THEN Assign id.sentiment.weight=-3 id.sentiment.weight = $\sum$1id. sentiment. weight

Step 20.        When (id.sentiment == "fear")
                THEN Assign id.sentiment.weight=-2 id.sentiment.weight = $\sum$1id. sentiment. weight

Step 21.        When (id.sentiment == "sad")
                THEN Assign id.sentiment.weight=-1 id.sentiment.weight = $\sum$1id. sentiment. weight
        End Loop End
                Loop
```

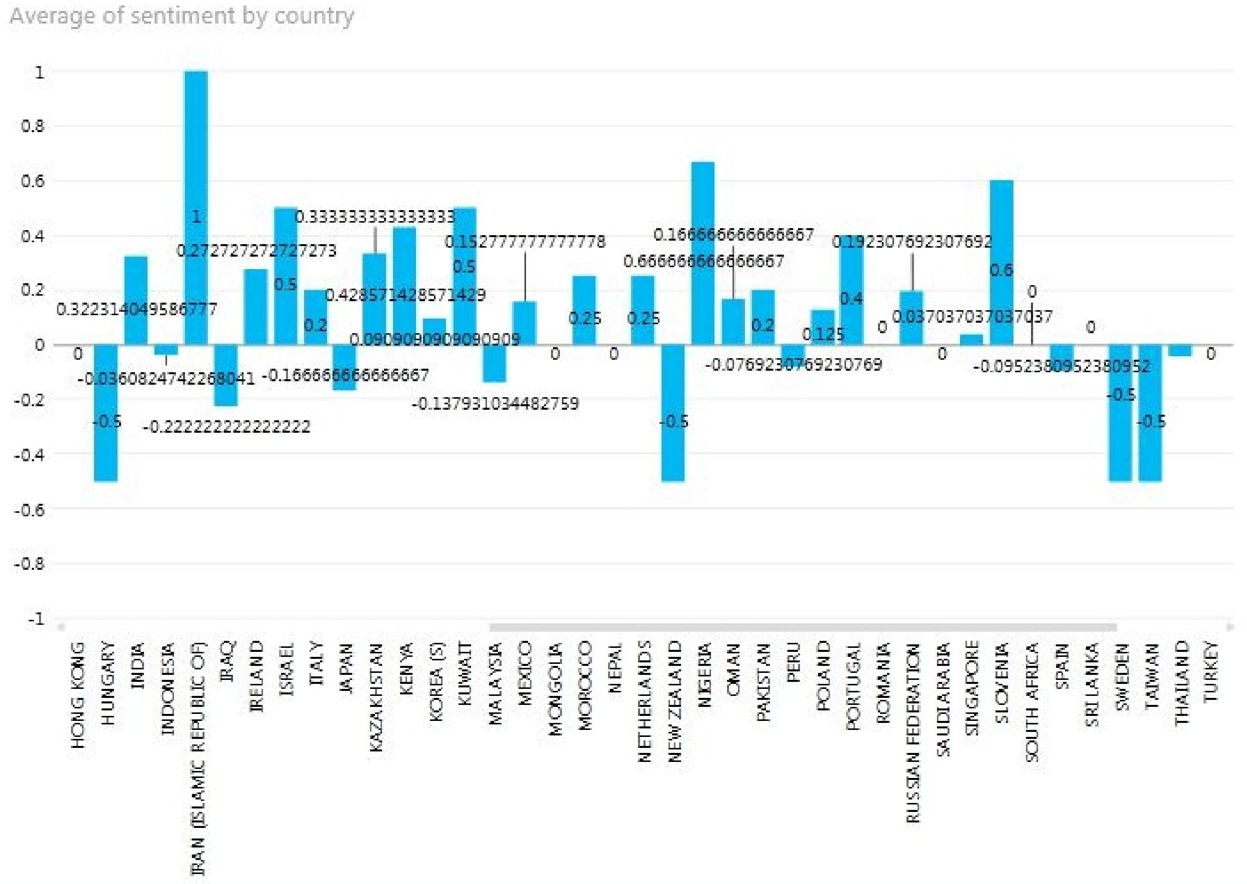
Findings

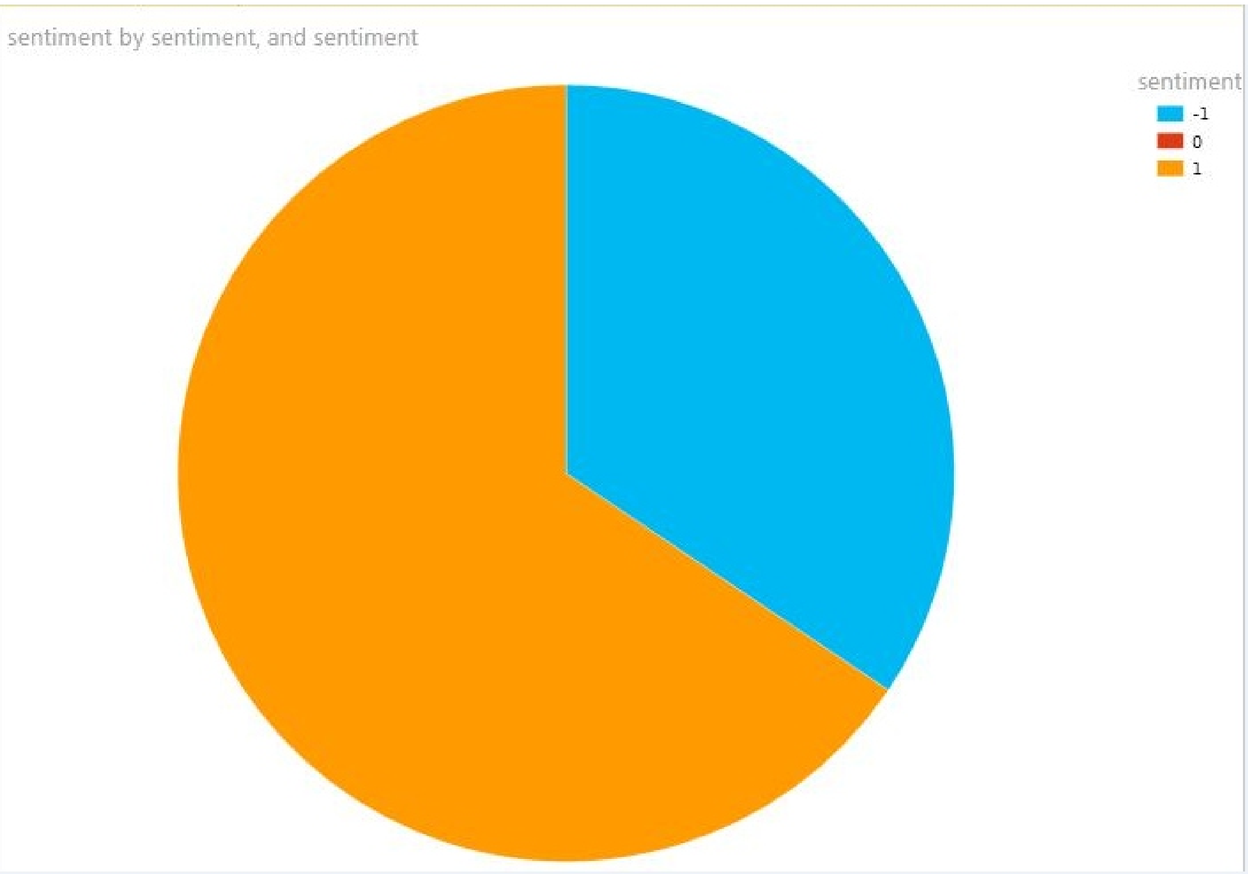


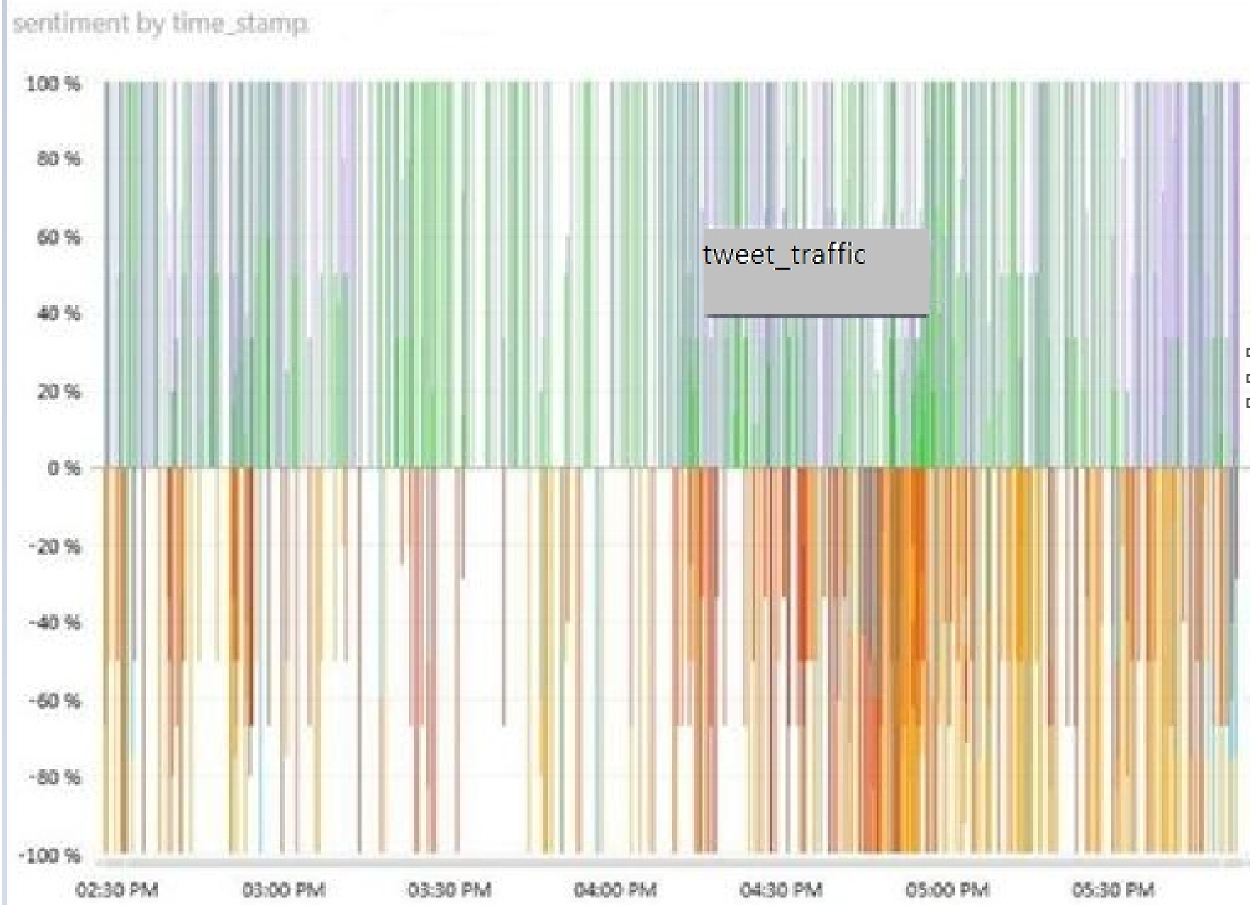

sentiment
● -1
● 0
● 1
GREENLAND
FINLAND
CANADA
SWEDEN
RUSSIAN FEDERATION
IRELAND
NETHERLANDS
BELARUS
UNITED KINGDOM
POLAND
AUSTRIA
SPAIN
FRANCE
SLOVENIA
KAZAKHSTAN
JAPAN
UNITED STATES
PORTUGAL
TURKEY
GREECE
KUWAIT
PAKISTAN
CHINA
KOREA (S)
ISRAEL
UNITED ARAB EMIRATES
MOROCCO
MEXICO
INDIA
BANGLADESH
TAIWAN
OMAN
HONG KONG
COLOMBIA
VIET NAM
NIGERIA
VENEZUELA
INDONESIA
SRI LANKA
ECUADOR
KENYA
SINGAPORE
PERU
BRAZIL
BOLIVIA


GREENLAND
SWEDEN
FINLAND
RUSSIAN FEDERATION
CANADA
UNITED KINGDOM
KAZAKHSTAN
MONGOLIA
SLOVENIA
UNITED STATES
PORTUGAL
GREECE
AZERBAIJAN
CHINA
JAPAN
IRAN (ISLAMIC REPUBLIC OF)
MOROCCO
NEPAL
MEXICO
UNITED ARAB EMIRATES
HONG KONG
THAILAND
GUAM
NIGERIA
SRI LANKA
VENEZUELA
ECUADOR
KENYA
SINGAPORE
BRAZIL
BOLIVIA
CHILE
ZIMBABWE
AUSTRALIA
ARGENTINA
SOUTH AFRICA
NEW ZEALAND
Fc
1
0
0


GREENLAND 0.065
Canada
CANADA 0.283
FINLAND -0.500
Russia
RUSSIAN FEDERATION 0.192
United
UNITED STATES 0.172
KAZAKHSTAN 0.333
MONGOLIA 0.000
GREECE 0.038
China
CHINA -0.068
JAPAN -0.167
MOROCCO 0.250
EGYPT 0.000
OMAN 0.167
India
INDIA 0.322
MEXICO 0.153
GUAM 0.000
NIGERIA 0.667
SRI LANKA 0.000
GUYANA 0.000
KENYA 0.429
INDONESIA -0.036
Brazil
BRAZIL 0.189
ZIMBABWE 0.000
Australia
AUSTRALIA 0.347
ARGENTINA 0.143
SOUTH AFRICA 0.000
NEW ZEALAND -0.500


Average of sentiment by country
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
0
-0.5
0.322314049586777
-0.0360824742268041
1
-0.222222222222222
0.272727272727273
0.5
0.2
-0.166666666666667
0.333333333333333
0.428571428571429
0.0909090909090909
0.5
-0.137931034482759
0.152777777777778
0
0.25
0
0.25
-0.5
0.666666666666667
0.166666666666667
0.2
-0.0769230769230769
0.125
0.4
0
0.192307692307692
0
0.037037037037037
0.6
0
-0.0952380952380952
0
-0.5
-0.5
0
HONG KONG
HUNGARY
INDIA
INDONESIA
IRAN (ISLAMIC REPUBLIC OF)
IRAQ
IRELAND
ISRAEL
ITALY
JAPAN
KAZAKHSTAN
KENYA
KOREA (S)
KUWAIT
MALAYSIA
MEXICO
MONGOLIA
MOROCCO
NEPAL
NETHERLANDS
NEW ZEALAND
NIGERIA
OMAN
PAKISTAN
PERU
POLAND
PORTUGAL
ROMANIA
RUSSIAN FEDERATION
SAUDI ARABIA
SINGAPORE
SLOVENIA
SOUTH AFRICA
SPAIN
SRI LANKA
SWEDEN
TAIWAN
THAILAND
TURKEY


sentiment by sentiment, and sentiment
sentiment
-1
0
1



Interpretation:

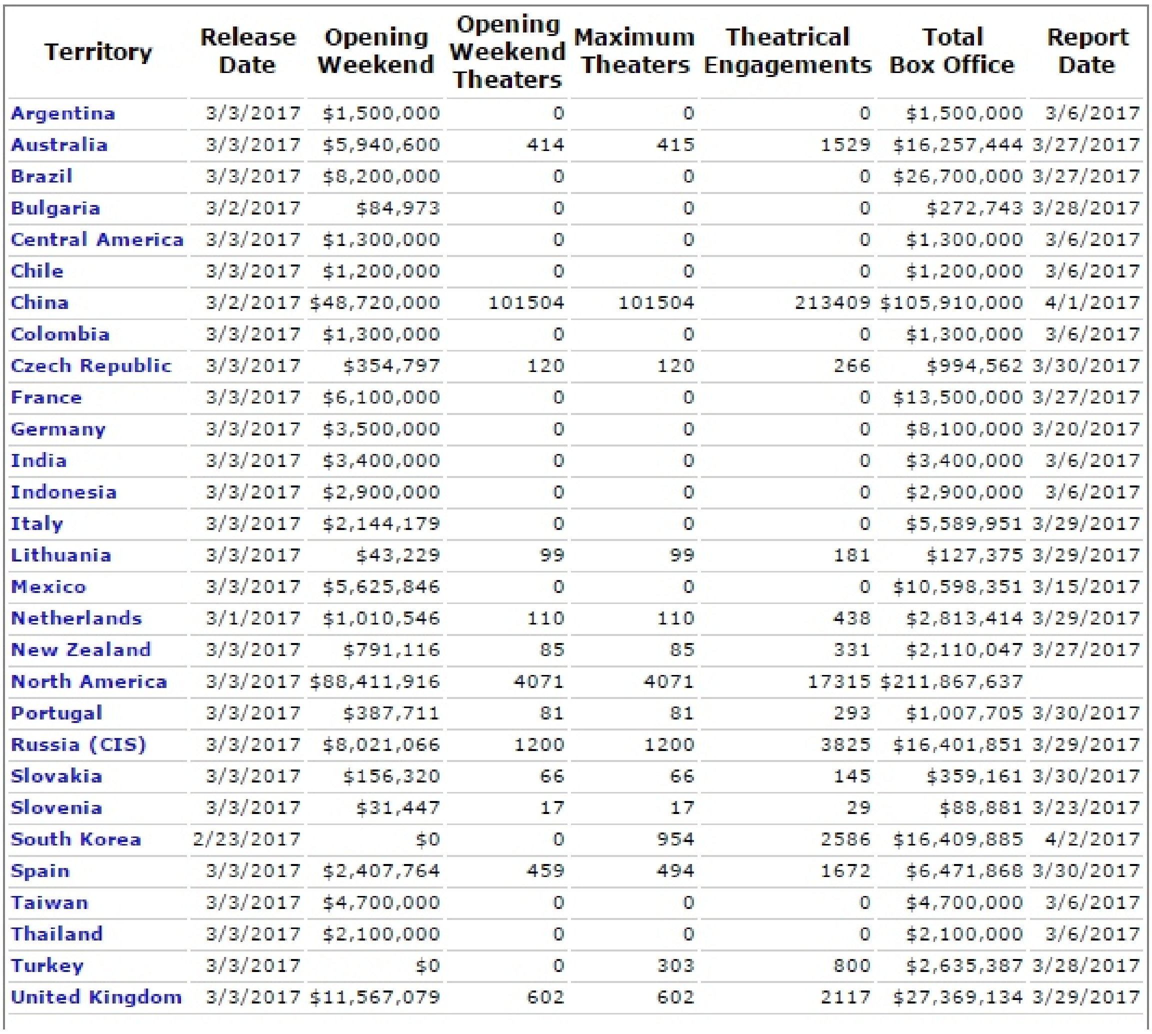
I have analyzed near about 1,00,000 tweets which is on behalf of Logan movie and here I have got some amazing results by doing sentiment analysis compare to the overall worlwide Box-office collection of the movie.

Here,from the sentiment analysis we can see that there is a negative sentiment present in the countries like China,Chile, New-Zeland,Thailand,Bolivia, Finland etc and the positive sentiment is present in the countries like USA, Canada, Brazil,UK, Greece,Rusia etc.

Although China has a negative sentiment for the movie but from the above data we can see that it has produced the most box-office collection and which is greater than UK,and Brazil which in-terms have a positive sentiment for this movie.

So, we can say that from the sentiment analysis, we can't always predict the Box-office collection of a movie as because it depends on various factors like size of the country, population of the country, number of theatres in which the movie is released etc.These factors have a great effect on Box-office collection. After watching any movie the movie fan could be unhappy and he/she could dislike the story, script or acting but due to these above mentioned factors a movie can generate high Box-office collection which can be in contrast with the sentiment analysis.

Russia is also another good example of this contradiction. It has a positive sentiment for the movie and it is also in the top 10 countries in terms of Box-office collection.



Territory	Release Date	Opening Weekend	Opening Weekend Theaters	Maximum Theaters	Theatrical Engagements	Total Box Office	Report Date
Argentina	3/3/2017	$1,500,000	0	0	0	$1,500,000	3/6/2017
Australia	3/3/2017	$5,940,600	414	415	1529	$16,257,444	3/27/2017
Brazil	3/3/2017	$8,200,000	0	0	0	$26,700,000	3/27/2017
Bulgaria	3/2/2017	$84,973	0	0	0	$272,743	3/28/2017
Central America	3/3/2017	$1,300,000	0	0	0	$1,300,000	3/6/2017
Chile	3/3/2017	$1,200,000	0	0	0	$1,200,000	3/6/2017
China	3/2/2017	$48,720,000	101504	101504	213409	$105,910,000	4/1/2017
Colombia	3/3/2017	$1,300,000	0	0	0	$1,300,000	3/6/2017
Czech Republic	3/3/2017	$354,797	120	120	266	$994,562	3/30/2017
France	3/3/2017	$6,100,000	0	0	0	$13,500,000	3/27/2017
Germany	3/3/2017	$3,500,000	0	0	0	$8,100,000	3/20/2017
India	3/3/2017	$3,400,000	0	0	0	$3,400,000	3/6/2017
Indonesia	3/3/2017	$2,900,000	0	0	0	$2,900,000	3/6/2017
Italy	3/3/2017	$2,144,179	0	0	0	$5,589,951	3/29/2017
Lithuania	3/3/2017	$43,229	99	99	181	$127,375	3/29/2017
Mexico	3/3/2017	$5,625,846	0	0	0	$10,598,351	3/15/2017
Netherlands	3/1/2017	$1,010,546	110	110	438	$2,813,414	3/29/2017
New Zealand	3/3/2017	$791,116	85	85	331	$2,110,047	3/27/2017
North America	3/3/2017	$88,411,916	4071	4071	17315	$211,867,637	
Portugal	3/3/2017	$387,711	81	81	293	$1,007,705	3/30/2017
Russia (CIS)	3/3/2017	$8,021,066	1200	1200	3825	$16,401,851	3/29/2017
Slovakia	3/3/2017	$156,320	66	66	145	$359,161	3/30/2017
Slovenia	3/3/2017	$31,447	17	17	29	$88,881	3/23/2017
South Korea	2/23/2017	$0	0	954	2586	$16,409,885	4/2/2017
Spain	3/3/2017	$2,407,764	459	494	1672	$6,471,868	3/30/2017
Taiwan	3/3/2017	$4,700,000	0	0	0	$4,700,000	3/6/2017
Thailand	3/3/2017	$2,100,000	0	0	0	$2,100,000	3/6/2017
Turkey	3/3/2017	$0	0	303	800	$2,635,387	3/28/2017
United Kingdom	3/3/2017	$11,567,079	602	602	2117	$27,369,134	3/29/2017

Another interesting pattern is that during 4pm to 6 pm the twitter traffic is very high.

Conclusion

In our project, we have discussed the sentiment analysis and sentiment mining in detail. Sentiment Analysis deals with understanding whether the expressed opinion about the entity has a positive or a negative orientation. We observed that opinion mining helps to get better insight once the particular thing is analyzed.

Future scope

In this project, we have used “Dictionary” based approach for Sentiment Analysis. In future, we would like to extend our work by incorporating machine learning in my analysis. For that, we will use the “Apache Mahout”, the machine learning and data mining framework of the Hadoop Eco System. As the same, we would like to shift our work to “Apache Spark” cluster for faster and real-time analysis.

References

1. Hsinchun Chen, Roger H. L. Chiang, Veda C. Storey in BUSINESS INTELLIGENCE AND ANALYTICS: FROM BIG DATA TO BIG IMPACT

2. Kotwal, Aishwarya, et al. "Improvement in Sentiment Analysis of Twitter Data Using Hadoop." Imperial Journal of Interdisciplinary Research 2.7 (2016).

3. Samariya, Durgesh, et al. "A Hybrid Approach for Big Data Analysis of Cricket Fan Sentiments in Twitter." Proceedings of International Conference on ICT for Sustainable Development. Springer Singapore, 2016.

4. Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding in Data Mining with Big Data

5. F. B. Viegas, M. Wattenberg, F. Van Ham, J. Kriss, and M. McKeon, IEEE Trans. Vis. Comput. Graph. 13, 1121 (2007).

6. Andrea De Mauro, Marco Greco, and Michele Grimaldi in What is big data? A consensual definition and a review of key research topics.

7. J. Manyika, M. Chui, B. Brown, and J. Bughin, Big Data: The next Frontier for Innovation, Competition, and Productivity (2011).

8. W. Xiong, Z. Yu, Z. Bei, J. Zhao, F. Zhang, Y. Zou, X. Bai, Y. Li, and C. Xu, in Big Data, 2013 IEEE Int.Conf. (2013), pp. 118–125.

9. T. Pearson and R. Wegener, Big Data: The Organizational Challenge, Bain & Company report (2013).

10. J. Dijcks, Big Data for the Enterprise, Oracle report (2012).

11. M. Schroeck, R. Shockley, J. Smart, D. Romero-Morales, and P. Tufano, Analytics: The Real-World Use of Big Data, IBM report (2012), pp. 1–20.

12. Intel, Big Data Analytics. Intel's IT Manager Survey on How Organizations Are Using Big Data, Intel report (2012).

13. Fernández, Alberto, et al. "Big Data with Cloud Computing: an insight on the computing environment, MapReduce, and programming frameworks." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 4.5 (2014): 380-409.

14. Kaushik, Chetan, and Atul Mishra. "A scalable, lexicon based technique for sentiment analysis." arXiv preprint arXiv:1410.2265 (2014).

15. http://www.tutorialspoint.com/tableau/