Sentiment Analysis of English Tweets Using Data Mining

Article *in* INTERNATIONAL JOURNAL OF COMPUTER SCIENCES AND ENGINEERING · October 2018

DOI: 10.26438/ijcse/v6i10.276284

2 authors, including:

Seema Baghla
Punjabi University, Patiala
7 PUBLICATIONS **10** CITATIONS

SEE PROFILE

IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-10, Oct 2018 E-ISSN: 2347-2693

Sentiment Analysis of English Tweets Using Data Mining

Amritpal Kaur[1*], Seema Baghla[2]

[1,2]Dept. of Computer Engineering Yadavindra College of Engineering, Punjabi University Guru Kashi Campus, Talwandi Sabo, Bathinda, Punjab, India

[*]*Corresponding Author: garg_seema238@yahoo.co.in, Mob.8146551446*

Available online at: www.ijcseonline.org

Accepted: 18/Oct/2018, Published: 31/Oct/2018

***Abstract*—**Social media has been used for expressing and sharing the thoughts of people with different events. Sentiment analysis is being used for computing and satisfying a view of a person given in a piece of a text, to identify persons thinking about any topic is positive, negative or neutral. In the present work sentiment analysis has been used to analyze people's sentiments, opinions, and emotions towards entities. In this work, sentiment 140 tools have been used for the collection of tweets on different topics. The collected tweets have been preprocessed. Different techniques have been used to present work. Classification technique has been used for the analysis of tweets how many positive, negative and neutral tweets. Sentiment analysis algorithm has been used to analyze tweets whether tweets are positive, negative or neutral. An autocorrect option has been also used to correct the sentence. Sentiment analysis has been used parameters such as accuracy, predictive and automation.

***Keywords*—**Data Mining, Sentiment Analysis, Twitter, Classification.

I. INTRODUCTION

Data mining extract the information from a huge set of data. This is a method for deciding plans and extracting the information from a huge set of data. It is the procedure of mining knowledge from data. Data mining is also called knowledge discovery in databases. It is the process of finding interesting and useful patterns in a large set of data [1].The main aim of the data mining process is to gather information from a large data set and convert it into clear form [1]. Sentiment analysis also called opinion mining that analyzes people's sentiments, opinions, and emotions towards entities [2]. These entities might be a thing or a film, surveys of people, products, issues, and topics that truly matters. Social sites, for example, Facebook and Twitter are that, where characters put the status or sentiments. People tweet in their twitter account on any event, product or topic [3].

II. RELATED WORK

Mehta and Jain (2016) [1] discussed sentiment mining and related classifiers: a review. Discussed all the fundamental details about opinion analysis. It compromises of current research and available scope of sentiment mining. Also, the information regarding basic workflow of the opinion mining process, recent trends, and applications of sentiment analysis has been explained amazingly. **Chandani et al. (2015) [2]** discussed sentiment analysis and its challenges. Discussed recently proposed algorithms, approaches and its applications briefly. The main goal of this paper was to provide a complete description of sentiment analysis of opinion mining technique and the challenges faced by it. Also, include the social issue regarding manipulation, privacy and economic impact faced during the development of information seeking services. **Rajan and victor (2014) [3]** discussed web sentiment analysis for scoring positive or negative words using tweeter data. The sentiment examination was done on a per tweet basis. The words in each tweet were compared with those in other tweets that were previously labeled as "positive" or "negative". This approach was used to obtain the significant features and to analyze the overall sentiment for each object.

Gupta et al. (2017) [4] discussed sentiment analysis of tweets using machine learning approach. This research has reviewed different machine learning approaches and compared their results. This examination concentrated on two machine learning calculations K- nearest neighbors (KNN) and support vector machines (SVM) in a hybrid manner. **Alvares et al. (2016) [5]** discussed sentiment analysis using opinion mining. The idea of this paper is to "process a set of user reviews for a given product, generating a summarization (quality, features) and aggregating of user opinion". This research aimed to shift focus on the user's opinion after semantically analyzing and mining the data to find the hidden sentiment it. The evaluation analysis showed

that the proposed hybrid approach was better in terms of accuracy and F- measures as compared to individual classifiers.

Bandgar and sheeja (2016) [6] discussed analysis of real time social tweets for opinion mining. Discussed original windows based easy to understand application in java to concentrate and process tweet using unstructured models. Tweets were obtained for sentiment analysis and divided into three different sentiments positive, negative and neutral by using unstructured algorithm such as elliptic curve cryptography (EEC) and inter process communication (IPC). Their results were compared using the confusion matrix, precision and accuracy parameters. The results were visualized using the pie graph. **Dattu et al. (2015) [7]** discussed a survey on sentiment analysis on twitter data using different techniques. Discussed interpreting the reasons of the sentiment change in public opinion mining and summarizing products reviews to solve the polarity shift problem. Different algorithms/models were used to perform the above tasks like latent dirichlet allocation (LDA), naïve bayes (NB) classifier and support vector machine (SVM).

Oleary (2015) [8] discussed twitter mining for discovery, prediction and causality: applications and methodologies. Discussed a portion of the methodologies used to accumulate data information from Twitter for Twitter mining. What's more, this paper audits some of the applications that utilize Twitter Mining, exploring Twitter data for expectation, disclosure what's more, as notify assumption of cause. **Sabarmathi and Chinnaiyan (2017) [9]** discussed reliable data mining tasks and techniques for industrial applications. Discussed various data mining applications and techniques. Thus data mining is said to be a promising interdisciplinary area in the field of research.

Rahmath (2014) [10] discussed opinion mining and sentiment analysis –challenges and applications. Discussed opinion mining and sentiment analysis extract and classify people's opinion automatically from the internet. This paper discusses various application and challenges related to the opinion mining and sentiment analysis. **Smeureanu and Bucur (2012) [11]** discussed applying supervised opinion mining techniques on online user reviews. Discussed an algorithm for detecting sentiments on movie user reviews, based on naïve bayes (NB) classifier. They also analyzed the opinion mining domain, techniques used in sentiment analysis and its applicability.

Umar and Chiroma (2016) [12] discussed data mining for social media analysis: using twitter to predict the 2016 US presidential election. Discussed a little- estimated information was removed from Twitter and investigated utilizing information mining arrangement calculation known as assessment examination. Also, the aftereffect of the examination was utilized to predict the result of the previously mentioned. **Sutar et al. (2016) [13]** discussed sentiment analysis: opinion mining of positive, negative or neutral twitter data using hadoop. This paper gave a method for analysis of Twitter data using AFFIN and EMOTICON for natural language processing. An open source framework Hadoop used to store and process large sentiments.

Sheela (2016) [14] discussed a review of sentiment analysis in twitter data using Hadoop. Discussed a technique which performed grouping of tweet sentiment in twitter was talked about. At long last, extensive tests will be conducted on real-world data, with an expectation to achieve comparable or greater accuracy than the proposed techniques in the literature. A method to predict or deduct the location of a tweet based on the tweet's information. **Hridoy et al. (2015) [15]** discussed localized twitter opinion mining using sentiment analysis. Discussed a procedure which allows utilization and interpretation of Twitter data to determine public opinions. Feature specific popularities and male-female specific analysis have been included. Mixed opinions were found but general consistency with outside reviews and comments were observed.

III. METHODOLOGY

There are two techniques used in proposed work.

I. Sentiment Analysis [5]

Sentiment analysis is done through two types of procedures as below:

a. Sentiment arrangement utilizing regulated learning supervised learning is actualized by making a classifier. It requires two arrangements of reports for order one is preparing set other is trying a set. This strategy is otherwise called machine learning technique.

b. Sentiment arrangement utilizing unsupervised learning: In the unsupervised order, the content is characterized by contrasting it and given words or dictionaries. The feeling an incentive for these words or dictionaries is already characterized.

II. Classification [5]

Classification assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data.

There are two flowcharts used for the proposed work. One algorithm is used for tweets analysis and another algorithm is for sentiment analysis. Java language has been used for proposed work. Java is high level object oriented programming language. Net Beans have been used for Java codes as a front end. It provides proper error handling mechanism. Structure query language (SQL) has been used as a database. Figure 1 represents the flow of the proposed work that has been used for the classification process.

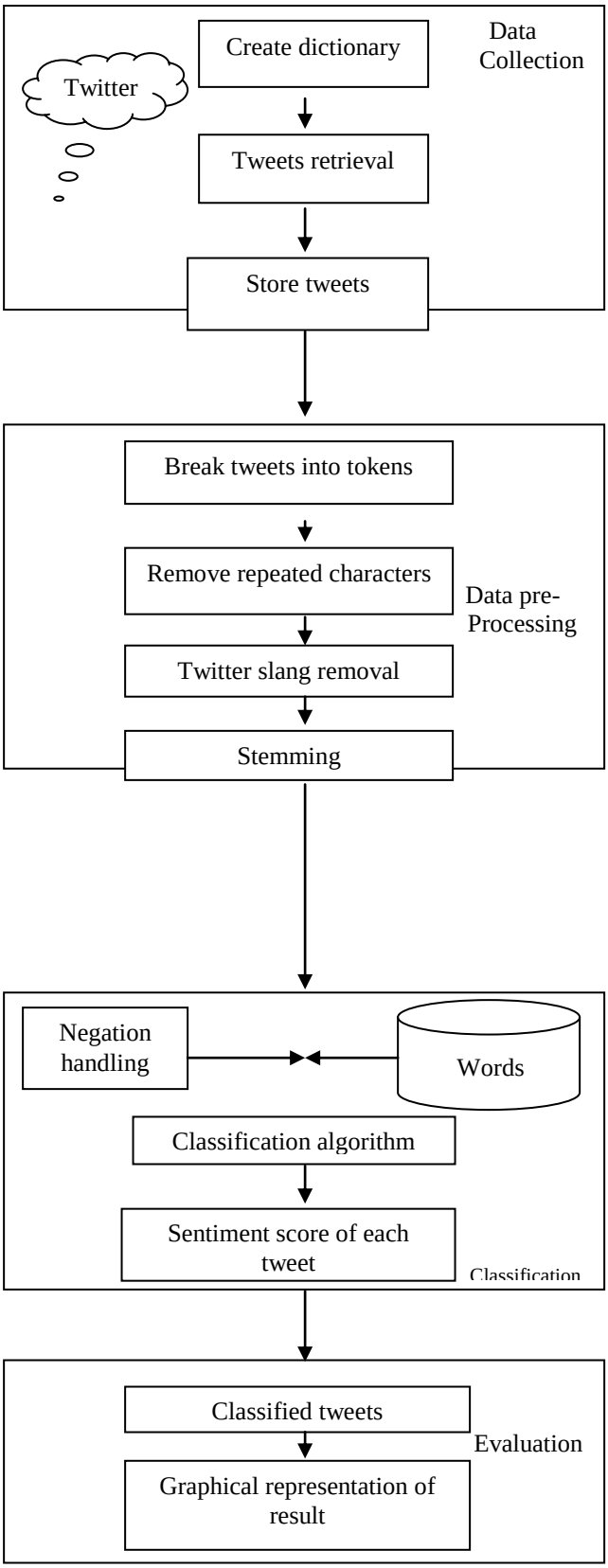

Figure 1. Flow chart of proposed work

I. Create dictionary: Make a dictionary of positive and negative words. Two different tables are created in the sentiment database one for positive words and other for negative words. Firstly make a dictionary of positive and negative words.

II. Tweets collection: The tweets are collected from the twitter. Firstly one has to create a twitter account then login to that account to collect the tweets. Tweets are collected from www.sentiment140.com website and structured query language database used to store these tweets.

III. Data pre-processing: Pre-processing removes stop word, handling of negation; misspell correction, positive word lists of each tweet and negative word lists of each tweet.

a. Filtering: Filtering is a process that removes unnecessary parts or information from the sentence. Filters used in many ways.
b. URL: Entire URL removed from the sentence or input file after checking the whole sentence or input file. These links are replaced by the empty space.
c. Usernames: Sometimes user used any username or @ symbol before any tweet. These types of usernames or @ replaced by empty space.
d. Duplicate or repeated characters: Users sometimes use informal language in tweets. For example, users mostly write 'baaad' in place of the bad word.

IV. Twitter slang removal: Most of the users prefer to write a short form of the actual words. These short forms of words called slang words. For example, thnx is used in place of thanks. These slang words then change into their actual meaning.

V. Stop words removal: Stop words are the words which are mainly used in tweets or comments but this does not add to the sentiment. Stop words are articles, prepositions etc. These should be removed from the document and replaced by the empty space.

VI. Negation handling: There are some words which change the meaning of sentence these words are known as negation words. Words like never, not, do not, no, nor are the negation words. If the tweet is positive these words change the sentiment of the tweet to negative. So these are handled with the proper method.

a. Negation word used with positive word and it make it negative: In this, if the whole sentiment of sentence is positive, but the positive word preceded by negation then the sentiment of sentence is changed to negative.

- Story of serial is good: This sentence gives the positive sentiment as the positive word good is present here.

- Story of serial is not good: This sentence has negation word 'not', which changes the sentiment of the sentence to negative sentence.

b. Negation word used with negative word and makes it positive: In this, if the whole sentiment of the sentence is negative, but the negative word preceded by negation then the sentiment of the sentence is changed to positive.

- Story of serial is bad: This sentence gives the negative sentiment as the negative word bad is present here. Now consider the case:
- Story of serial is not bad: This sentence has negation word 'not', which changes the sentiment of the sentence to positive sentence.

VII. Stemming: It is the process to convert the words into their original form. Sometimes users use the stemmed words for the original words which should be replaced by actual words. For example, hate hated, hates, hating all belong to the single word hate. It has been increased the efficiency of the software.

VIII. Calculating sentiment score: Sentiment score is calculated by comparing the words from the tweets with the dictionary words.

IX.

Figure 2 represents the flow chart for sentiment analysis. A list of tweets collected from Twitter; calculate sentiment score for each tweet.

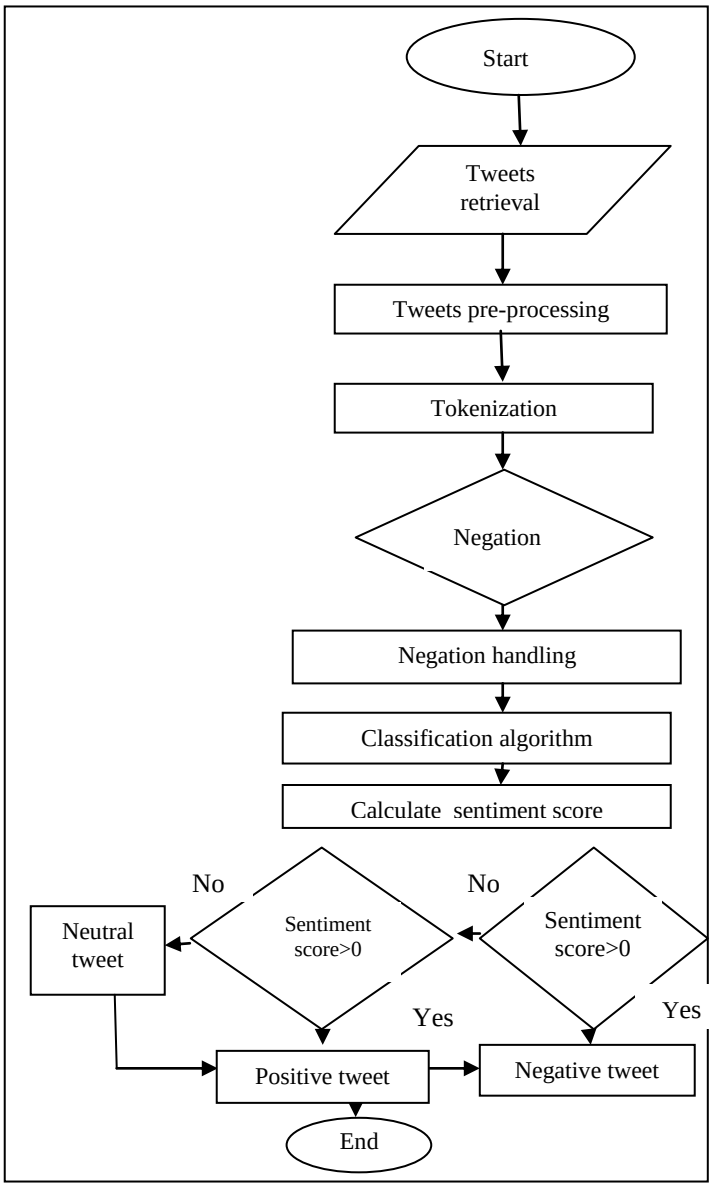

Figure 2. Flow chart for sentiment analysis

IV. RESULTS AND DISCUSSIONS

I. There have been four datasets. Different tweets have been collected for each dataset.

a. Results for demonetization dataset: Total 74 tweets have been collected. The algorithm has been applied to them. The overall sentiment of tweets shows that the opinion of the public towards the demonetization is positive. 43 tweets from the total tweets are calculated with the wrong sentiment, 22 as positive tweets, 20 as negative tweets and 32 as neutral tweets. The software calculated the sentiment with the efficiency of 42%. Figure 3 shows the analysis of the demonetization tweets.

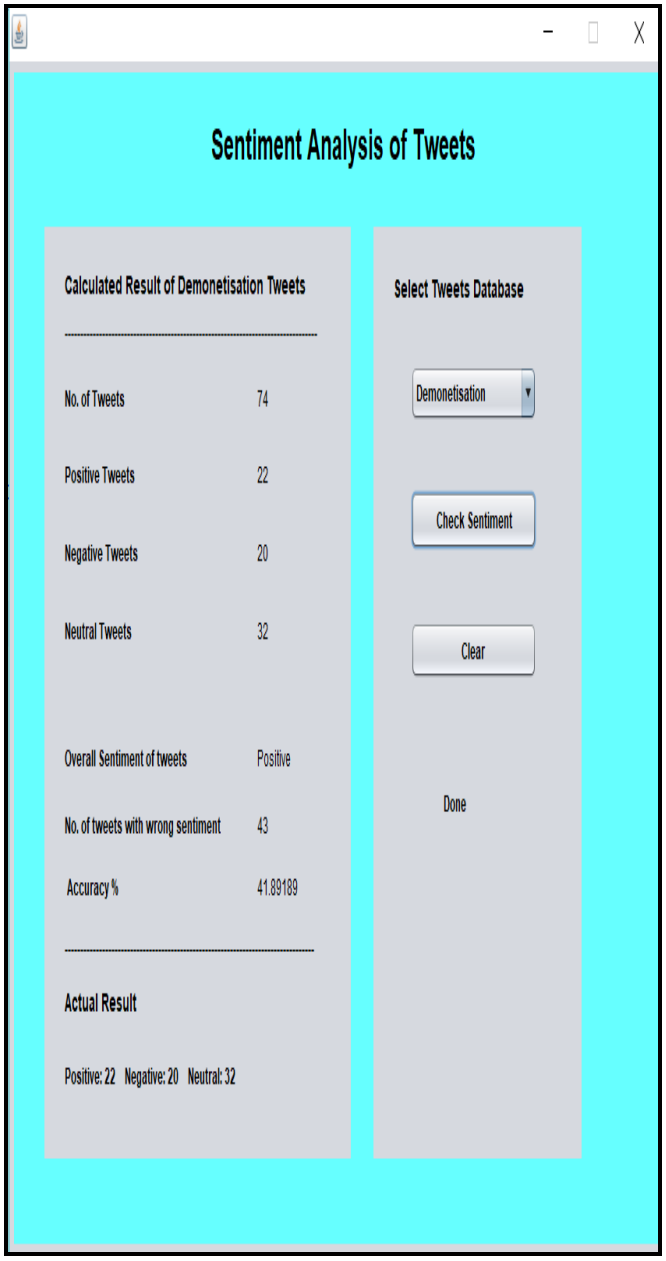

Figure 3. Results of demonetization tweets

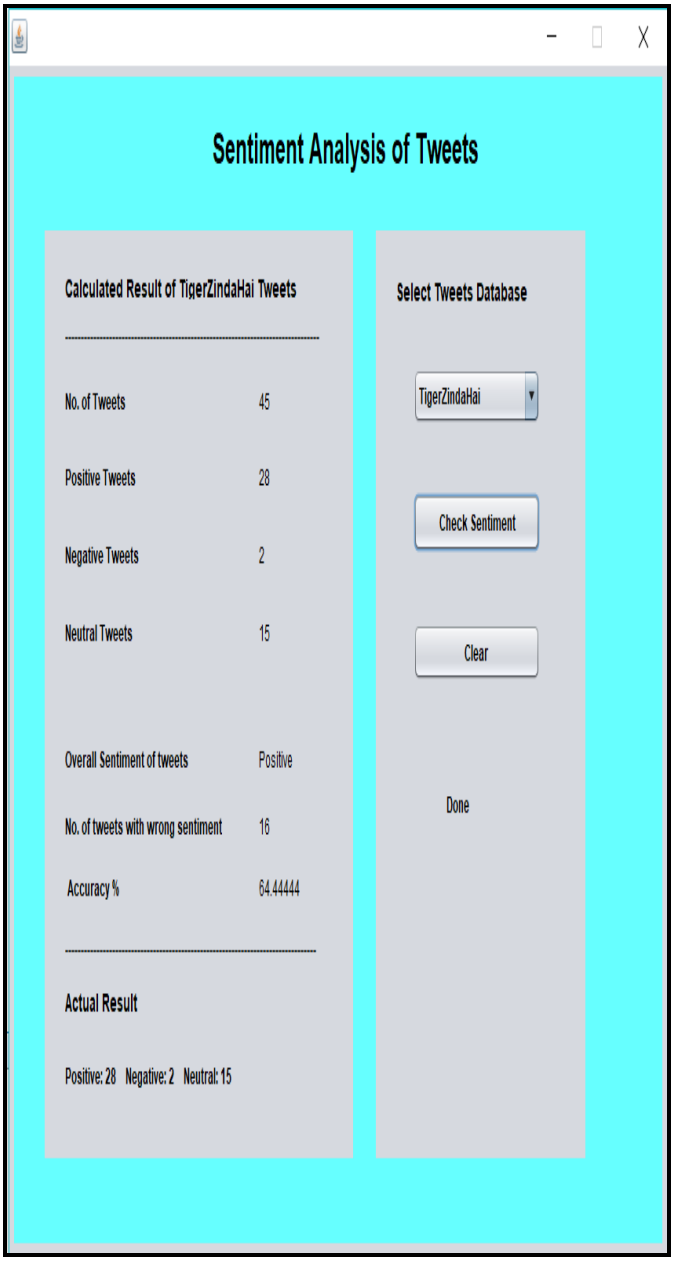

Figure 4. Results of tiger zinda hai movie tweets

b. Results for tiger zinda hai movie dataset: Total 45 tweets have been collected. The algorithm has been applied to them. The overall sentiment of tweets shows that the opinion of the public towards the movie is positive. 16 tweets from the total tweets are calculated with the wrong sentiment, 28 as positive tweets, 2 as negative tweets and 15 as neutral tweets. The software calculated the sentiment with the efficiency of 64.44%. Figure 4 shows the analysis of the tiger zinda hai movie tweets.

c. Results of padmaavat movie dataset: Total 21 tweets have been collected. The algorithm has been applied to them. The overall sentiment of tweets shows that the opinion of the public towards the padmaavat movie is positive. 9 tweets from the total tweets are calculated with the wrong sentiment, 10 as positive tweets, 2 as negative tweets and 9 as neutral tweets. The software calculated the sentiment with the efficiency of 57.14%. Figure 5 shows the analysis of the padmaavat movie tweets.

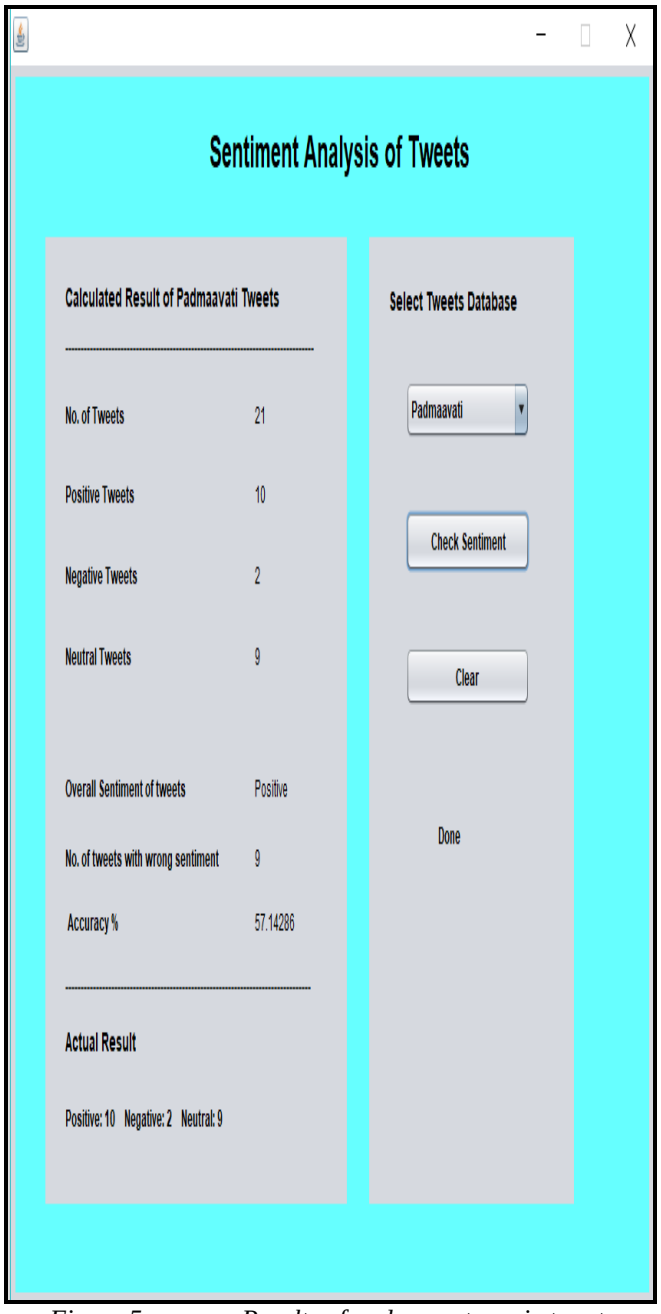

Figure 5. Results of padmaavat movie tweets

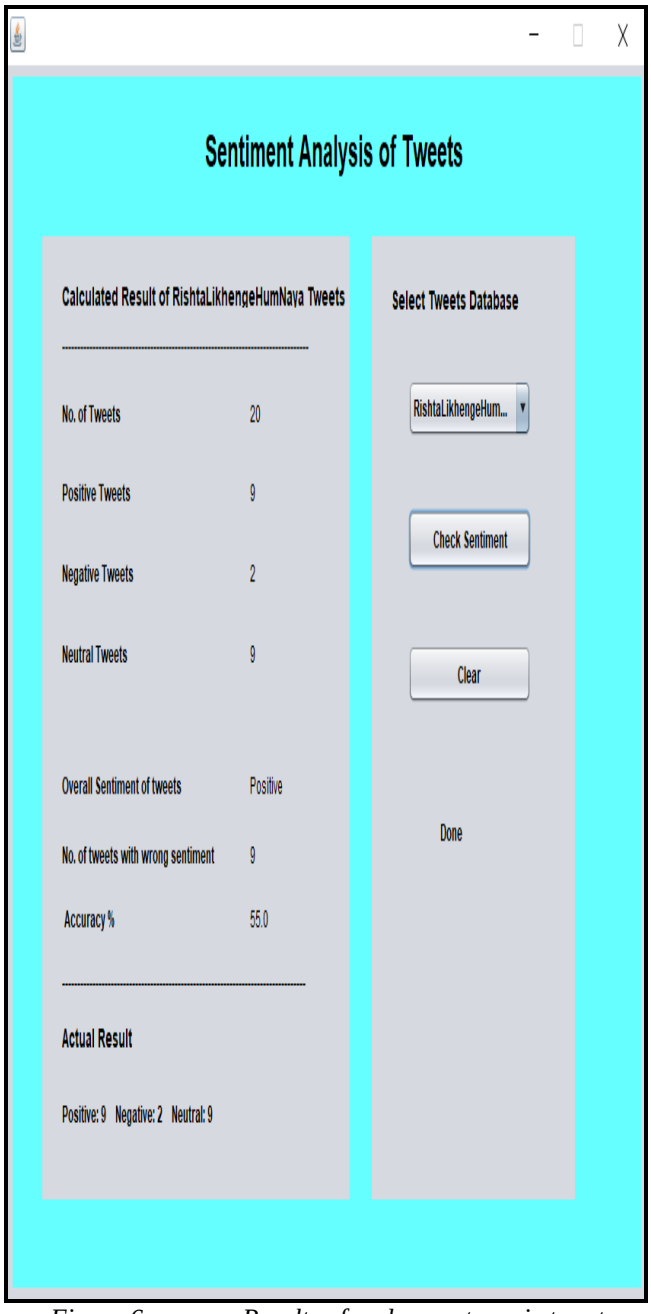

Figure 6. Results of padmaavat movie tweets

d. Results of rishta likhenge hum naya Hindi serial dataset: Total 20 tweets have been collected. The algorithm has been applied to them. The overall sentiment of tweets shows that the opinion of the public towards the padmaavat movie is positive. 9 tweets from the total tweets are calculated with the wrong sentiment, 9 as positive tweets, 2 as negative tweets and 9 as neutral tweets. The software calculated the sentiment with the efficiency of 55%. Figure 6 shows the analysis of the rishta likhenge hum naya Hindi serial tweets.

II. Also results have been shown in the pie charts. Each dataset has been a different pie chart with positive, negative and neutral tweets.

a. Pie chart of demonetization movie tweets: Figure 7 represents the pie chart of demonetization tweets. Here 43% tweets as neutral, 30% tweets as positive and 27% tweets as negative.

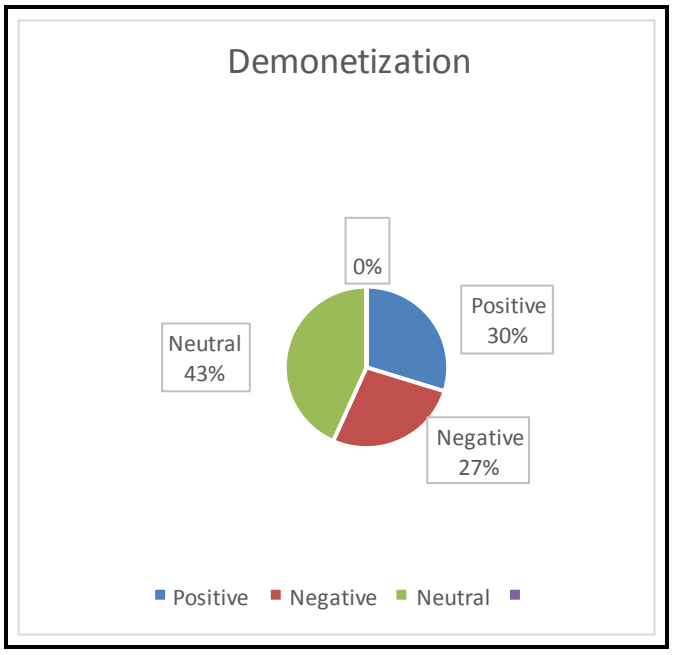

Figure 7. Pie chart of demonetization tweets

b. Pie chart of tiger zinda hai movie tweets: Figure 8 represents the pie chart of tiger zinda hai movie tweets. Here 33% tweets as neutral, 62% tweets as positive and 5% tweets as negative.

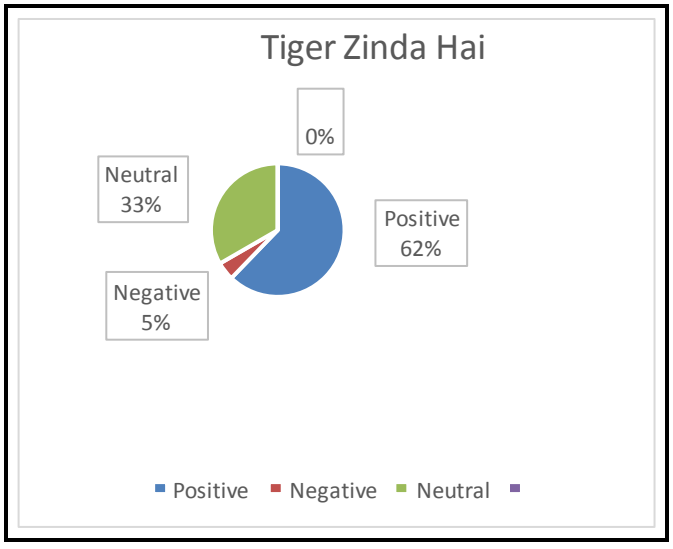

Figure 8. Pie chart of tiger zinda hai movie tweets

c. Pie chart of padmaavat movie tweets: Figure 9 represents the pie chart of padmaavat movie tweets. Here 43% tweets as neutral, 48% tweets as positive and 9% tweets as negative.

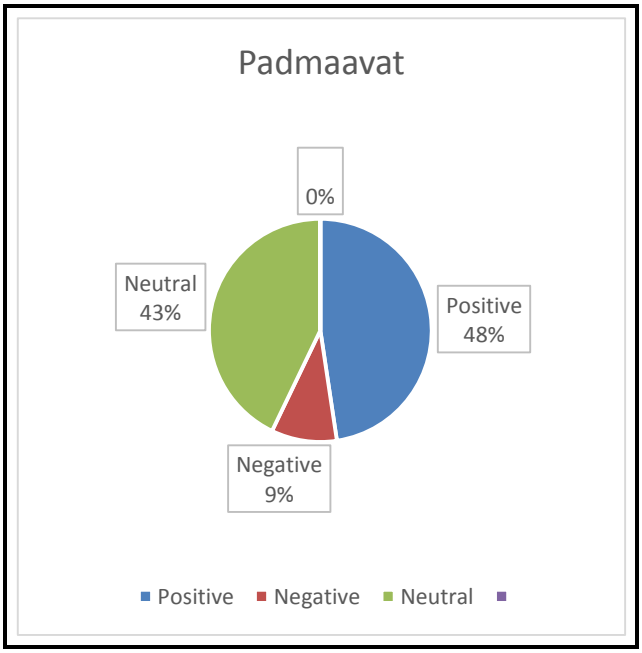

Figure 9. Pie chart of padmaavat movie tweets

d. Pie chart of rishta likhenge hum naya Hindi serial tweets: Figure 10 represents the pie chart of rishta likhenge hum naya Hindi serial tweets. Here 45% tweets as neutral, 45% tweets as positive and 10% tweets as negative.

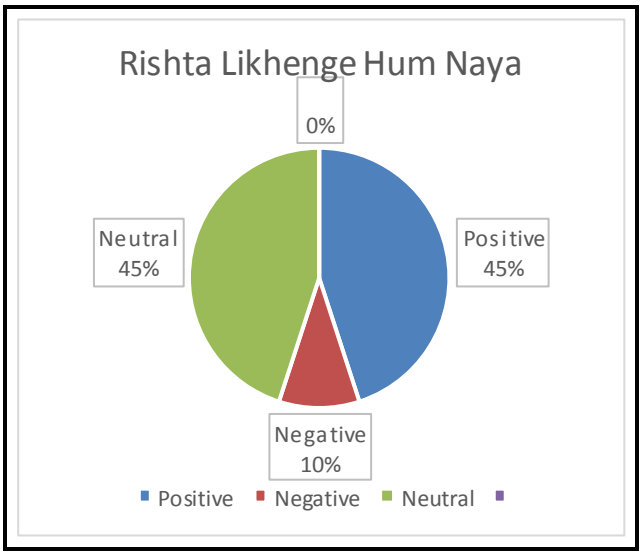

Figure 10. Pie chart of rishta likhenge hum naya Hindi serial tweets

III. Accuracy comparison of different datsets: Figure 11 represents the accuracy of demonetization, tiger zinda hai movie, padmaavat movie and rishta likhenge hum naya Hindi serial tweets.

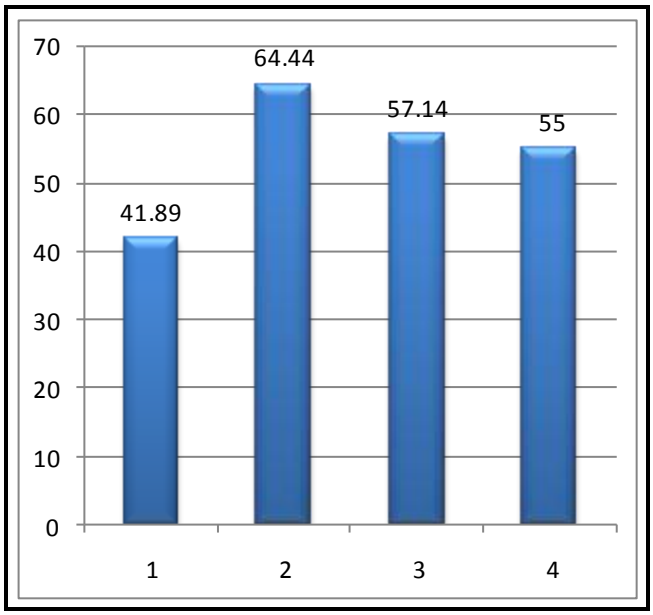

Figure 11. Accuracy comparison of different datasets

Table 1. Describing accuracy of different datasets

Approaches	Classification of tweets			
	Positive tweets	**Negative tweets**	**Neutral tweets**	**Accuracy in %**
KNN	17	77	19	67.78
SVM	22	77	14	67.78
Proposed model	69	26	65	86

IV. Graphical representation of results: Figure 12 represents the results of different datasets in graphical form.

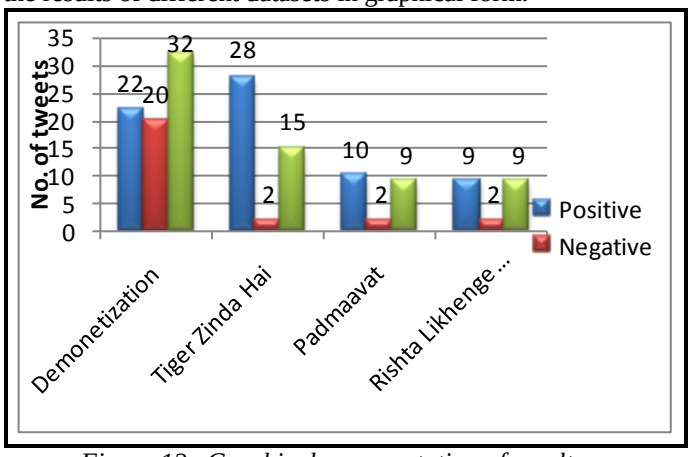

Figure 12. Graphical representation of results

V. CONCLUSION AND FUTURE SCOPE

Sentiment analysis is the emerging field that is mainly used in many application areas. Its scope is increasing. So a need arises to create or develop an algorithm that could properly find the sentiment of the public tweets or opinion. This work shows a new algorithm that is developed in the Java language. The algorithm is applied on tweets and efficiency is calculated based on the accuracy rate of the algorithm. The approximate efficiency of the algorithm is 86%. The accuracy of the algorithm is checking by taking the comments from other websites. Evaluation of two or more products or brands is also done for better performance. A rich lexicon dictionary is created for enhanced processing of the algorithm. Sentiment analysis is applied to further more datasets for better analysis. The work can be extended by collecting the tweets from different blogs and sites and apply different types of classifiers on the dataset and their accuracy can be compared to know which classifier is helpful for achieving better efficiency.

REFERENCES

[1] R. Mehta, D.S. Jain, "Sentiment Mining and Related Classifiers: A Review", IOSR Journal of Computer Engineering, Vol.18, Issue.1, pp.50-54, 2016.

[2] Chandni, N. Chndra, S. Gupta, R. Pahade, "Sentiment Analysis and its Challenges", International Journal of Engineering Research & Technology, Vol.4, Issue.3, pp.968-970, 2015.

[3] A.P. Rajan, S.P. Victor, "Web Sentiment Analysis for Scoring Positive or Negative Words Using Tweeter Data", International Journal of Computer Applications, Vol.96, Issue.6, pp.33-37, 2014.

[4] A. Gupta, J. Pruthi, N. Sahu, "Sentiment Analysis of Tweets Using Machine Learning Approach", International Journal of Computer Science and Mobile Computing, Vol.6, Issue.4, pp.444-458, 2017.

[5] B. Alvares, N. Thakur, S. Patil, D. Fernandes, K. Jain, "Sentiment Analysis Using Opinion Mining", International Journal of Engineering Research & Technology, Vol.5, Issue.4, pp.88-91, 2016.

[6] B.M. Bandgar, D.S. Sheeja, "Analysis of Real Time Social Tweets for Opinion Mining", International Journal of Applied Engineering Research, Vol.11, Issue.2, pp.1404-1407, 2016.

[7] B.S. Dattu, P. Deipali, V.Gore, "A Survey on Sentiment Analysis on Twitter Data Using Different Techniques", International Journal of Computer Science and Technologies, Vol.6, Issue.6, pp.5358-5362, 2015.

[8] D.E. Oleary, "Twitter Mining for Discovery, Prediction and Causality: Applications and Methodologies", International Journal of Intelligent Systems in Accounting and Finance Management, Vol.22, Issue.3, pp.222-247, 2015.

[9] G. Sabarmathi, D.R. Chinnaiyan, "Reliable Data Mining Tasks and Techniques for Industrial Applications", IAETSD Journal for Advanced Research in Applied Sciences, Vol.4, Issue.7, pp.138-142, 2017.

[10] H.P. Rahmath, "Opinion Mining and Sentiment Analysis-Challenges and Applications", International Journal of Application or Innovation in Engineering & Management, Vol.3, Issue.5, pp.401-403, 2014.

[11] I. Smeureanu, C. Bucur, "Applying Supervised Opinion Mining Techniques on Online User Reviews", Informatica Economică, Vol.16, Issue.2, pp.81-91, 2012.

[12] K.I. Umar, F. Chiroma, "Data Mining for Social Media Analysis: Using Twitter to Predict the 2016 US Presidential Election", International Journal of Scientific & Engineering Research, Vol.7, Issue.10, pp.1972-1980, 2016.

[13] K. Sutar, S. Kasab, S. Kindare, P. Dhule, "Sentiment Analysis: Opinion Mining of Positive, Negative or Neutral Twitter Data

Using Hadoop", International Journal of Computer Science and Network, Vol.5, Issue.1, pp. 177-180, 2016.
[14] L.J. Sheela, "A Review of Sentiment Analysis in Twitter Data Using Hadoop", International Journal of Database Theory and Application, Vol.9, Issue.1, pp.77-86, 2016.
[15] S.A.A. Hridoy, M.T. Ekram, M.S. Islam, F. Ahemed, R.M. Rahman, "Localized Twitter Opinion Mining Using Sentiment Analysis", Decision Analytics, Vol.2, Issue.1, pp.1-19, 2015.

Authors Profile

Seema Baghla is presently working as Assistant Professor in Computer Engineering at Yadavindra College of Engineering, Punjabi University Guru Kashi Campus, Talwandi Sabo (Distt Bathinda) Punjab w.e.f. August 2008. She has almost 14 years of teaching experience of teaching M.Tech. (CE), B.Tech. (CSE) and MCA Classes. She previously worked as Senior Lecturer & Head, Department of Computer Science & Engineering at BMSCE, Muktsar and Lecturer (CSE) at Government Polytechnic College, Bathinda. She completed her Bachelor of Technology (CSE) from Institute of Engineering & Technology, Bhaddal, Ropar, Punjab in year 2004 holding 9th merit position in the University. She completed Master of Technology in Computer Engineering in year 2007 from Punjabi University, Patiala, Punjab. She has guided almost 35 M.Tech. Dissertations and a number of B.Tech. Projects. She has more than 50 research publications in reputed International/ National Journals and Conferences. Her research areas include Data Mining, Big Data Management, Digital Image Processing, Optimization, networking etc.

Amritpal Kaur is presently working as Computer teacher at D.M.Group of Institutions Kararwala (Distt Bathinda). She has been working here since 2016. She completed her Bachelor of Technology (CSE) from GRDIET lehra bega Bathinda. She is pursuing M.tech (CE) Part time from Yadavindra College of Engineering Punjabi University Guru Kashi Campus. She has submitted her dissertation. Her research area include Data Mining.

