International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 6 Issue: 2
ISSN: 2454-4248
01 – 08

Sentiment Analysis on Social media network

B. Edukondalu[1], P. Neelima[2],
PG scholar[1], Associate Professor[2],
Department of Computer science and Engineering[1, 2]
SRKR Engineering College, Bhimavaram[1,2].
edukondalu510@gmail.com[1] neelima.p47@gmail.com

***Abstract*--**Detecting changes in a data stream is very important area of research with several applications. In this project, we use a method for the detection and estimation of change. This strategy mainly dealing with distribution change when learning from data sequences which may vary with time. We use sliding window whose size, rather than being fixed a priori, is recomputed according to the rate of change determined from the data in the window itself. This delivers the user to guess a time-scale for change within the data stream.

In this project we tend to use jio tweets in twitter as data stream. Reliance Jio network offers cost free services; the 100% satisfaction of its customer could be a doubtful one. Though the customers are availing Jio services, they spend some amount for using other networks. If Reliance Jio fails to give the full satisfaction to its customer, it is tough to sustain its image in the systematic nation. Hence the study is undertaken for the aim of analyzing the satisfaction level of the customer of Jio network.

From Twitter, we gather tweets using Twitter API based on keywords #jio. This project can verify the sentiment orientation of the tweets and also detect the changes in tweeted words in terms of frequencies by applying ADWIN sliding window algorithm. Further we can visualize these results by plotting graphs and can understand how many people are positive and negative towards jio.

*****

I. Introduction

Dealing with data whose nature changes over time is one of the core problems in data mining and machine learning. To mine or learn such data, one needs strategies for the following three tasks, at least: 1) detecting when change occurs 2) deciding which examples to keep and which ones to forget (or, more in general, keeping updated sufficient statistics), and 3) revising the current model(s) when significant change has been detected handle English language. We use ADWIN sliding window algorithm to detect changes in the streaming data. . The algorithm automatically grows the window when no change is apparent, and shrinks it when data changes. Finally, it signals alarm when change is detected.

In this paper we make a first step towards formally introducing and quantifying change in a data stream. We view the data as being generated by some underlying probability distribution, one data point at a time, in an independent fashion. Our goal is to detect when this data-generating distribution changes, and to quantify and describe this change.

Twitter is a "micro-blogging" social networking website that has a large and rapidly growing user base. Those who use Twitter can write short 140 characters long or less updates called „tweets". „Tweets" are seen by those who „follow" the person who „tweeted". Due to the growing popularity of the website, Twitter can provide a rich bank of data in the form of harvested "tweets". Twitter by its very nature, allows people to convey their opinions and thoughts openly about whatever topic, discussion point or product that they are interested in sharing their opinions about. Therefore, Twitter is a good medium to search for potentially interesting trends regarding prominent topics in the news or popular culture.

R studio is one of the varied programs that provide packages which can analyze data, however, R studio works well with math problems and incorporates a simple interface.

Sentiment analysis (or opinion mining) refers to natural language, text analysis and linguistics to identify and extract subjective knowledge at intervals the provision material.

The value of Twitter in recent years has risen as businesses, political groups and curious internet users alike have begun to assess the public"s general sentiment for his or her merchandise and services from Twitter posts. Sentiment analysis provides a way of pursuit opinions and attitudes on the net and determines if they are completely or negatively received by the final public. The aim of Text mining is to technique unstructured (textual) knowledge and to extract pregnant numeric indices from the text, allowing the applying of varied processing algorithms to elucidate the matter data set. The classification model that this project will develop will verify whether or not or not the tweet standing updates (which cannot exceed hundred and forty characters) reflects the positive opinion or negative opinion on the behalf of the one that tweeted.

 ISSN: 2454-4248
01 – 08

This project will use a hybrid of information based totally sentiment analysis methodologies that area unit plenty of traditionally used, and other people of machine learning methodologies that used a plenty of intuitive approach to sentiment.

II. Related Work

The concept drift detection problem has a classic statistical interpretation: given a sample of data, does this sample represent a single homogeneous distribution or is there some point in the data (i.e the concept change point) at which the data distribution has undergone a significant shift from a statistical point of view? All concept change detection approaches in the literature formulate the problem from this viewpoint but the models and the algorithms used to solve this problem differ greatly in their detail. Sebastiao and Gama [7] present a concise survey on change detection methods. They point out that methods used fall into four basic categories: Statistical Process Control (SPC), Adaptive Windowing, Fixed Cumulative Windowing Schemes and fi- nally other classic statistical change detection methods. Early Drift Detection Method (EDDM) [8] works on the same basic principle as the authors earlier work but uses different statistics to detect change. More recently Bifet et al [6] proposed an adaptive windowing scheme called ADWIN that is based on the use of the Hoeffding bound to detect concept change. The ADWIN algorithm was shown to outperform the SPC approach and has the attractive property of providing rigorous guarantees on false positive and false negative rates. ADWIN maintains a window (W) of instances at a given time and compares the mean difference of any two sub windows (W0 of older instances and W1 of recent instances) from W. If the mean difference is statistically significant, then ADWIN removes all instances of W0 considered to represent the old concept and only carries W1 forward to the next test.

Twitter is different to other forms of raw data which are used for sentiment analysis as sentiments are conveyed in one or two sentence blurbs rather than paragraphs. Twitter is much more informal and less consistent in terms of language. Users cover a wide array of topics which interest them and use many symbols such as emoticons to express their views on many aspects of their life (Agarwal et al. 2011). When using human generated status updates, sentiment is not always obvious; many tweets are ambiguous and can use humour to maximize the opinion to other human readers but deflect the opinion to a machine learning algorithm. (Agarwal et al. 2011).

Another consideration when using a dataset generated from Twitter is that a considerably large number of tweets which convey no sentiment such as linking to a news article, which can lead to difficulties in data gathering, training and testing. Parikh, Movassate (2009). Sentiment analysis provides a means of tracking opinions and attitudes on the web and determines if they are positively or negatively received by the public.

According to Mejova (2009) Sentiment analysis is usually conducted between two levels; a coarse level and a fine level. Coarse level sentiment analysis deals with determining the sentiment of an entire document and Fine level deals with attribute level sentiment analysis. Neethu, Rajasree (2013) Sentence level sentiment analysis comes in between these two. Sentiment analysis in Twitter provides a dramatically different data set where multiple interesting challenges can arise. According to Boiyetal (2007), Symbolic techniques and Machine Learning techniques are the two basic methodologies used in sentiment analysis from text.

Symbolic techniques in supervised classification models make use of available lexical resources. In his sentiment analysis Turney (2002) used bag-of-words approach. In this approach, the document was treated as a collection of words where relationships between words are not considered important. To determine the overall sentiment, sentiments of every word are given a value and using aggregation functions, those values are combined. Where tuples are, phrases having adjectives or adverbs which may be considered positive or negative, Turney (2002) found the polarity of a review was based on the average semantic orientation of tuples extracted from the review.

WordNet which is a database consists of words and their relative synonyms were used by Kamps etal. (2004). In this study, a distance metric was developed on WordNet and the semantic orientation of adjectives was determined from this metric.

In their study, Balahur et al. (2012) introduced a conceptual representation of text, which stored the structure and the semantics of real events, in a system called EmotiNet, Emotinet was able to identify the emotional responses triggered by actions with the information it stored.The difficulty with using a Knowledge base approach however that is it requires of a large lexical database. This has become harder and harder to provide as the language of social networks is so trend dependent and changeable that lexicon datasets cannot keep up. Therefore, Knowledge based approaches to sentiment analysis are not as popular as they used to be.

 ISSN: 2454-4248
01 – 08

III. Implementation Methodology

Design and Architecture

The system design that's projected for this project is shown in Figure. This project proposes a hybrid approach involving each information primarily {based} methodologies and machine learning based methodologies to analysis the sentiment orientation of the tweets. Tweets are accessed through a Twitter API. The collected reviews are projected in R associate degree saved to a surpass file. The words were extracted and keep in an exceedingly feature vector. The words were scored into their relative sentiment orientation employing a sentiment lexicon that was sourced from the net and was in public offered. The surpass file was then entered into R and a corpus was created. The corpus was then divided into a „training set" and a „test set" and have were extracted from every severally. From the „training set- feature extraction" a machine learning algorithmic program was created from the „test set-feature extraction" a model classifier within the sort of Naive Bayes classifier was created. Naive Bayes doesn't take into account the relationships between options like emotional keywords and emoticons. Usually this can be ideal for sentiment analysis as often these options don't invariably relate to 1 another like within the use of an emoticon facial gesture at the tip of a negative tweet. Naive Bayes Classifier[3] analyses every of the options of the feature vector one by one because it assumes that they're equally freelance of every alternative.The conditional probability for Naive Bayes can be defined as

$P(X|yj) = \Pi m\ P(xi|yj)$

"X" is the feature vector defined as X={x1 ,x2 ,....xm } and yj is the class label. In the tweets collected for this project there are different independent features such as emoticons, emotional keyword, which are treated as either positive or negative and so are utilized by Naive Bayes classifier for classification

Finally, we apply ADWIN algorithm for positive tweets to detect changes in the distribution. In ADWIN, a window is maintained that keeps the most recently read examples, and from which older examples are dropped according to some set of rules. For this three tasks, the content of the window can be used. The algorithm is shown below:

Adaptive Windowing Algorithm(ADWIN)
1 Initialize Window W
2 for eacht>0
3 do W←W∪{xt} (i.e., add xt to the head of W)
4 repeat Drop elements from the tail of W
5 until|ˆμW0−ˆμW1|≥εcut holds
6 for every split of W into W=W0·W1
7 output ˆMw

ADWIN keeps a variable-length window of recently seen items, with the property that the window has the maximal length statistically consistent with the hypothesis "there has been no change in the average value inside the window". More precisely, an older fragment of the window is dropped if and only if there is enough evidence that its average value differs from that of the rest of the window.

This has two consequences: one, that change is reliably detected whenever the window shrinks; and two, that at any time the average over the existing window can be reliably taken as an estimation of the current average in the stream.

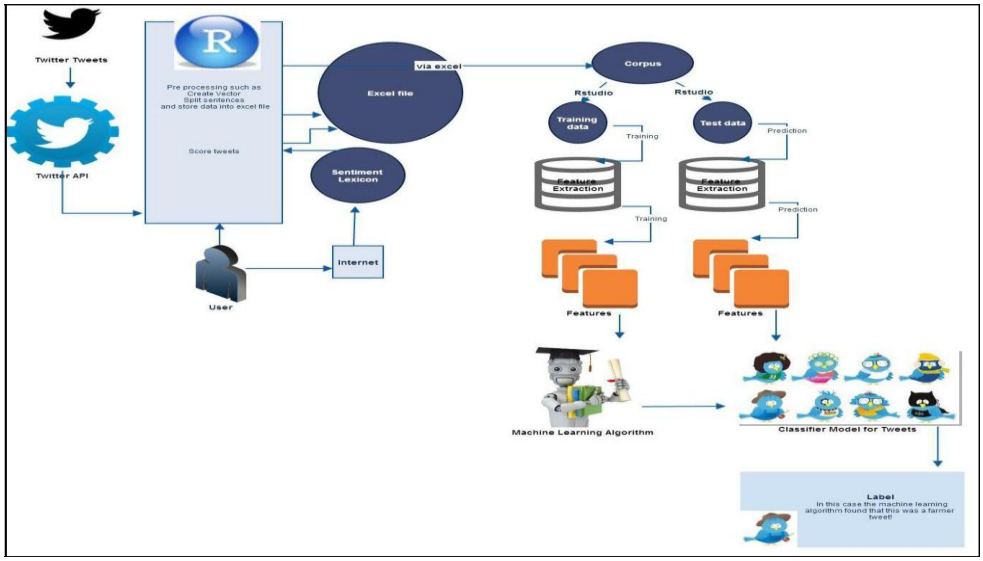


Figure: The machine learning algorithm is then applied to the model classifier and a label is produced

 ISSN: 2454-4248
01 – 08

IV. Objectives and Contribution to the Knowledge

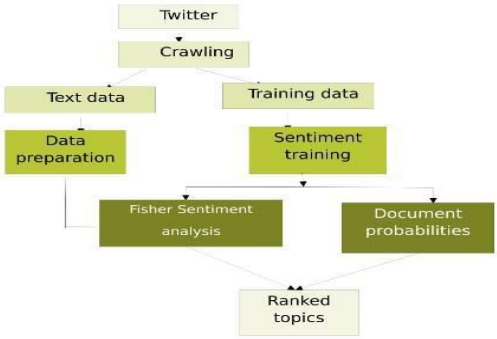
This project will use a hybrid of knowledge based sentiment analysis methodologies which have been more traditionally used, and those of machine learning methodologies which used a more intuitive approach to sentiment. The results of these two methodologies will be used to perform a thorough analysis of the dataset.

In order to conduct any kind of analysis on twitter the construction of a suitable dataset of tweets needs to be built. Twitter API is an app which extracts tweets from twitter and loads them into a dataset**.**

The aim of this project is to use the results from the knowledge based techniques and those of the machine learning techniques to ensure a thorough analysis of the dataset.

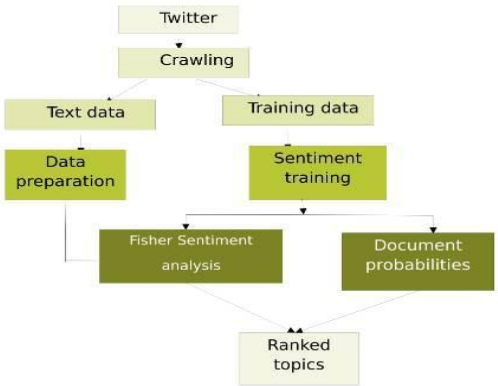
Systems/Datasets

The process of sentiment analysis for this project is outlined in the diagram below figure.



Overview

The process of sentiment analysis for this project is outlined in the diagram below:



The figure 3.2 shows the process will begin with an acquisition program applied to Twitter. Specified keywords will be taken from the data that is retrieved from Twitter. In order to support a derivation to a conceptual level, the data segments are then fragmented, assuming that every message will only contain a single concept. Three distinct categories have been chosen as follows

Tweets will be evenly split into three sets of text types;

Positive: those opinions who favour to jio has said or who react positively to his proposal comments (note these texts may be negative towards those who oppose her).

Negative: tweets that are not in favour to jio has said and who react negatively towards jio (these tweets may be positive towards other groups)

Neutral: Objective tweets or those which do not state an opinion.

V. General Description

Twitter API

Twitter exposes its data via an Application Programming Interface, (API). The Twitter API has two different flavours: Restful and Streaming. The Streaming API works by making a request for a specific type of data; filtered by keyword, user, geographic area, or a random sample, and then keeping the connection open as long as there are no errors in the connection.

The Restful API is useful for getting things like lists of followers and those who follow a particular user, and is what most Twitter clients are built off of. However, one of the main drawbacks is that the Restful API is that only tweets from 7 days preceding can be searched, and queries are limited to approximately 10 per minute at the time of writing. For this project, I am going to focus on the Streaming API.

Product Functions

A vital aspect of this project is Document preparation. This allows for the different aspects we want to when representing our document. A full text string representation is not very useful, because it is hard to find similarities between two text strings.

Therefore the „Bag of Words" text string model is used which vitally ignores the ordering of words, but instead counts of the number of occurrences of the words in the document. Some information may be lost in this system however, the bag of words model is still commonly used, and performs very strongly. It is computationally simple and in many applications, much of the information required for learning is captured by this representation. According to. Bespalov, Bai, and. Shokoufandeh 2011, the Bag of Words mode is a natural predecessor to the bag of N-gram. This system counts for groups of consecutive words of size n. This is important as it can eliminate ambiguities that can occur

 ISSN: 2454-4248
01 – 08

in bag of words models, such as “gay rights" being significantly different to “The gay flower leaves a shadow which falls to the right”. This system allows us the advantage of increased string length and therefore a greater context.

User Characteristics

The intended user will be politicians, media broadcasters and general public who are interested in the sentiment of the Twitter population with respect to the opinions formed by @jio.

Users are not expected to have a very high level of technical expertise.

VI. General Constraints

Personal Data

If a User has not made information public, Twitter does not return that data. Any Personal information that is collected from Twitter will not be stored or used in any way.

Twitter Data

The application must comply with the Twitter Developer terms of service. This includes the Following:

Defining an application privacy policy (what we do with tweets, user data, etc.)

Not redistributing Tweets

Providing a link to Twitter sign-up if user does not have a registered Twitter account Specific Constraints

Specific Constraints

Specific Constraints that this product may encounter when dealing with the users Jose, Bhatia and Krishna (2010).

Negative sentences: many people would write their tweets with negation before the adjective or verb, which complicates the data. For example: a sentence such as Not satisfied with the Situation of Gay Marriage. Has the adjective satisfied, which can assign a polarity positive without considering the negation in the sentence.

Confusing polarity: for certain tweets, there will be a confusion or disagreement for the polarity to be assigned. For instance, jio defeats other networks is positive when taken from others point of view its negative query

Dealing with emoticons: Our data should contain clean labels and emoticons are deemed. A noisy label. However, emoticons are popular on Twitter therefore the data will have to clean these out.

Casual language: Tweets contain very casual language. For example, a user may want to right the word happy as: happppppyyy happpiieee happy hap-e besides showing that people are happy, this emphasizes the c asual nature of Twitter. Usage of links:

Users very often include links in their tweets. Thus, there is a need to classify. This type of tweet by using keywords such as URL. But even then, it is difficult to extract the Sentiment sometimes as sometimes it may not be given or it is unclear.

VII. Sentimental Analysis Algorithm

1. Creating a Twitter Application using twitter developer API.(application program interface).
2. Working on RStudio- Building the corpus.
3. Search the keywords and then Save the Tweets.
4. Clean the Tweets and save them.
5. Apply Sentiment function to tweets.
6. Then Import the csv file.
7. Apply polarity function to Visualizing the tweets.
8. Perform Text Analysis.
9. plot Word Could, locationwise plot emotion plot and polarity graph.

VIII. Flowchart for Sentimental Analysis Algorithm

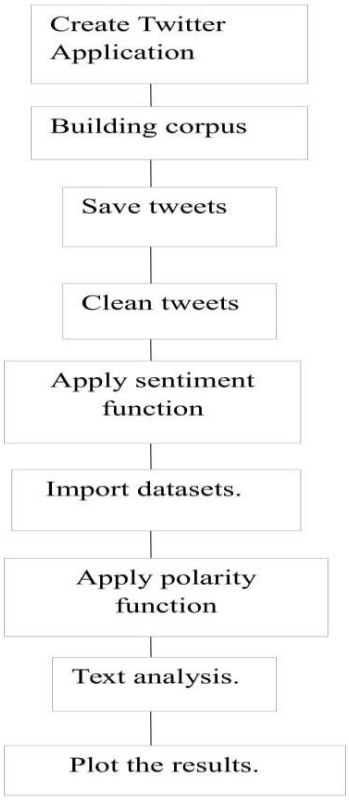
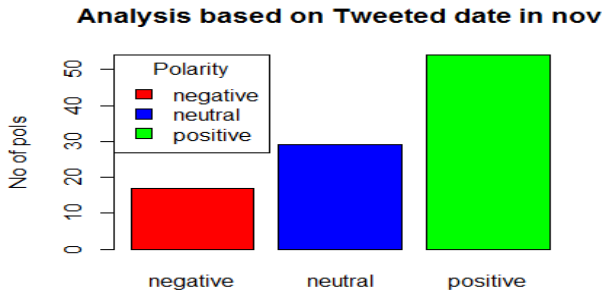


Figure: the figure shows the flow of the sentimental analysis algorithm.

 ISSN: 2454-4248
01 – 08

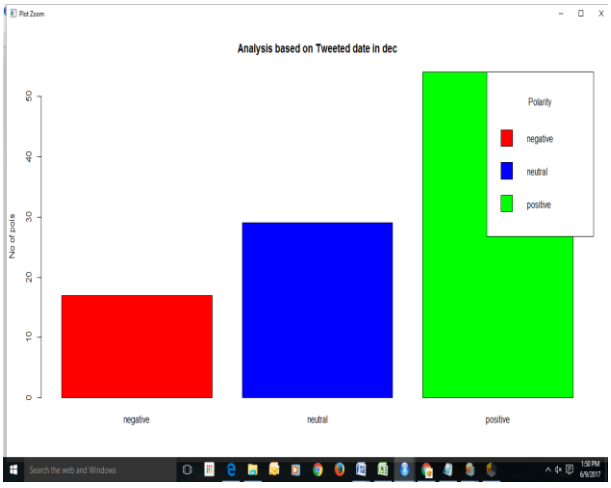
IX. Results Analysis

The aim of this project was to analysis the results of a sentiment orientation on the keyword #jio. The plot below shows the polarity related to the jio tweets. The x-axis shows the score of each tweet as a negative and positive integer or zero. A positive score represents positive or good sentiments associated with that particular tweet whereas a negative score represents negative or bad sentiments associated with that tweet. A score of zero indicates a neutral sentiment. The more positive the score, the more positive the sentiments of the person tweeting and vice-versa.



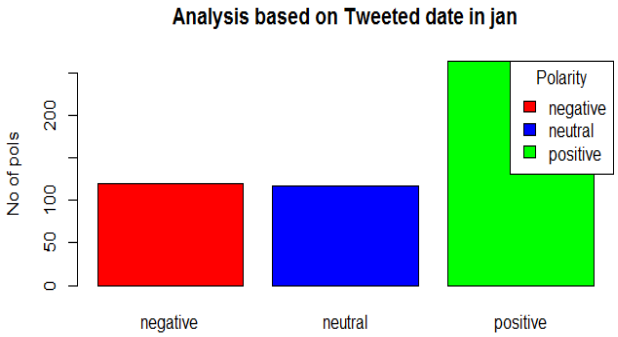
November month

The above histogram is slightly skewed towards negative score which shows that the sentiments of people regarding jio are overwhelming negative with a slight skew towards positive.



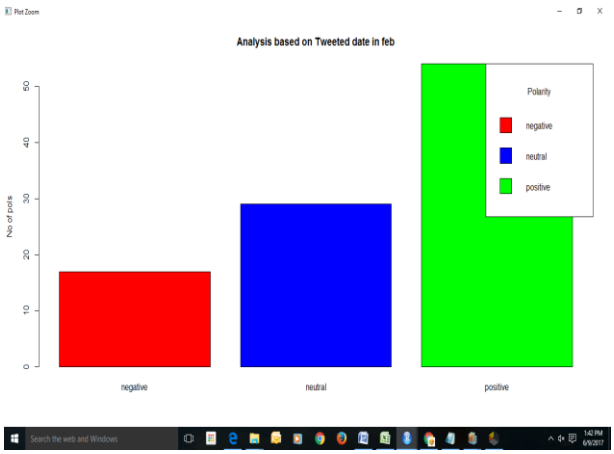
December month

The above histogram is slightly skewed towards positive score which shows that the sentiments of people regarding jio are overwhelming positive with a slight skew towards negative.



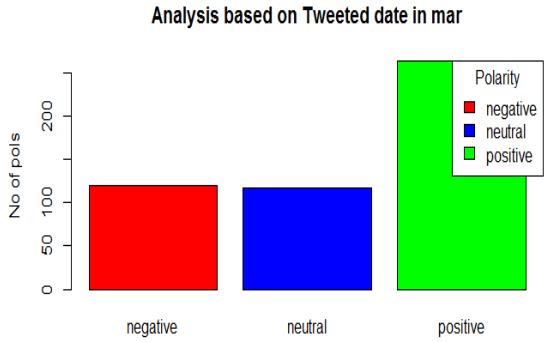
January month:

The above histogram is slightly skewed towards negative score which shows that the sentiments of people regarding jio are overwhelming negative with a slight skew towards positive.



February month:

The above histogram is slightly skewed towards positive score which shows that the sentiments of people regarding jio are overwhelming positive with a slight skew towards negative.



March month

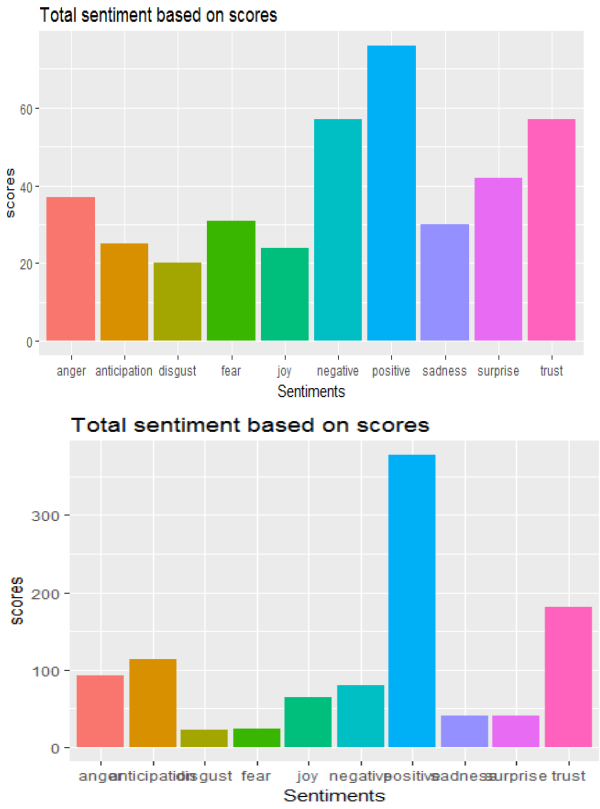
The above histogram is slightly skewed towards positive score which shows that the sentiments of

ISSN: 2454-4248
01 – 08

people regarding #jio overwhelming positive with a slight skew towards positive.

Emotions

The below plots (figures 9.6) shows the frequency of tweets with respect of scores allotted to each tweet in the months of November of 2016, January, February and march of 2017.

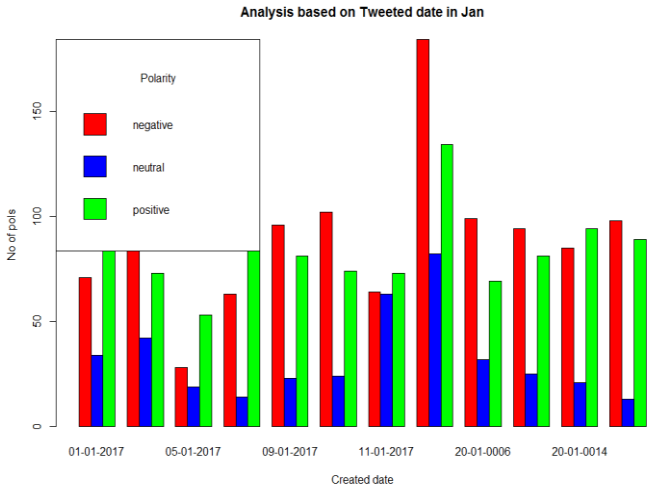


The above figures plot shows the frequency of tweets with respect of scores allotted to each tweet . The x-axis shows the score of each tweet as anger, joy, surprise, disgust, fear and sadness. Based on the tweeted text the "classify_emotion" function classified emotion.

The above plot is slightly skewed towards "joy, which shows that the sentiments of people regarding jio are overwhelming "joy". the positive response of jio is increased in august due to introducing new phone they are very positive towards jio.

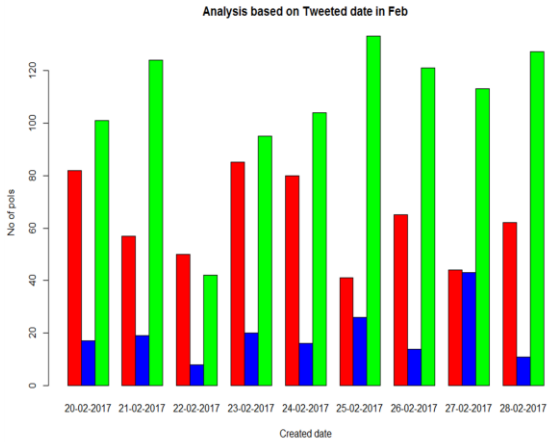
Date wise plots

The plot below shows the date wise emotions related to the jio tweets in different months from November 2016 to March 2017.



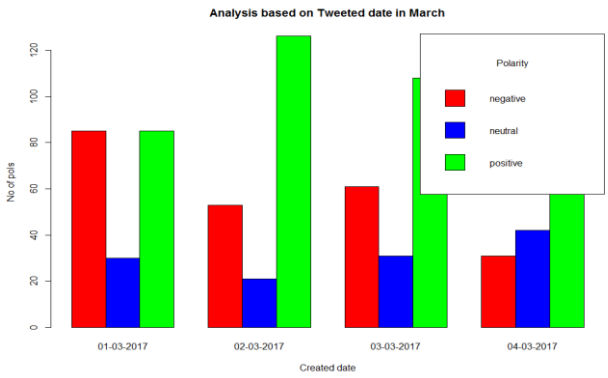
The plot below shows the date wise emotions related to the jio tweets in January month.

February month:



March month:

The plot below shows the date wise emotions related to the jio tweets in February month. In February month 21, 25 and 28 date has high positive tweets related to jio tweets.



The plot below shows the date wise emotions related to the #jio tweets in march month.

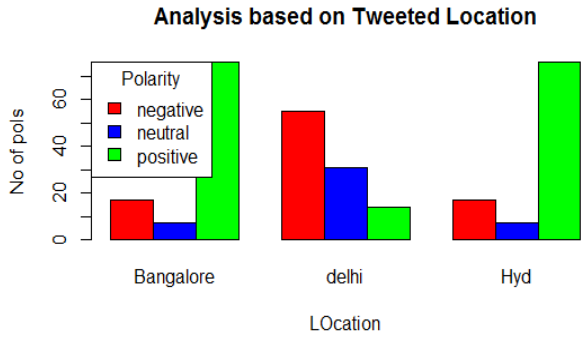
The above plots show the twitter analysis based on different months. Here, polarity is placed according to different months. And graphs were plotted by positive,

 ISSN: 2454-4248
01 – 08

negative and neutral aspects accordingly. We can clearly see the analysis performed on tweets of different months in their respective graphs.

Location wise

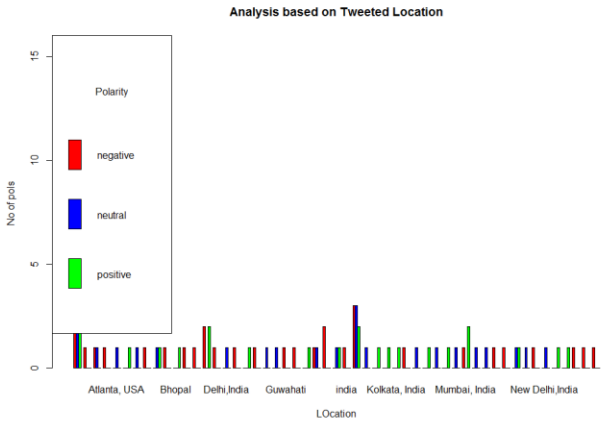
The plot shows the location wise analysis on jio tweets. The plot shows the twitter analysis based on different locations. We took three locations (Bangalore, Delhi, Hyderabad) as a sample in the plot. For this we used geocode attribute in the "search Twitter" function. Based on the geocode value we can retrieve those location tweets.



We can observe that the tweets in Bangalore and Hyderabad were positively scored and Delhi were negatively scored.

Location wise

The below plot shows the world-wide tweets analysis. For this we used birdIQ site to retrieve tweets.



X. Conclusions

Reliance Jio launched in India in September, the company's offering has been restricted to Smartphone users as it is a 4G only network, and to make calls ,a phone has to support the VoLTE technology, which is usually not found on basic phones so they are slightly positive towards jio. Now a phone has to support the VoLTE technology, which is usually not found on basic phones functionality. The phone comes with a number of apps and browser installed aside from Jio"s offerings too with affordable cost. So customers are very positive towards jio network.

References

[1]. Apoorv Agarwal, Fadi Biadsy, and Kathleen Mckeown. 2009. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams. Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), pages 24–32, March.

[2]. Luciano Barbosa and Junlan Feng. 2010. Robust sentiment detection on twitter from biased and noisy data. Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 36–44.

[3]. Adam Bermingham and Alan Smeaton. 2010. Classifying sentiment in microblogs: is brevity an advantage is brevity an advantage? ACM, pages 1833–1836.

[4]. C. Fellbaum. 1998. Wordnet, an electronic lexical database. MIT Press.

[5]. Michael Gamon. 2004. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. Proceedings of the 20th international conference on Computational Linguistics.

[6]. Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. Technical report, Stanford.

[7]. David Haussler. 1999. Convolution kernels on discrete structures. Technical report, University of California at Santa Cruz.

[8]. M Hu and B Liu. 2004. Mining and summarizing customer reviews. KDD.

[9]. S M Kim and E Hovy. 2004. Determining the sentiment of opinions. Coling.

[10]. Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. Proceedings of the 41st Meeting of the Association for Computational Linguistics, pages 423–430.

[11]. Alessandro Moschitti. 2006. Efficient convolution kernels for dependency and constituent syntactic trees. In Proceedings of the 17th European Conference on Machine Learning.

[12]. Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of LREC.

[13]. B. Pang and L. Lee. 2004. A sentimental education: Sentiment analysis using subjectivity analysis using subjectivity summarization based on minimum cuts. ACL.

[14]. P. Turney. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. ACL.

[15]. C M Whissel. 1989. The dictionary of Affect in Language. Emotion: theory research and experience, Acad press London.

[16]. T.Wilson, J.Wiebe, and P. Hoffman. 2005. Recognizing contextual polarity in phrase level sentiment analysis. ACL.