IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-1, Jan 2019 E-ISSN: 2347-2693

Implementation of a Generalized, Real Time and Natural Language Processing Based Opinion Mining System for Twitter

Urmita Sharma[1*], Dhanraj Verma[2]

[1,2]Department of Computer Science, Dr A.P.J Abdul Kalam University, Indore, India

[*]*Corresponding Author: urmitasharma2018@gmail.com, Tel.: +91-97523-67914*

Available online at: www.ijcseonline.org

Accepted: 14/Jan/2019, Published: 31/Jan/2019

***Abstract*—**Success of any company or product depends on customer's satisfaction. If customers do not satisfied with the services or product provided by company, then certainly company needs to improve it. Opinion mining (OM) can help in doing this. OM is the process of computationally identifying and categorizing opinions from piece of text and determines whether the writer's attitude towards a particular topic or the product is positive, negative or neutral. This paper proposed a training model using sentdex data set to train the OM algorithm. This algorithm is based on supervised machine learning model to calculate OM of given text. Entire system is developed to calculate opinion from tweeters feeds. This system is working on real time data. Proposed system is designed for open field. One can take opinion of many field like political issue, product, company, person etc. this paper also presented the comparison of proposed results with well known python textblob API. textblob is used to perform many texts based operations. Sentiment analysis (OM) is one of them. In many OM systems this API is used.

***Keywords*—**Opinion Mining, Machine Learning, NLP, textblob, sentdex, NLTK.

I. INTRODUCTION

OM is a computation treatment of opinion and subjectivity in any text. Internet and web has made possible to find out opinion and experience of people. Most of the population post web content about their thinking. Most of the people do research on internet before purchasing any product, if the product reviews are good on internet than it will help us to take decision about the product. Now a day's customers are more interested to purchase five star rating items. OM can help us to take decision.

By the use of OM one can determine the current states of any product. Due to these companies are also interested in OM systems. Company can know about their product and services given to customers through a sophisticated OM system. OM can be use know the mental status of any person by analyzing his/her blogs and tweets [11].

Fact based analysis traditionally classifies the documents on the basis of topics. There can be many categories and for a given task it may be a binary classification or it could be thousands of classes based on the taxonomy used for classification. But in the OM classes are limited for example five star rating is considered on a positive rating. Also, the different classes and topics based calculation can be completely unrelated, while opinion labels can be setting numerical categories. Feature-based classification is also efficient in OM [12].

One of the most difficult parts while developing the OM is determining whether the text be analyzed contain subject or object text, or even both and where exactly in documenting its located. Opining mining is as reform disagreement and consensus point should also be represented and identify that form where this negative and positive sentiments coming form. [1]

Due to the short length text and unstructured content social media platform text analysis has many challenges. It is challenging task to recognition of name entities such as person, organization and location. Noun and pronoun phrase reorganization is also the typical task. Some time use of verbs is also confusing. For the good classification a sufficient quality training data set is required. But in social media platform data is irregular and has random pattern. So it is difficult to get sufficient data for training set. Other than these there are so many other factors affect the OM like spelling mistakes, use of abbreviations, poor punctuation, grammatical mistakes, incomplete sentences. Due to all this challenges accuracy of OM system is always under the question marks.

Social media and micro-blogging sites are huge source of information. These platforms are auto updated with current

issues and data. Public post everything, every time, about anything on it. Means one can find post related to everything. Frequency of post may vary but people continually post on social media. It is the easiest way to express emotions on public domain. So it is straightforward to understand the feeling of public about any issue, product or person. Companies are frequently using social media platform to get feedback and review of their products. Political parties use this platform to spread their ideas and thought to the media and also get the public opinion through this social media. But it is difficult and time-consuming process to do this manually. So it motivates used to develop a system that can calculate opinion polarities of any issue, person or product.
Rest of the paper organized as follows Section 2 contain work related on OM in the form of literature review. Section 3 described the proposed system. In section 4 methodologies used to implement the tool is explained. All the experimental results and analysis of results are presented in section 5. All research activity is concluded in section 6.

II. RELATED WORK

In Maria Del et al. (2016) [2] proposed an OM method for classification of features and news polarity. Proposed method is based on ontology-driven approach. Their research is focused on financial news OM. Authors address the problem of feature-based opining mining. In this work ontology approach was manual so that it is time-consuming and typical. Authors recommended developing ontology automatically or semi- automatically. Authors also recommended using this approach with some other domains like product or hotel review.

Nuno Olivera et al. (2017) [3] proposed a technique to visualize the impact of micro-blogging knowledge on exchange values. Their methodology used sentiments and a spotlight indicates from social media. They used a Kalman Filter to merge survey sources and micro-blogging. Their experimental results clearly show that micro-blogging knowledge place impact on exchange costs. Financial studies shows that social media opinion can affect medium to long-term returns. But authors result analysis proved that it can also affect the short term returns. Micro-blogging opinion has short term predictive power.

Shiliany Sun et al. (2017) [4] presented a study of natural language process (NLP) techniques for sentiment analysis. They mentioned general information science technique that is needed for text pre-processing like tokenization, word segmentation, a part of speech (PoS), tagging, parsing etc. in step with authors sentiment analysis is divided into 3 levels, first document level, second sentence level and third is fine-grained level. Authors conjointly enclosed completely different OM techniques in their analysis paper.

R. Piryani et al. (2017) [5] presented a study of OM and OM on analysis done from the year 2000 to 2016. They enclosed several papers indexed in web of Science (WoS). In their study they found several approaches of OM like machine learning and lexicon primarily based Storm Troops. They conjointly notice totally different level of research (document, sentiment or aspect-level). Authors gave a close analysis mapping of OM.

Mangi Kang et al. (2018) [6] proposed a text based hidden markov model (Text HMMs), sentiment analysis method. This method uses classification method using training set. This method defines hidden variables in Text HMMs. Authors presented experimental results on movie rating data set and found high accuracy. They found that their method is calculating opinion forms the sentences where no opinion words are presents. Authors also compared accuracy of proposed methods with some other methods. Like CNN-non-static, CNN-and, CNN-Multi, MV-RNN, RAE, NBSVM, MMB, Tree-CRF etc.

M. Rathan et al. (2018) [7] presented a feature level sentiment analysis model for micro-blogging site (Twitter). Emoji detection, spelling correction and emotion detection features were included for data mining. Automatic training data labeling was used to develop the training model. Lexicon based approach was used by author to design the training algorithm. This system was specially designed for smart phone. This system classifies the tweets extracted according to particular mobile phone. Support Vector Machine (SVM) classifier is used for classification. By the experimental results author shows that automated training data provide good accuracy then manual training data set. Multiple model of SVM classifier is used to get accuracy. Author also suggested that voting scheme based classifier can be used to get more accuracy.

Betul Dundar et al. (2017) [8] proposed a generic recommender system. This system is based on combination of OM and fuzzy quantification methods. This proposed system operated on two aspects. First is semantic orientation computation method that is used to reduce the features extracted and also reduce the opinion expressions. Second, summary opinions are presented by system using fuzzy quantification. Restaurant review data sets were used by authors to experiment their proposed work. The main objective of researchers in this literature to establish a decision tool for generating short summary opinion polarity from restaurant reviews. Their proposed system works in four phases i.e. pre-processing, feature extracting, feature reduction and opinion classification. In this work two main ideas are suggested. First, generation of fuzzy quantified sentences from qualitative data and second is web domain based PMI_IR method.

Soujanya Poria et al. (2016) [9] presented the deep learning approach in extracting opinions from text. Opinion holder parses the text to calculate opinion about products or service. Authors also developed some linguistic patters and integrated them with neural network. In this research work author proposed a deep CNN architecture with seven layers. One input layer, two convolution layers, two max-pooling layers, one fully connected layers and one output layer. Each convolution layer is followed by one max-pooling layer.

III. PROPOSED SYSTEM

Micro-blogging website (Twitter) has become a very popular tool of communication on internet. Information generated by one person and is consumed by many internet users. Through this implementation we have developed a tool that can be used to get public opinion through Twitter about any issue, product or person.

A. Tool Features

- *Real Time:* implemented tool takes real time tweets from twitter handler so that it will give real time opinion about any search.
- *Generalized*: This tool is not specific for any domain. This works for any issue, person product or topics. Anyone can randomly pickup a topic and start OM.
- *Text based OM*: Proposed tool uses twitter text material for OM. This current version of tool does not consider the images, audio, video and other multimedia content for OM.
- *Uses Supervised machine learning approach*: Proposed tool uses NLTK data set to train the algorithm. Then algorithm classified the tweets in different categories (positive, negative and neutral). Classification algorithm is based on natural language processing approach.
- *Opinion level*: Algorithm not only classifies the tweets into positive, negative and neutral. It also calculates the opinion level in numeric form. Positive number for positive polarity, negative number for negative polarity, and zero represents the neutral.
- *Overall and individual tweets analysis*: Proposed tool give both types of analysis. Individual tweets analysis (polarities and subjectivity) and overall opinion analysis of all sample tweets used for mining

B. Interface Designed

Entire tool is designed and developed in python programming language. Tkinter python module is used to design the interface of application tools. Complete interface is divided into four sections. First section is input section, in this section user has to give two main inputs, number of tweets (sample size) and search text (tweet filter). User has to select real time category of mining by checking the real time radio button. "Calculate opinion" button click starts the entire analysis process. Figure 1 shows the first section of interface

Figure 1: Input Section of OM Tool

Second section of interface is tweets section. In this section all the individual tweets are visible with their tweet number, retrieval date, retrieval time, tweet text (in Jason format), and opinion polarity of proposed algorithm with subjectivity of tweet. Figure 2 shows the Tweets section of application tool.

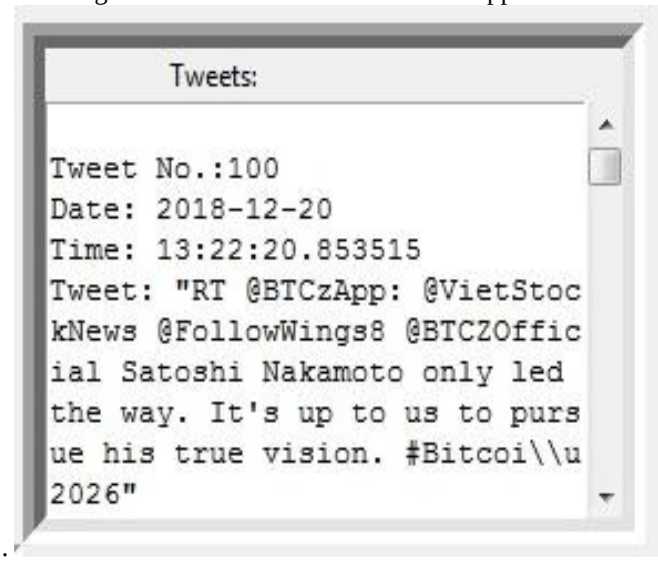

Figure 2: Tweet Section of OM Tool

Third section of interface represents the tweets analysis by counting both algorithms opinion in the form of positive, negative and neutral opinions as shown in figure 3. Existing label shows the result analysis of textblob based algorithm and proposed label shows the sentdex API based algorithm results. Two buttons (Show Graph) are also available to show the results in form of bar chart. Figure 4 and 5 shows the sample bar chart of both algorithms results.

Fourth section of interface shows the overall analysis results of OM. There are mainly three information is represented, Opinion value of proposed algorithm (Sentdex based), opinion value of existing algorithm (textblob based) and accuracy of proposed algorithm. Figure 6 shows the overall analysis section of tool.

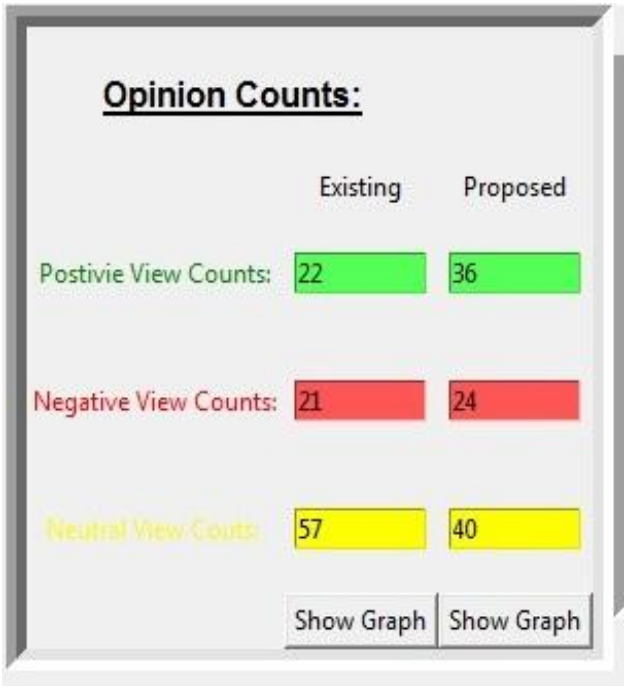

Figure 3: Analysis Section of OM Tool

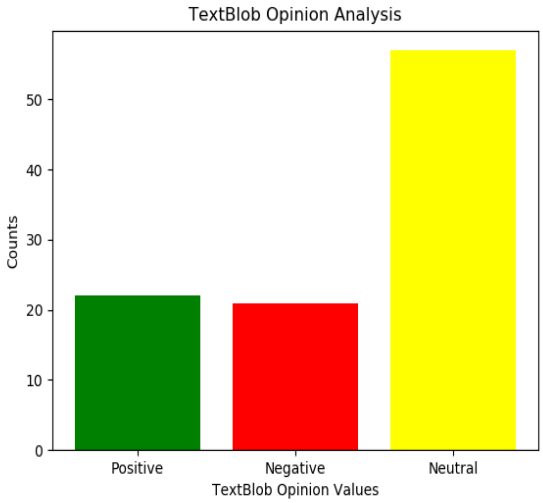

Figure 4: Barchart of textblob API Based Algorithm

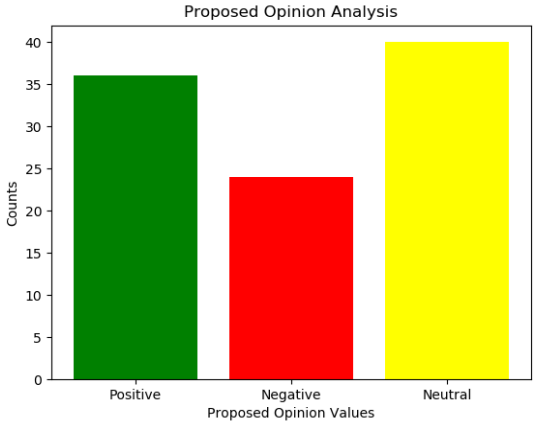

Figure 5: Barchart of sentdex API Based Algorithm

Fourth section of interface shows the overall analysis results of OM. There are mainly three information is represented, Opinion value of proposed algorithm (sentdex based), opinion value of existing algorithm (textblob based) and accuracy of proposed algorithm. Figure 6 shows the overall analysis section of tool.

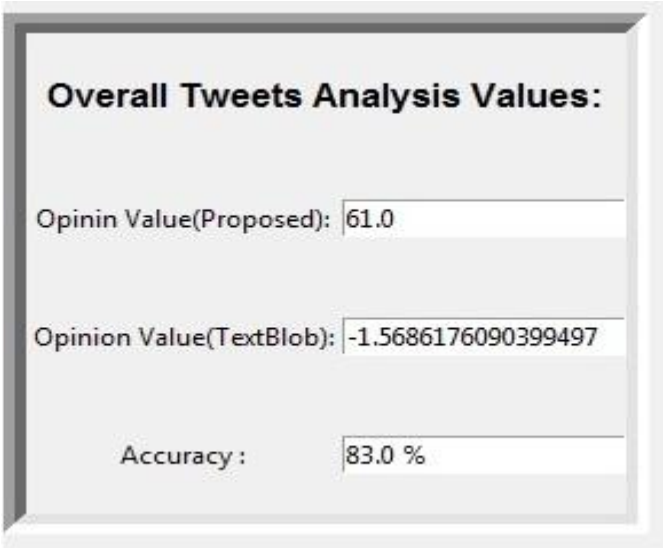

Figure 6: Overall Analysis Section of OM Tool

IV. METHODOLOGY

This section covers the tools and technology used to develop the proposed application. This section also covers the data set details and application flow.

A. Tools and Technology Used:

Proposed implementation is developed using python programming language. Python 3.6 version is used to implement this version. Although any python editor can be use to write python script but we have used JetBrains

PyCharm Community 2018 edition to write the python script. Windows 7 ultimate with service pack 1 or higher version can be used to execute this tool. Following common python modules are used in implementation.

- *tweepy*- module provides all supporting classes and method that can be used to access twitter data.
- *tkinter*- module is used to design the interface of application.
- *nltk*- is natural language toolkit. It provides data set for training the algorithm.
- *textblob*- module provides the various text operations on text. Sentiment calculation is one of them.
- *matplotlib*- module is used to represents the results in graphical form like pi chart, bar chart etc.

B. Data Set:

This tool is developed for real time twitter data. So that we are not using any predefined data sets for experiments. Application directly extracts data form twitter and calculates opinion on that data. This data is mostly text format data. We have not included multimedia data for mining the tweets. This application is devoted to text mining. In the application, data is represented in the Jason format, as shown in figure 7. NLTK data set [10] is used as a training data set to train the algorithms training models.


```
Tweet No.:20
Date: 2018-12-20
Time: 13:20:41.500000
Tweet: "@AfDimBundestag @TilSchweiger 12.350 ? \\ud83e\
 mit dem Geld gemacht? Bitcoin im Rausch gekauft?\",\"c
04],\"source\":\"\\u003ca href=\\\"http:\\/\\/twitter.c
\"\\u003eTwitter Web Client\\u003c\\/a\\u003e\",\"trunc
o_status_id\":1075415131960549377,\"in_reply_to_status_
49377\",\"in_reply_to_user_id\":910913583168573440,\"ir
"910913583168573440\",\"in_reply_to_screen_name\":\"AfD
"id\":45252204,\"id_str\":\"45252204\",\"name\":\"Seb G
sebseb7\",\"location\":\"Germany\",\"url\":\"https:\\/\
cription\":\"blockchain craftsmanship  http:\\/\\/crypt
ive-Streams YT:http:\\/\\/www.youtube.com\\/c\\/sebGree
"none\",\"protected\":false,\"verified\":false,\"follow
s_count\":1226,\"listed_count\":26,\"favourites_count\"
25282,\"created_at\":\"Sun Jun 07 00:58:14 +0000 2009\"
me_zone\":null,\"geo_enabled\":true,\"lang\":\"en\",\"c
se,\"is_translator\":false,\"profile_background_color\"
```


Figure 7: Jason Format of Data Retrieval from Twitter

C. Application Flow

Figure 8 shows the flow of application execution. Main steps of application execution are.

- Input Data: Sample size and filter text are required as a input.
- Extracting Tweets: Tweets from twitter handler needs to be extracted from opinion calculation.

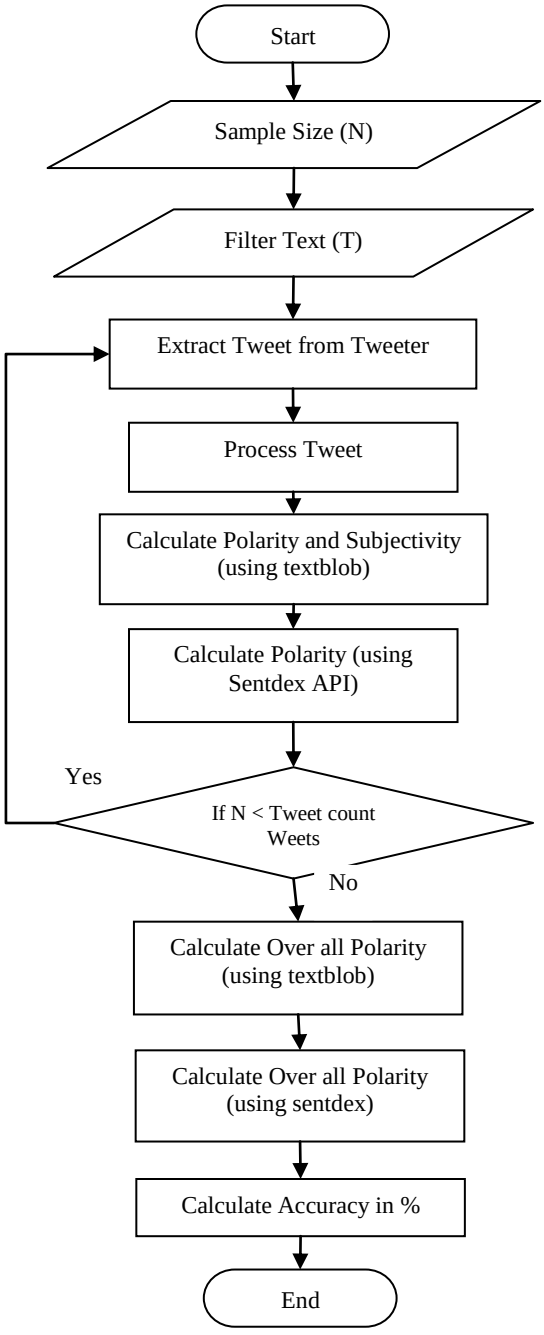

Figure 8: Flow Chart of Entire Process of Execution

- Process the Tweets: individual tweets are processed to calculate the polarity and subjectivity by textblob API and after sentdex API for polarity calculation.
- Repeat Step ii and iii: Need to repeat step ii and iii until tweets count reaches to sample size.
- Overall Polarity Calculation: Overall polarity of textblob and sentdex API are calculated. Overall polarity is average of individual polarities and it is calculated using equation 1.

$$P = (\sum_{i=0}^{n} Pi)/n \quad (1)$$

Where P is overall polarity, Pi is polarity of i^{th} tweet and n sample size.

Accuracy Calculation: Proposed tool's accuracy calculated and display in this step. Accuracy of proposed sentdex API based algorithm is calculated by taking the textblob API based algorithm as a reference.

V. RESULT ANALYSIS

This section covers the experimental results of proposed tool. This section also included some observations taken during the development and execution of application.

A. Experimental Results

This OM tool is tested on some different domain subjects as shown in table 1. Table shows the information of 10 executions on some topics like bitcoin (Crypto-currency), cricket (game), AQUAMAN (movie), iPhone x (Mobile phone) and Mortal Engine (Movie). Table 1 also represent the date and time of execution with sample size (number of Tweets) in each execution.

Table 1: Experimental Data set Information

S.No.	Date	Start Time	End Time	Analysis Text	Number of Tweets
1	23-10-18	1:38:00 PM	1:41:00 PM	Bitcoin	100
2	23-10-18	1:43:00 PM	1:46:00 PM	Bitcoin	50
3	23-10-18	1:48:00 PM	1:53:00 PM	Bitcoin	100
4	15-12-18	1:35:11 PM	1:46:01 PM	Bitcoin	300
5	15-12-18	2:03:13 PM	2:03:13 PM	cricket	90
6	15-12-18	2:05:04 PM	2:09:22 PM	cricket	100
7	15-12-18	2:18:21 PM	2:20:57 PM	cricket	100
8	15-12-18	2:34:24 PM	2:36:02 PM	AQUAMAN	100
9	15-12-18	2:36:12 PM	2:47:47 PM	iPhone x	93
10	15-12-18	2:41:46 PM	3:11:46 PM	Mortal Engines	39

Table 2 shows the textblob API based OM results. Table 2 contains the positive, negative and neutral opinion about particular topics.

Table 2: Experimental Results of Textblob API based OM

S.No.	Analysis Text	Positive	Negative	Neutral
1	Bitcoin	37	17	46
2	Bitcoin	19	5	26
3	Bitcoin	38	26	36
4	Bitcoin	77	85	138
5	cricket	45	19	26
6	cricket	44	27	29
7	cricket	36	37	27
8	AQUAMAN	24	12	64
9	iphone x	17	13	63
10	Mortal Engines	9	11	19

Table 3 shows the sentdex API based OM. Table 3 also contains the polarity counts of different tweets.

Table 3: Experimental Results of sentdex API based OM

S.No.	Analysis Text	Positive	Negative	Neutral
1	Bitcoin	50	19	31
2	Bitcoin	23	7	20
3	Bitcoin	57	16	27
4	Bitcoin	121	69	110
5	cricket	36	14	40
6	cricket	45	20	35
7	cricket	35	27	38
8	AQUAMAN	19	11	70
9	iphone x	20	12	61
10	Mortal Engines	7	9	23

Table 4 represents the overall opinion values of both algorithm and this table also represents the accuracy of sentdex API based algorithm results with reference to textblob API based results.

Table 4: Overall Opinion Value and Accuracy of Proposed Algorithm

Analysis Text	Opinion (Textblob)	Opinion (sentdex)	Accuracy(%)
Bitcoin	10.39	134	85%
Bitcoin	4.64	49	88%
Bitcoin	6.5	163	81%
Bitcoin	-2.3182	218	85.33%
cricket	5.40756	105	82.00%
cricket	3.39413	63	93.00%
cricket	-2.3329	70	89.00%
AQUAMAN	2.33641	22	94.00%
iphone x	3.52534	47	96.77%
Mortal Engines	-1.2746	-16	89.74%

Figure 9, 10 and 11 shows the comparison of positive, negative and neutral opinion of textblob and sendex API based algorithm respectively

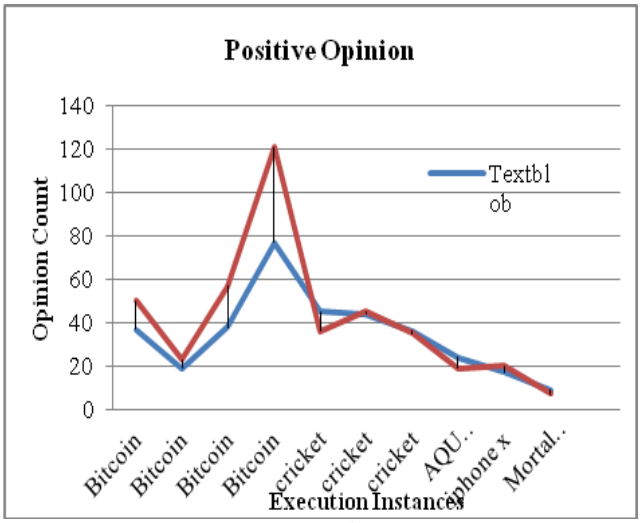

Figure 9: Positive Opinion of textblob and sentdex

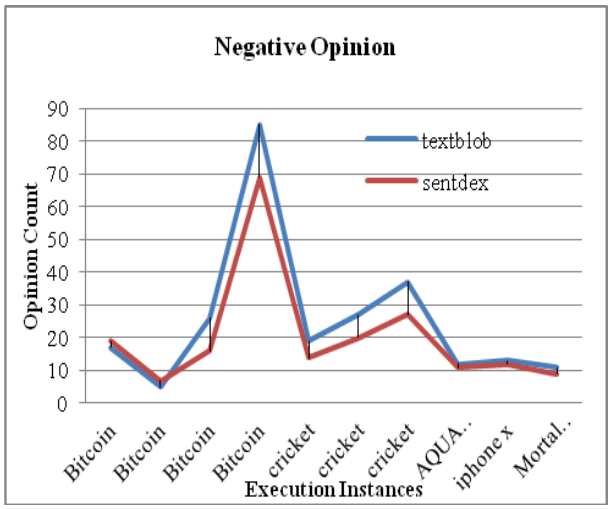

Figure 10: Negative Opinion of textblob and sentdex

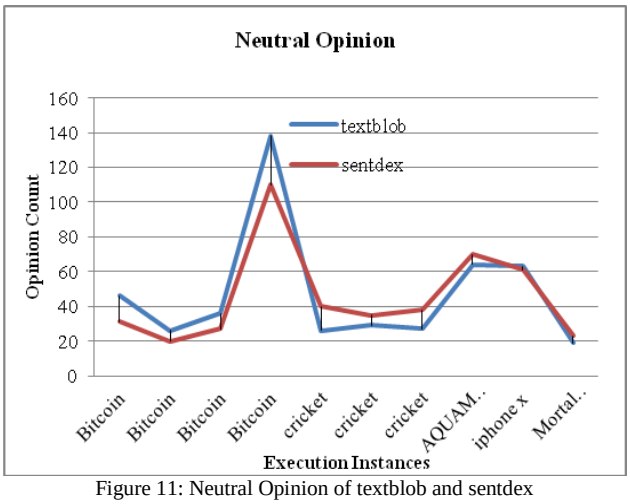

Figure 11: Neutral Opinion of textblob and sentdex

B. Observations

During the implementation and experiments certain observations has been recorded.

- This is a real time analysis application, so that it works on real time twitter data. Application fetch data from twitter while execution. Application is highly dependent on internet connection. If internet connection is weak or discontinue than application generate exception and won't be able to generate results of actual sample size selected by user. But it will give analysis results of extracted tweets.
- Twitter is open platform for OM but it is seen that some post related to current issues are more frequent than other. Same thing also affect proposed tool performance also. As we observed that tweet retrieval is fast for current issue, but it is slow for others.
- Twitter text are very unstructured and there are lots of other impurities in content like spelling mistakes, grammatical mistakes, use of abbreviations, multi mining words etc. Due to this many times algorithms are not able to calculate actual polarity of text. In that case algorithm shows the zero polarity of text and tweet will be considered as neutral tweet. It also affects the result of opinion. It is also difficult to distinguish between tweets with zero polarity and tweets with no results.
- It there are more than one issue related to same keywords than opinion results of one issue may affected by others.

VI. CONCLUSION AND FUTURE ENHANCEMENTS

The proposed implemented tool provides generalized OM. Through this tool analysis of any trendy topic can be done easily. Proposed tool gives real time OM. It uses supervised machine learning algorithm to calculate opinion. Two different algorithms comparative results are calculated and presented in this tool. NLTK data set is used as a training data set. textblob and sentdex API is used for calculating opinion. 88% accuracy is archived through sentdex API. For calculation of accuracy we have considered textblob OM as ideal analysis. Proposed tool works on real time data. So internet connectivity is required to run tool.

Input data is taken form twitter. Data on twitter are vary unstructured that can affect the results of analysis. Tool performance also affected through trendiness of subject. If subject is trendy data extraction and analysis will be fast and if subject is not form current issues than analysis will take time.

Current implementation takes only text data for analysis. In future images, audio, videos and other multimedia information can be considered for analysis. This tool work on real time data, similarly this tool can be extending for other stored data. As of now twitter is only data source for this tool. We can also extend this tool for other data sources like news websites, other social media platforms etc.

REFERENCES

[1] Farhan Hassan Khan, Saba Bashir and Usman Qamar, "*TOM: Twitter opinion mining framework using hybrid classification scheme*", Decision Support Systems, Vol. **57**, pp. **245–257** , **2014**.

[2] Maria del Pilar Salas, Rafael Valencia, Antonio Ruiz and Ricardo Colomo, "*Feature-based opinion mining in financial news: An ontology-driven approach*", Journal of Information Science, Vol. **34**, Issue **4**, pp. **458-479**, **2016.**

[3] Nuno Oliveira, Paulo Cortez and Nelson Areal, "*The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices*", Expert Systems With Applications, Vol. **73**, pp. **125–144, 2017.**

[4] Shiliang Sun, Chen Luo, Junyu Chen, *"A Review of Natural Language Processing Techniques for Opinion Mining Systems*", Information Fusion, Vol. **36**, pp. **10-25**, **2017**.

[5] R. Piryani, D. Madhavi and V.K. Singh, "*Analytical mapping of opinion mining and sentiment analysis research during 2000–2015*", Information Processing and Management, Vol. **53**, pp. **122-150, 2017**.

[6] Mangi Kang, Jaelim Ahn and Kichun Lee, "*Opinion mining using ensemble text hidden Markov models for text classiÞcation*", Expert Systems With Applications ,Vol. **94**, pp. **218-227**, **2018**.

[7] M. Rathan, Vishwanath R. Hulipalled, K.R. Venugopal and L.M. Patnaik, "*Consumer Insight Mining: Aspect Based Twitter Opinion Mining of Mobile Phone Reviews*", Applied Soft Computing Journal, Vol. **68**, pp. **765-773**, **2018**.

[8] Betoul Duondar,Diyar Akay, Fatih Emre Boran and Suat Ozdemir, "*Fuzzy Quantification and Opinion Mining on Qualitative Data using Feature Reduction*", International Journal of Intelligent System, Vol. **33**, Issue **9**, pp. **1840–1857**, **2017**.

[9] Soujanya Poria, Erik Cambria and Alexander Gelbukh, "*Aspect Extraction for Opinion Mining with a Deep Convolutional Neural Network*", Knowledge-Based Systems, Vol. **108**, pp. **42-49**, **2016**.

[10] Bird, Steven, Edward Loper and Ewan Klein, "*Natural Language Processing with Python*", O'Reilly Media Inc., **2009**.

[11] Shrija Madhu, "*An approach to analyze suicidal tendency in blogs and tweets using Sentiment Analysis*", International Journal of Scientific Research in Computer Science and Engineering, Vol.**6**, Issue.**4**, pp.**34-36**, **2018.**

[12] Ketan Sarvakar, Urvashi K Kuchara, "Sentiment Analysis of movie reviews: A new feature-based sentiment classification", Vol.**6**, Issue.**3**, pp. **8-12** , **2018**

Authors Profile

Mrs Urmita Sharma pursed Bachelor of Science from Devi Ahilya University of Indore, in 1998 and Master of Computer Management from Shri Vaishnav Institute of Management and Science, DAVV University in 2001. She is currently pursuing Ph.D. from Dr A.P.J Abdul Kalam University and currently working as Assistant Professor in Department of Computer Science at S.J.H.S Gujarati Innovative College of Commerce and Science, since 2001. Her main research work focuses on Machine Learning Algorithms, Opinion Mining,Data Mining.She has 18 years of teaching experience and 2 years of Research Experience.

Mr Dhanraj Verma pursed Bachelor of Science from Vikram University of Ujjain, in 1997 and Master of Technology from Devi Ahilya University in 2007. Ph.D. from BU University in 2013.Currently working as Professor in Department of Computer Science & Engineering at Dr A.P.J Abdul Kalam University. He is a member of IEEE & IEEE computer Society since 2012. Life member of CSI Since 2012. He has published more than 24 research paper national/international journals. His main research work focuses on Network Security,Cloud Security & Privacy, Big data Analytics,IoT & Computational intelligence.He has 18 years of teaching experience and 6 years of Research Experience.

