Swathieswari Mohanraj et al, International Journal of Computer Science and Mobile Applications,
Vol.5 Issue. 9, September- 2017, pg. 39-47 ISSN: 2321-8363
Impact Factor: 4.123

REVIEW PREDICTOR ON SOCIAL MEDIA FOR SPECIFIC TOPIC TO FIND ITS IMPACT ON PEOPLE: SENTIMENT ANALYSIS APPROACH

Mrs. Swathieswari Mohanraj, Dr. K.Thyagarajan
Research Scholar, Department of Computer Science, swathieswari.mca@gmail.com
Associate Professor, Department of Computer Science, k_thyagarajan@gmail.com
A.V.C College (Autonomous), Mannampandal, Mayiladuthurai

Abstract

In this paper, we predict the reviews of people on social media like twitter and facebook to know the topic's impact. For eg: how many positive and negative reviews found on jallikattu as state wise. Now a days present industries and some survey companies are mainly taking decisions by data obtained from web. As we see WWW is a rich collection of data that is mainly in the form of unstructured data from which we can do analysis on those data which is collected on some situation or on a particular thing.

Simply this review predictor collects all the tweets and facebook posts based on the keyword which is in unstructured format. We need to make it structured by using hadoop, php Search api for twitter and Graph api for facebook posts

Keywords— PHP, jQuery, Hadoop, TwitterAPI, GraphApi, AFFIN dictionary, Flume, Pig, Sentiment analysis, Big data

1. Introduction

Humans are subjective creatures and opinions are important. Being able to interact with people on that level has many advantages for information systems. Comparatively few categories (positive/negative, 3 stars, etc) compared to text categorization Crosses domains, topics, and users Categories not independent (opposing or regression-like) Characteristics of answers to opinion- based questions are different from fact- based questions, so opinion-based information extraction differs from trad information extraction.

Some of the challenges in Sentiment Analysis are: People express opinions in complex ways, in opinion texts, lexical content alone can be misleading. Another challenge can be in the form of Intra-textual and sub-sentential reversals, negation, topic change common. Humans tend to express a lot of remarks in the form of sarcasm, irony, implication, etc. which is very difficult to interpret. For Example- "How can someone sit through the movie" is extremely negative sentiment yet contains no negative lexographic word. Even if a opinion word is present in the text, their can be cases where a opinion word that is considered to be positive in one situation may be considered negative in another situation. People can be contradictory in their statements. Most reviews will have both positive and negative comments, which is somewhat manageable by analyzing sentences one at a time. However, in the more informal medium like twitter or blogs(Social media), the more likely people are to combine different opinions in the same sentence which is easy for a human to understand, but more difficult for a computer to parse. Sometimes even other people have difficulty understanding what someone thought based on a short piece of text because it lacks context. A good example would be “The laptop is good but I would prefer, the operating system which I was using”.

There is a huge demand of sentiment analysis. Before buying any product its a practice now, to review its rating as rated by other persons who are using it. Online advice and recommendations the data reveals is not the only reason behind the buzz in this area. There are other reasons like the company wants to know “How Successful was their last campaign or product launch” based upon the sentiments of the customers on social media. .Sentiment analysis concentrates on attitudes, whereas traditional text mining focuses on the analysis of facts. There are few main fields of research predominate in Sentiment analysis: sentiment classification, feature based Sentiment classification and opinion summarization.

Apache's Hadoop framework has become synonymous with the big data movement and is it designed to become the dominant data management platform for us all. Present situation is people completely are expressing their thoughts through online blogs, discussion forms and also some online applications like Facebook, Twitter, etc.

If we take Twitter as our example nearly 1TB of text data is generating within a week in the form of tweets. So, by this it is understand clearly how this Internet is changing the way of living and style of people. Among these tweets can be categorized by the hash value tags for which they are commenting and posting their tweets. So, now many companies and also the survey companies are using this for doing some analytics such that we can predict the success rate of their product or also we can show the different view from the data that we have collected for analysis.

There are different ways to get Twitter data or any other online streaming data where they want to code lines of coding and that streaming process could not be a dynamic one to achieve this. That will not be a real-time. Coming to this paper we have achieved by this problem statement and solving u BIGDATA by using Php, Hadoop and its Eco Systems. And finally we have done sentiment analysis on the Twitter and facebook data that is stored in HDFS. So, here the data taken in real-time also we are getting data dynamically based on user needs.

2. Literature Review

Numerous study has been done to determine and classify sentiment of tweets in twitter. Both supervised and unsupervised techniques are used. Supervised such as Naïve Bayes Algorithm[1].Some other papers shown an AAVN based sentiment analysis technique deploying linguistic analysis of adverbs, adjective, abstract npoun and categorized verb, the paper defines a setoff general axioms for opinion analysis to determine a functional value of the sentiment analysis[2]. In this paper Based on the supervised techniques such as Naïve bayes to detect its polarity[3]. In particular, analysis on online reviews has become a hot research field. Survey on latest development in analysis, and makes an in-depth. Introduction on its research and application in business [4].

Our day to day life has always been influence by what people think. Ideas and opinion of others have always affected our own opinion. Effective analysis is the computational treatment of opinions, sentiments and subjectivity of text [5]. In this review we take look at the various challenges and applications of analysis.

There are different ways to get Twitter data or any other online streaming data where they want to code lines of coding to achieve this. Here the processing time taken is also very less compared to the previous methods because Hadoop Map Reduce and Hive are the best methods to process large amount of data in a small time [6].

One fundamental problem in sentiment analysis is categorization of sentiment polarity [10,11-12]. Given a piece of written text, the problem is to categorize the text into one specific sentiment polarity, positive or negative (or neutral). Based on the scope of the text, there are three levels of sentiment polarity categorization, namely the document level, the sentence level, and the entity and aspect level [13]. The document level concerns whether a document, as a whole, expresses negative or positive sentiment, while the sentence level deals with each sentence's sentiment categorization; The entity and aspect level then targets on what exactly people like or dislike from their opinions.

Since reviews of much work on sentiment analysis have already been included in [13], in this section, we will only review some previous work, upon which our research is based on only the topic.

Our day to day life has always been influence by what people think. Ideas and opinion of others have always affected our own opinion. Effective analysis is the computational treatment of opinions, sentiments and subjectivity of text [5]. In this review we take look at the various challenges and applications of analysis. There are different ways to get Twitter data or any other online streaming data where they want to code lines of coding to achieve this. Here the processing time taken is also very less compared to the previous methods because Hadoop Map Reduce and Hive are the best methods to process large amount of data in a small time [6].

3. Design Aspects

As we have already discussed about some of the ways of getting data and also performing the sentiment analysis on those data. Here they are going to use some coding techniques for crawling the data from the twitter where they can extract the data from the Twitter web pages by using some code that may be written either in JAVA, Python or PHP etc. For those they are going to download the libraries that are provided by the twitter guys by using this they are crawling the data that we want particularly [7]. After getting raw data they will filter by using some old techniques and also they will find out the positive, negative and moderate words from the list of collected words in a text file. All these words should be collected by us to filter out or do some sentiment analysis on the filtered data [2] [8]. These words can be called as a dictionary set by which they will perform sentiment analysis. Also, after performing all these things and they want to store these in a database and coming to here they can use RDBMS [9] where they are having limitations in creating tables and also accessing the tables effectively.

3.1 Existing System

In Existing System we have size issues i.e In RDBMS, we have limitation for the records in the table. Then the execution time to get the results. Apart from this basic features, polarity detection is not done so that we here find the polarity of each and every text. Here only twitter social media is handled. To increase more accuracy in the reviews we used one more social media such as facebook here.

Here only one api twitterAPI is used[8]. But in our paper three api's had been used to collect the appropriate user data based on the keyword search.

Based on the existing system issues, the new system is proposed with certain modification to increase the efficiency and accuracy of the system.

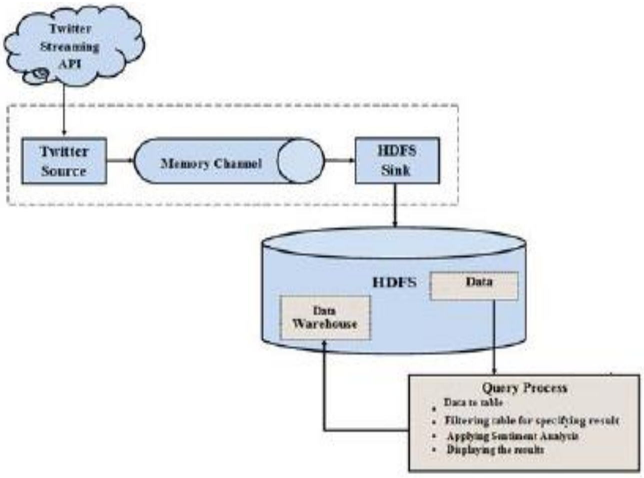


Fig A: Architecture of Existing System

3.2 Proposed System

We want to overcome limitations in creating, accessing and maintaining tables, here we use Big Data problem statement also we use Hadoop and its Ecosystems, for getting raw data from the Twitter and facebook using php search api and graph api as a json file then we store in the hdfs flume using Hadoop [2]. Then we process the json and separate the words by using pigjsonloader, after that we use the AFFIN dictionary to find its weight. For eg: “I like my home”. In this like is a positive word so here we give some +values to add its sentence weight. Likewise if we found any negative words we will give some negative values to the sentence. In this way weightage calculated for each tweets and facebook posts.

3.3 Methodology

STEP 1 : Invoking PHP code to get the twitter tweets and facebook posts . Here we use Twitter search API and TwitterAPI, Facebook Graph API to get the tweets and facebook posts as a json file.

Why search API

We want to select specific tweets (ie trending tweets such as big boss,jallikattu,methane scheme) by using keyword and on specific time and date. Eg: Since Jan 2016 to Feb 2016, how many tweets came for the specific keyword.

But in Facebook, we could not download specific facebook post. PHP code will produce all our facebook page posts as a json. We again process the json using Jquery to get specific facebook post.

STEP 2: Storing these json in a HDFS sink via flume

What is Json and Why we have to use it here

Json is JavaScript object Notation is a lightweight format that is used for data interchanging. In php, json file can be easily parsed for further process

HDFS Sink:

HDFS is a java-based file system that provides scalable and reliable data storage, and it was designed to span large clusters of commodity servers. This sink removes the event from the channel and puts into an external repository like HDFS via flume`

What is Flume
Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of streaming data into the Hadoop Distributed File System (HDFS).

STEP 3: Here we going to use hive, to extract the id and text
Why Hive
It is a platform used to develop SQL type scripts It is same to a SQL but it is for big data operations ie for bigger size datas using HQL.

STEP 4: Process using AFFIN dictionary, then find the weightage to each and every tweets and posts. Weightage assigned will from 1 to 5.

What is AFFIN dictionary
AFINN is a list of English words rated for valence with an integer between minus five (negative) and plus five (positive).

4. Figures and Explanations

4.1 Figures

In Figure 1, structure of the methodology is explained then in Figure 2, the GUI window for php search is given. In Figure 3, the extracted json file is attached with the tweets and facebook posts. In Figure 4, we processed the each text of tweet and facebook post using Affin dictionary. In Fig 5,Result of the each text by each user based on their id. In Fig 6, we displayed the result as a reviews using Hive.

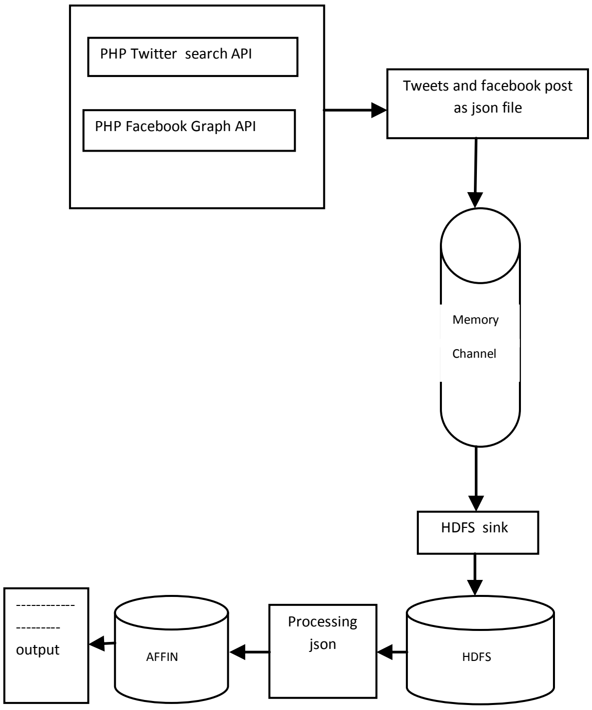
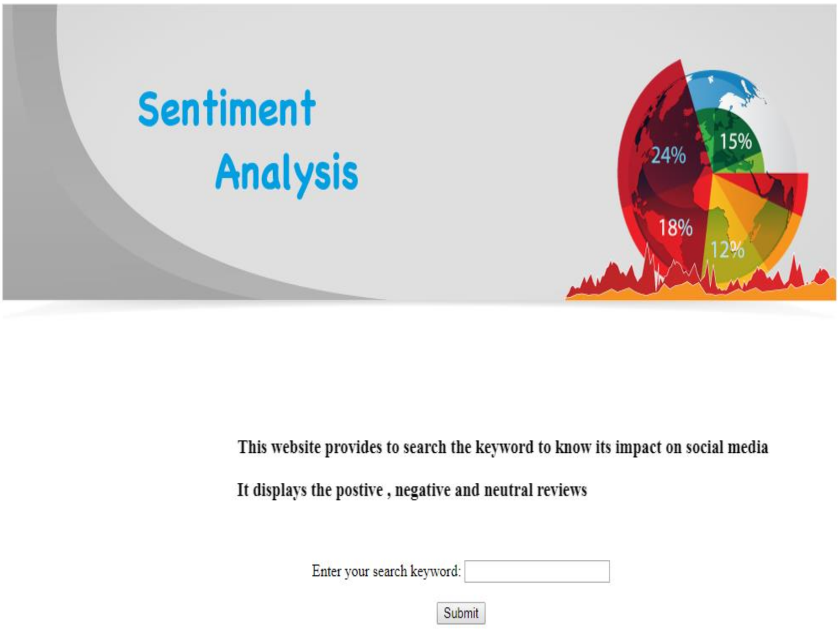


Fig. 1 : Structure of Review Predictor

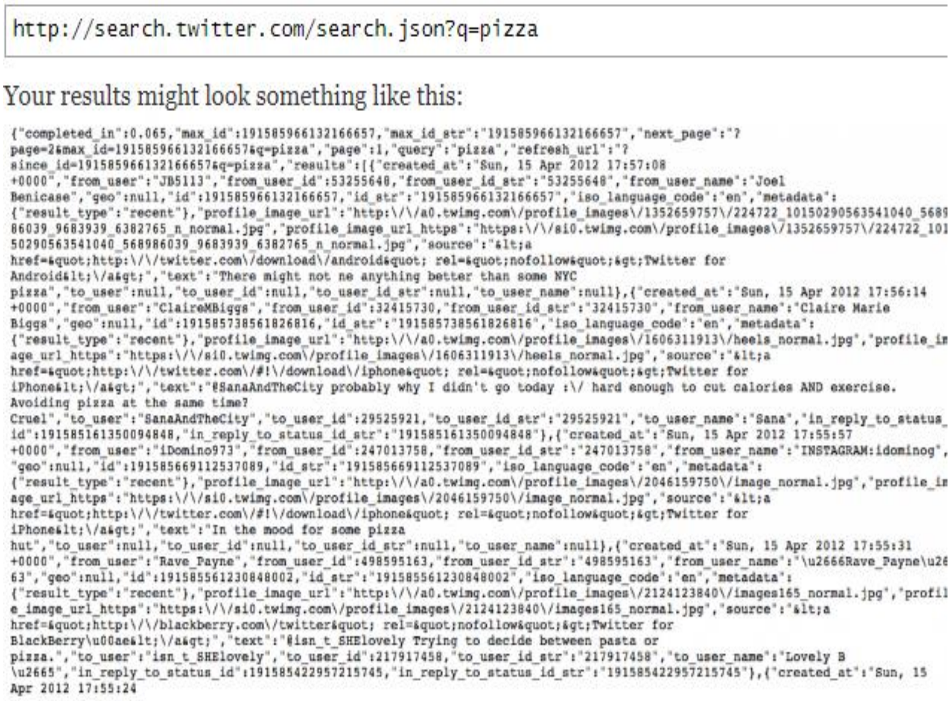


This website provides to search the keyword to know its impact on social media

It displays the postive , negative and neutral reviews

Enter your search keyword:

Fig 2: GUI Window for Sentiment analysis

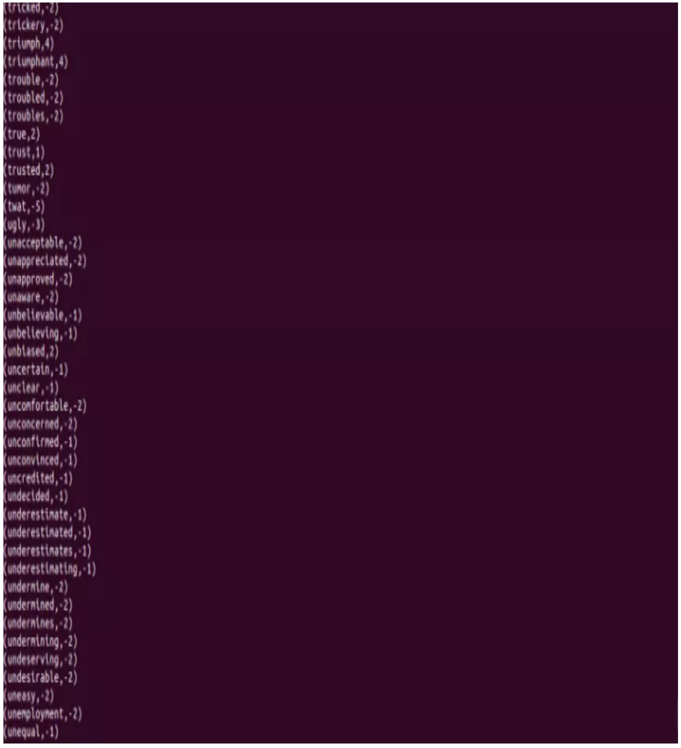


http://search.twitter.com/search.json?q=pizza

Your results might look something like this:

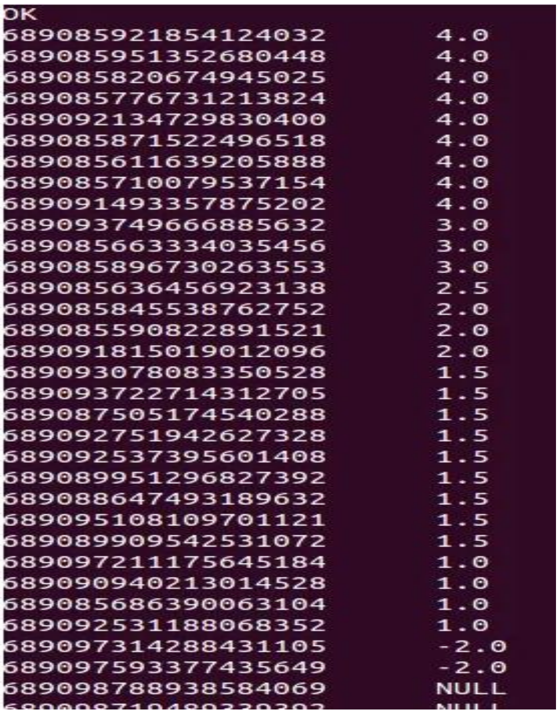
```
{"completed_in":0.065,"max_id":191585966132166657,"max_id_str":"191585966132166657","next_page":"?
page=2&max_id=191585966132166657&q=pizza","page":1,"query":"pizza","refresh_url":"?
since_id=191585966132166657&q=pizza","results":[{"created_at":"Sun, 15 Apr 2012 17:57:08
+0000","from_user":"JB5113","from_user_id":53255648,"from_user_id_str":"53255648","from_user_name":"Joel
Benicase","geo":null,"id":191585966132166657,"id_str":"191585966132166657","iso_language_code":"en","metadata":
{"result_type":"recent"},"profile_image_url":"http:\/\/a0.twimg.com\/profile_images\/1352659757\/224722_10150290563541040_5689
86039_9683939_6382765_n_normal.jpg","profile_image_url_https":"https:\/\/si0.twimg.com\/profile_images\/1352659757\/224722_101
50290563541040_568986039_9683939_6382765_n_normal.jpg","source":"&lt;a
href=&quot;http:\/\/twitter.com\/download\/android&quot; rel=&quot;nofollow&quot;&gt;Twitter for
Android&lt;\/a&gt;","text":"There might not ne anything better than some NYC
pizza","to_user":null,"to_user_id":null,"to_user_id_str":null,"to_user_name":null},{"created_at":"Sun, 15 Apr 2012 17:56:14
+0000","from_user":"ClaireMBiggs","from_user_id":32415730,"from_user_id_str":"32415730","from_user_name":"Claire Marie
Biggs","geo":null,"id":191585738561826816,"id_str":"191585738561826816","iso_language_code":"en","metadata":
{"result_type":"recent"},"profile_image_url":"http:\/\/a0.twimg.com\/profile_images\/1606311913\/heels_normal.jpg","profile_im
age_url_https":"https:\/\/si0.twimg.com\/profile_images\/1606311913\/heels_normal.jpg","source":"&lt;a
href=&quot;http:\/\/twitter.com\/#!\/download\/iphone&quot; rel=&quot;nofollow&quot;&gt;Twitter for
iPhone&lt;\/a&gt;","text":"@SanaAndTheCity probably why I didn't go today :\/ hard enough to cut calories AND exercise.
Avoiding pizza at the same time?
Cruel","to_user":"SanaAndTheCity","to_user_id":29525921,"to_user_id_str":"29525921","to_user_name":"Sana","in_reply_to_status_
id":191585161350094848,"in_reply_to_status_id_str":"191585161350094848"},{"created_at":"Sun, 15 Apr 2012 17:55:57
+0000","from_user":"iDomino973","from_user_id":247013758,"from_user_id_str":"247013758","from_user_name":"INSTAGRAM:idominog",
"geo":null,"id":191585669112537089,"id_str":"191585669112537089","iso_language_code":"en","metadata":
{"result_type":"recent"},"profile_image_url":"http:\/\/a0.twimg.com\/profile_images\/2046159750\/image_normal.jpg","profile_im
age_url_https":"https:\/\/si0.twimg.com\/profile_images\/2046159750\/image_normal.jpg","source":"&lt;a
href=&quot;http:\/\/twitter.com\/#!\/download\/iphone&quot; rel=&quot;nofollow&quot;&gt;Twitter for
iPhone&lt;\/a&gt;","text":"In the mood for some pizza
hut","to_user":null,"to_user_id":null,"to_user_id_str":null,"to_user_name":null},{"created_at":"Sun, 15 Apr 2012 17:55:31
+0000","from_user":"Rave_Payne","from_user_id":498595163,"from_user_id_str":"498595163","from_user_name":"\u2666Rave_Payne\u26
63","geo":null,"id":191585561230848002,"id_str":"191585561230848002","iso_language_code":"en","metadata":
{"result_type":"recent"},"profile_image_url":"http:\/\/a0.twimg.com\/profile_images\/2124123840\/images165_normal.jpg","profil
e_image_url_https":"https:\/\/si0.twimg.com\/profile_images\/2124123840\/images165_normal.jpg","source":"&lt;a
href=&quot;http:\/\/blackberry.com\/twitter&quot; rel=&quot;nofollow&quot;&gt;Twitter for
BlackBerry\u00ae&lt;\/a&gt;","text":"@isn_t_SHElovely Trying to decide between pasta or
pizza.","to_user":"isn_t_SHElovely","to_user_id":217917458,"to_user_id_str":"217917458","to_user_name":"Lovely B
\u2665","in_reply_to_status_id":191585422957215745,"in_reply_to_status_id_str":"191585422957215745"},{"created_at":"Sun, 15
Apr 2012 17:55:24
```

Fig 3: structure of json going to store in hdfs directory



```
(tricked,-2)
(trickery,-2)
(triumph,4)
(triumphant,4)
(trouble,-2)
(troubled,-2)
(troubles,-2)
(true,2)
(trust,1)
(trusted,2)
(tumor,-2)
(twat,-5)
(ugly,-3)
(unacceptable,-2)
(unappreciated,-2)
(unapproved,-2)
(unaware,-2)
(unbelievable,-1)
(unbelieving,-1)
(unbiased,2)
(uncertain,-1)
(unclear,-1)
(uncomfortable,-2)
(unconcerned,-2)
(unconfirmed,-1)
(unconvinced,-1)
(uncredited,-1)
(undecided,-1)
(underestimate,-1)
(underestimated,-1)
(underestimates,-1)
(underestimating,-1)
(undermine,-2)
(undermined,-2)
(undermines,-2)
(undermining,-2)
(undeserving,-2)
(undesirable,-2)
(uneasy,-2)
(unemployment,-2)
(unequal,-1)
```

Fig 4: Processed texts using affin



```
OK
689085921854124032        4.0
689085951352680448        4.0
689085820674945025        4.0
689085776731213824        4.0
689092134729830400        4.0
689085871522496518        4.0
689085611639205888        4.0
689085710079537154        4.0
689091493357875202        4.0
689093749666885632        3.0
689085663334035456        3.0
689085896730263553        3.0
689085636456923138        2.5
689085845538762752        2.0
689085590822891521        2.0
689091815019012096        2.0
689093078083350528        1.5
689093722714312705        1.5
689087505174540288        1.5
689092751942627328        1.5
689092537395601408        1.5
689089951296827392        1.5
689088647493189632        1.5
689095108109701121        1.5
689089909542531072        1.5
689097211175645184        1.0
689090940213014528        1.0
689085686390063104        1.0
689092531188068352        1.0
689097314288431105        -2.0
689097593377435649        -2.0
689098788938584069        NULL
```

Fig 5: Resultant display of weightage for each user id

Positive_reviews	20%
Negative_reviews	70%
Neutral_reviews	10%

Fig 6: Result

5. Future Scope

The future of sentiment analysis lies in resolving the challenges faced and forming an effective sentiment analysis tools. The tool can scale and learn, once it has relevant data across various platforms. One has to make sure that the tool adapts through the changing needs of the brands across various time domains and for the years to come. One can generate a tool that adds various other sentiments in addition to the existing ones. Add wishes, caveats, comparisons and preferences to the existing sentiments. The tool must be able to identify false messages that are used to portray the brand in a positive way. These messages are computer generated.

In future we can search by location wise and we can use multimodal sentimental analysis like processing text, processing images, processing videos and processing the smileys etc to get the further more accuracy in sentimental analysis.

6. Conclusion

There are different ways to get Twitter data or any other online streaming data and also they want to perform the analysis on the stored data which will be helpful for the application user. After creating an account on twitter developer, authentication keys (API key, API Secret key, Access token and Access token secret) are generated by twitter server. These keys are useful for authentication. The keyword to be searched for tweets is passed from application by using twitter API towards twitter server which will intern returns a set of recent tweets. Whatever data fetched from twitter server is raw data. Raw data has to be converted to original tweets by using Java script object notation.

Based on the new Api's list, we can add so many api to get the appropriate results and also we can include so many social media accounts to more prior reviews.

References

[1] Mohd Naim Mohd Ibrahim and Mohd Zaliman Mohd Yusoff, "*Twitter Sentiment Classification Using Naive Bayes Based on Trainer Perception", 2015 IEEE Conference on eLearning, eManagement and e-Services.

[2] Souvik Sarkar, Partho Mallick and Tapas Kr. Mitra, *"A Novel Machine Learning Approach for Sentiment Analysis Based on Adverb - Adjective - Noun - Verb (AANV) Combination"*, Int. J. on Recent Trends in Engineering and Technology,Vol 7, No.1,July 2012.

[3] Sentiment and Emotion Analysis for Context Sensitive Information Retrieval of Social Networking Sites: A Survey D.I. George Amalarethinam, V. Jude Nirmal International Journal of Computer Applications (0975 – 8887) Volume 100– No.10, August 2014.

[4] Eman M.G. Younis *"Sentiment Analysis and Text Mining for Social Media Microblogs using Open Source Tools: An Empirical Study".* International Journal of Computer Applications (0975 – 8887) Volume 112 – No. 5, February 2015 .

[5] Nann, Stefan, Krauss,Jonas, Schoder, Detlef " *Predictive analytics on public data – the case of stock markets"* Proceedings of the 21st European Conference on Information Systems.2013 .

[6] Chlo´e Clavel, Catherine Pelachaud, Magalie *Ochs "User's sentiment analysis in face-to-face human-agent interactions – prospects" 2013.*

[7] Jiehan Zhou, Changrong Yu, Jukka Riekki, Elise Kärkkäinen." *AmE Framework: a Model for Emotionaware Ambient Intelligence"* 2010.

[8]Suchita M. Patil, R. P. Mirajkar, Neeta B. Patil, Sagar B. Patil , *"Prioritization of Data Using Sentiment Analysis in Calamitous Situation*" International Research Journal of Engineering and Technology (IRJET) Volume: 04 Issue: 01 | Jan-2017

[9] U. Krcadinac, P. Pasquier, J. Jovanovic, and V. Devedzic, "Synesketch*: An open source library for sentence-based emotion recognition,*" IEEE Transactions on Affective Computing, vol. 4, pp. 312–325, 2013.

[10] Sagar Patil, Neelesh Tippe, Pravin Patil, "Ubiquitous Adoption of Telemedicine to extend patient care beyond the office," International Journal of Emerging Pang B, Lee L (2008) Opinion mining and sentiment analysis. Found Trends Inf Retr2(1-2): 1–135

[12] Chesley P, Vincent B, Xu L, Srihari RK (2006) Using verbs and adjectives to automatically classify blog sentiment. Training580(263): 233.

[13] Tan LK-W, Na J-C, Theng Y-L, Chang K (2011) Sentence-level sentiment polarity classification using a linguistic approach In: Digital Libraries: For Cultural Heritage, Knowledge Dissemination, and Future Creation, 77–87.. Springer, Heidelberg, Germany.

[14] Liu B (2012) Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Pu

A Brief Author Biography

Mrs. Swathieswari Ramdass – currently pursuing M.phil. degree in Computer Science at A.V.C College(Autonomous), Mayiladuthurai. She received MCA degree in Gandhigram Rural University(Deemed University),Dindigul in 2012 . His Interested area includes Artificial Intelligence and Robotics

Dr. K. Thyagajan – Recieved Ph.d in Computer Science and working as a Head of the Department of Computer Science in A.V.C College(Autonomous),Mayiladuthurai. His Interested area Includes Data Mining.