



iJRASET

International Journal For Research in
Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 Issue: VIII Month of publication: August 2017

DOI: <http://doi.org/10.22214/ijraset.2017.8029>

www.ijraset.com

Call: ☎ 08813907089

E-mail ID: ijraset@gmail.com

Citizen Opinion Mining Initiatives in Government Decision Making by Categorizing Citizen Sentiments

Shikha Parashar ¹, R.K. Gupta²

¹Department of Computer Science & Information Technology, Madhav Institute of Technology & Science

Abstract: Citizen Opinion Mining (C.O.M) is a type of natural language processing (NLP) for finding the sentiments of citizen on various targeted government policies. Opinion mining techniques can be used for measuring influence facts like citizen behavior, desire, needs that help to improve government services. It's includes building a system to gather and examine opinions for specific policy, made in weblog posts, comments, stories or tweets. Proposed approach gives brief description how citizens opinions are play important role to make strong decision for any government policy .In our proposed approach we are categorized all the citizen opinion in different categories such as based on their profession (student, professor, common man, businessman) etc. Because if government wants to make decision for specific group of citizen then that group of citizens opinion are most important. The aim of our proposed approach is to make strong decision for citizens.

Keywords: decision making, data mining, SVM algorithm with RBF kernal function

I. INTRODUCTION

Citizen Opinion Mining (C.O.M) is a promising discipline which is defined as combination of understanding problem, computation linguistic and information retrievals offers with the citizen opinion expressed in a record. The field essential objectives are fixing the issues concerning opinions about products, politics, policies in newsgroup posts, assessment web sites, and so on. There are unique techniques for summarizing citizen stories like Data Mining, Information Retrieval, Text Classification and Text Summarization, before WWW citizen asked for opinions of his family members and friends before purchase any product. In the very same way when any government organization need to take the decision for their new policies they had to conduct various surveys on the focused object and they had discuss on various issues with external consultants to take advice. Web2.0 provide convenience for the government decision makers to take decision for any policies by reviewing the citizen posted comments. Citizens can post reports on different web blogs, dialogue forums, twitters, blogs, product's site these feedback are referred to as person generated contents. Web2.0 is taking part in a relevant function in data extracting source in opinion mining. It facilitates government to know about the policies from other citizen's reviews who has strong view point for specific policy. An automated opinion summarization mannequin is required to whole these tasks. This technique is used to easily identify the positive, negative or neutral citizen sentiment summary from unstructured data. It involves text subjectivity and computational management of opinion. The below Diagram describes the process flow model of Opinion Mining [1].

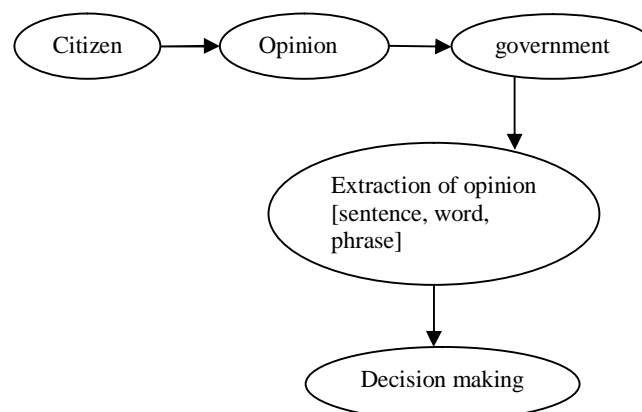


Fig.1. Opinion Mining Model

In above diagram, general process of citizen opinion extraction is given. Three most important add-ons of Opinion Mining are:

- A. *Opinion Holder*: Opinion holder is the citizen or any organization that expresses their opinion on any object.
- B. *Opinion Object*: It represents a feature on which the citizen is express their opinion.
- C. *Opinion Orientation*: In this we check whether the citizen sentiment about an object gives positive views, negative views or neutral views.

For example “This government does a great work”. In this review, Opinion Holder is the citizen who has written this review. Opinion object is defined “the work done by government” and the sentiment word is “great” which represent positive orientation of opinion. Determination of semantic orientation is an undertaking of finishing up whether a sentence or report has either positive or negative orientation [2].

II. DATA FOR CITIZEN OPINION MINING

Data from different sources are mentioned to extract opinins and ensure good recommendation for specific applications. Most common data sources are blogs and opinion review sites.

Many blogs have reviews on products, government policies, issues etc. Review sites can be a factor considered to make strong decisions by user before buying is to know comments by previous buyers.

Reviewer’s data in most sentiment classification approach are summarized from different websites like www.amazon.com, www.yelp.com, [www.CNET download.com](http://www.CNETdownload.com) and so on. Opinions which are summarized from different web sites are play very important role to make strong government decision. We are using twitter data (tweets) for opinion classification in our approach.

III. TWITTER DATA

Millions of public opinions are appearing daily in general web sites that furnish offerings for micro blogging such as Twitter1, Tumblr2, Facebook3. These opinions of citizen are help to make strong decision for any government policy. As increasingly citizen submit about merchandise and services they use, or express their political and devout views, micro blogging internet-websites become priceless sources of citizen opinion and sentiments.

Such knowledge may also be effectively used for advertising and marketing or social stories we use a dataset formed of accrued messages from twitter. Twitter contains large number of very short opinions created by the users of this micro blogging platform. Table 1 shows examples of typical posts from Twitter. For example, government may be interested in the following questions:

- A. What do people think about our policies or services?
- B. How positive (or negative) are people about our policies?
- C. What would people support our policies to be like?

Political parties may be interested to know if people support their program or not. Social organizations may ask people’s opinion on current debates.

All this information can be obtained from micro blogging services, as their users post everyday what they like/dislike, and their opinions on many aspects of their life. We use micro blogging and extra specifically twitter for the next explanations:

- D. Micro blogging systems are utilized by one-of-a-kind people to precise their opinion about one of a kind topics, hence it is a useful source of people’s opinions.
- E. Twitter involves a massive quantity of text posts and it grows every day. The amassed corpus will also be arbitrarily gigantic
- F. Twitter’s users are varies from common citizen to celebrities, enterprise representatives, business man, politicians, and even country presidents. Accordingly, it is easily to extract text posts of citizens from exclusive social and interests groups
- G. Twitter’s viewers are represented by means of customers from many countries. Even though customers from U.S. Are prevailing, it is easy to summarized data in distinctive languages.

We accumulated a corpus of 300000 text posts from Twitter evenly cut up automatically between three sets of texts:

- 1) Tweets containing positive sentiments, akin to happiness, entertainment or pleasure
- 2) Tweets containing negative sentiment, similar to unhappiness, anger or disagree
- 3) texts in that best state a reality or do not specific any emotions we participate in a linguistic evaluation of our corpus and we exhibit the way to build a sentiment classifier that makes use of the accumulated corpus as training data [4].

Funkeyr: @redeychicago I think narendra modi can visit might've scaled the the victory for U.S.A
Vcurve: I love that how Google celebrates all things like this: Google.co.jp honors Confucius Birthday – japan Probe
Mattfellows: Hai world, I don't like faulty hardware on control systems where politics prevents you from moving software to less faulty systems.
Brroookly: I like the sound quality of my mobile makes when I shake to shuffle it. Boo bee boo
MeganWilloughby: like a Disney buff, can you found out about the new Alice in Wonderland movie. Official trailer I like the Cheshire Cat.

Table 1. Examples of posts with expressed users opinions

IV. LITERATURE SURVEY

Lavanya T[2016] et al. It is proposed to plan an algorithm which removes conclusion targets and opinion words utilizing word arrangement display for online audits extricated from Twitter. An opinion target is represented as the topic about which users shows their opinions. An opinion words are describe as the words that are used to represent user's opinions point of the venture is to decide the contemplations of blog or audit author regarding some theme or the general relevant extremity of online surveys utilizing word arrangement display. The aim of this project is to design an algorithm that predict all kind of opinion words and opinion targets for analyzing the market status of a product by mining user reviews posted on internetworking site namely the Twitter [5].

Jumadi et al. [2016] This paper talks about an approach where a plugged stream of tweets from the Twitter microblogging website are preprocessed and ordered in light of their enthusiastic substance as positive, negative and immaterial; and examinations the processing of different classifying algorithms in view of their accuracy and review in such cases. Assist, the paper exemplifies the utilizations of this examination and its limitations [6].

MD. Azza F. Yatim et al. [2016] this method uses lexicon for identifying sentiment of particular object within related data sets. This paper solely concentrates on building the lexicon for the method. By focusing on Indonesian politic issues, we create a new corpus approach to build a contextual lexicon which uses news articles as corpora. We determine the initial or basic seed words and have it analyzed by domain experts for our experiment. On the basis of tests which we have done, we find that 51.79 % of the terms in our lexicon are relevant to our research domain. We use this finding to evaluate and improve our method as we continue the research to obtain more relevant result [7].

Shokoufeh Salem Minab[2015] et al. The reason for this paper is to demonstrate the past works online analysis of sentiment on Twitter. Social media like Twitter create space to explain the thoughts and opinions on various topics and different events, millions of users can share their ideas in this Micro blog. Therefore Twitter is changed over as a main source to investigation of data; settle on a choice and an examination of supposition. There is a mean in the maximum part of the writings, yet it is more essential to give techniques to acquiring appropriate guaging and improved utilization of data for anticipating supposition. Likewise twitter data process after the stream demonstrate. In this process, data were arrived base at rapid to destination in form of result, data mining algorithm should be capable to find user feeling in immediate time under limited space and time constraints [8].

O'nder C[2015] et al. In this study, our real objective is to determine positive or negative extremity of Turkish Twitter bolsters by utilizing content characterization techniques for sentiment analysis. Bag of Words and N-Gram approach are mainly used to separate the substance of content in highlight extraction phase. Distinctive closeness measurements are broke down to enhance the processing of the KNN classifier on both Reuters-8 and Turkish Twitter Feeds data. The Reuters-8 data used to identify impact of content dialect and length on classification comes about. The examinations are directed on six distinct blends of highlight extraction models and weighting techniques. Experimental results about demonstrate that IT-Sim gives better execution contrasted with other grouping measurements and Tf-Idf is the good weighting technique. The accuracy of the KNN classifier is relied on upon mix includes extraction display with various weighting techniques and the estimations of k parameter [9].

Malhar Anjaria[2014] et al. We likewise propose a hybrid approach of separating opinion utilizing direct and indirect components of Twitter opinions. We combined Principal Component of Analysis (PCA) with SVM trying to perform dimensionality decrease. This paper is proposed to describe two distinctive contextual analyses of completely unique social situations, Election of US Presidential 2012 and Karnataka Assembly Elections 2013. We close the conditions under which Twitter may fall flat or prevail with regards to

anticipating the decisions result. Exploratory outcomes exhibit that Support Vector Machines(SVM) beat every single other classifier with greatest effective expectation accuracy approx of 88% if there should be an occurrence of election of US Presidential held in November 2012 and maximum result of accuracy is approx of 58% in case of Karnataka State Assembly Elections held in May 2013 [10].

V. ARCHITECTURAL DIGRAM OF OUR APPROACH

Given diagram represent our work flow approach of citizen opinion mining.

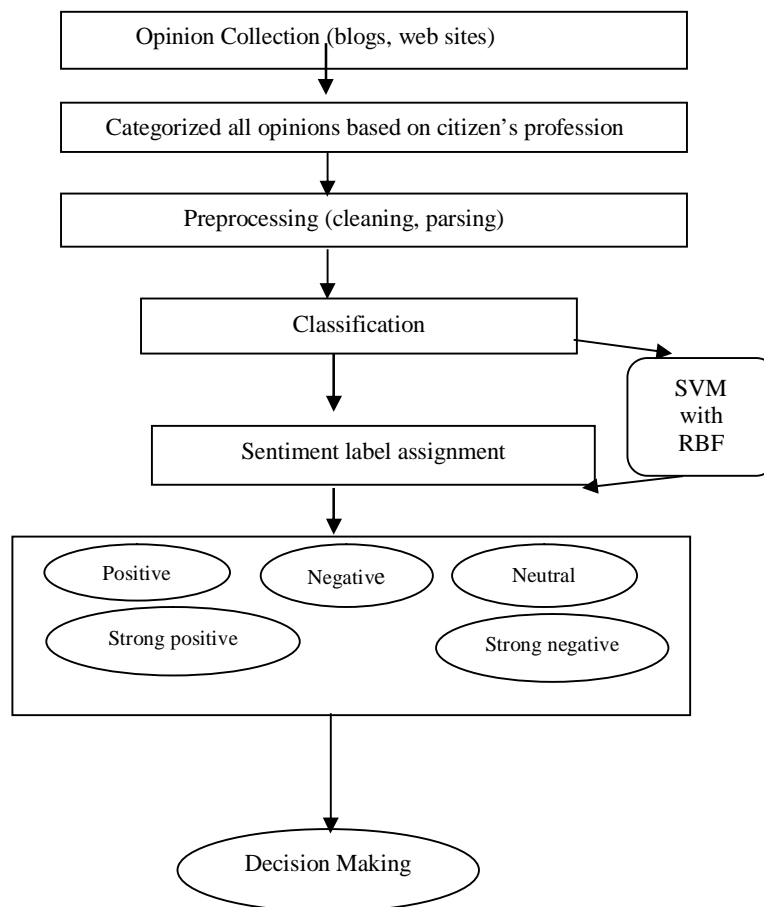


Fig.2.Work Flow Model

VI. PROPOSED APPROACH

In our proposed approach we are collect all the tweets from twitter or other micro blogging websites. All the tweets are contains citizen name, age and citizen profession.

Our data is on black money. We trained data with the format of

< Category, aspect, opinion word, sentiment, political interest government employee>

For example for the below tweets-

“narendra modi took a great decision on black money”

Category->decision, aspect->black money, opinion word->great, sentiment->positive, political interest->yes, government employee->yes

After collecting all tweets from different web sites our next step is categorized opinion in different citizen categories. We are categorized opinions based on citizen profession (student, professor, employee, common man, politician) etc. Reason behind perform categorization of opinion is- If government wants to make decision for specific group (student, professor, employee, common man, politician) than that group of citizen opinion are most important for strong decision making.

When we categorized all opinion our next step is preprocessing because information extracted from different web sites is in raw form that consist different irregularities in data. In order to use further processing, we need to clean and modify it in more usable

structured form of data. Tweet cleaning is the first step towards data transformation. After cleaning the data, the selected tweets are then parsed.

For example-words like 'a', 'the' might not be so significant. These words are known as Stop Words, and are removed while pre-processing. Also it doesn't matter whether it is 'great' or 'Great', so all the text can be brought to lower case. After cleaning the data, the selected tweets are then parsed. After data transformation we have structured data which is ready for further processing. After data transformation we get structured data which is ready for further processing.

After preprocessing of all the opinion we perform classification on each category. For classification we use support vector machine with neural network RBF function. SVM provide better result in terms of complexity and accuracy than the other classification algorithm like naïve byes algorithm. SVM is correctly work on linear Separable elements in which all data points are plotted in n-dimensional space or plane where n is the sum of all features and we have to select correct decision boundary that classify all data point in different classes. Major problem with SVM is it can't works in high dimensional feature space because large number of feature can't be classified linearly. To solve this problem we use RBF kernel function. It solves the problem that occurs in linear separable opinion classification. The reason behind using radial basis function is increase the accuracy of SVM algorithm. RBF is an artificial neural network function which is used for classification. SVM (RBF) function assign class label to each opinion.

In our approach sentiment label assignment is done in five categories. It classifies opinion in positive, negative, neutral, strong positive and strong negative. We are define all classes below-

- A. Positive-simple positive views {I love this decision very much}
- B. Negative- simple negative views {I don't agree with government decision}
- C. Neutral-not clear views {hop for best}
- D. Strong positive-positive view with positive reason { This move will force PeoPle to take this money to banks if they want to keep it, so this is right decision}
- E. Strong negative-negative view with valid reason {There are many who get their salaries in cash and do not have bank accounts so what they people do}

To calculate classification result we are using confusion matrix because this matrix is used to show accuracy of classification result. It is square matrix with equal row and column and this matrix is draw between predicted class and actual class. With the help of confusion matrix we will compare accuracy of each group that represents the specific group of citizen sentiments.

	A	B	C	D	E
A	TP _A	E _{AB}	E _{AC}	E _{AD}	E _{AE}
B	E _{BA}	TP _B	E _{BC}	E _{BD}	E _{BE}
C	E _{CA}	E _{CB}	TP _C	E _{CD}	E _{CE}
D	E _{DA}	E _{DB}	E _{DC}	TP _D	E _{DE}
E	E _{EA}	E _{EB}	E _{EC}	E _{ED}	TP _E

Table2.confusion matrix representation for five classes

Accuracy= sum of all diagonal elements/total elements in matrix

VII. RESULT ANALYSIS

In previous opinion classification method all opinion are classified by applying support vector machine but in our approach we are using SVM with RBF kernel function to improve the result of text classification.

The major drawback has overcome, SVM performs better on limited opinions because we already categorized all citizen sentiments in different categories and we are applying SVM independently on each category. For analysis we had collected citizen reviews and tweets on black money or demonetization decision. Here we will compare classification results of previous approach and our proposed approach.

A. Base approach results

Given below results show the result of previous classification (SVM) method which we are implement on Matlab 3.2 platform. In previous approach, classification is apply on all citizen opinions without categorizing them. Fig.3 shows confusion matrix of opinion classification.

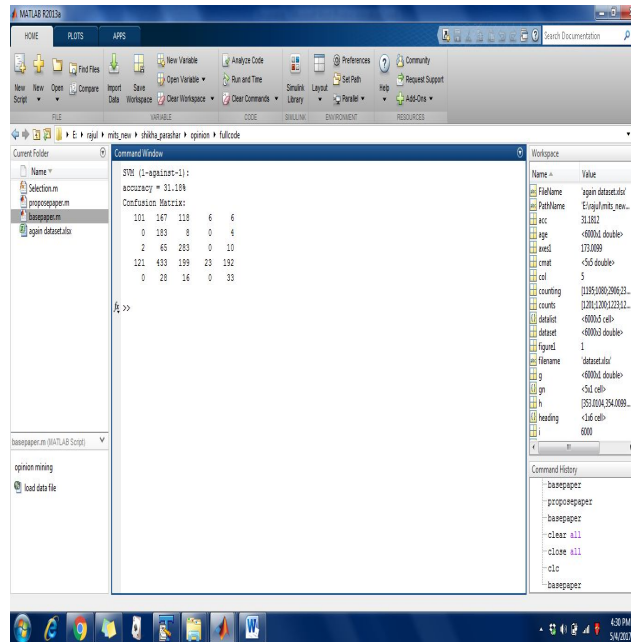


Fig.3. Resultant confusion matrix by simple SVM

Resultant matrix SVM (1-against-1)

101	167	118	6	6
0	183	8	0	4
2	65	283	0	10
121	433	199	23	192
0	28	16	0	33

Table3. Opinion classification

Accuracy = 31.18% are representing above confusion matrix in table format. This matrix shows all citizen opinion classification result of our previous approach and it is used to calculate the accuracy of support vector machine.

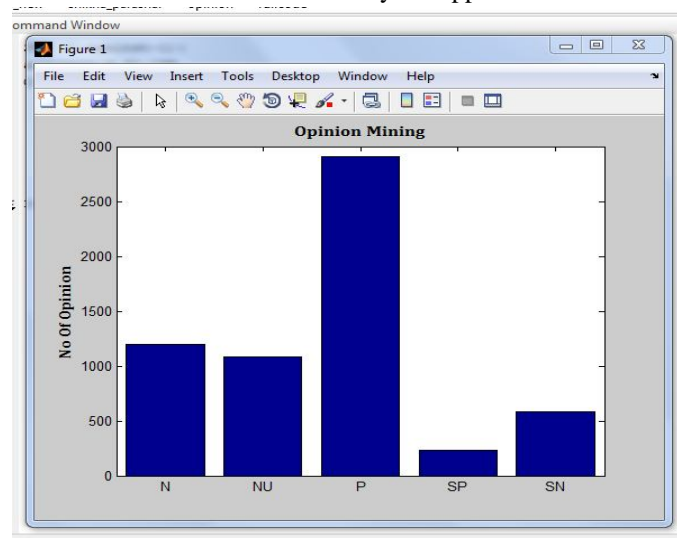


Fig.4.Graph on siple opinion mining using SVM

The above graph shows the opinion classification of 3000 citizen tweets in positive, negative, strong positive, strong negative and neutral opinion. This graph generated by simple support vector machine for opinion classification.

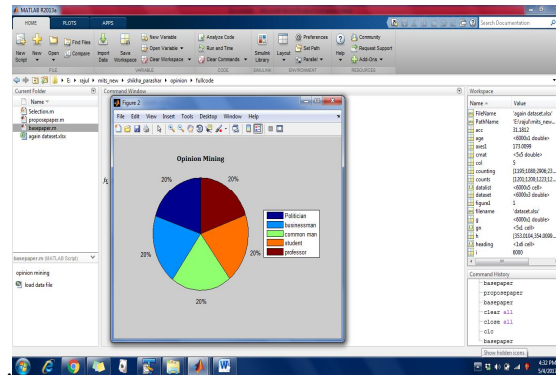


Fig.5. Pie graph representation of citizen categories

The above pie graph shows total number of public opinion that we had taken in our data set for classification and represents five different citizen categories.

B. Proposed approach results

When we categorized all the opinion in different category then each group of citizen give our own result. Here we can clearly see difference between two citizen categories businessman and common man on the same government decision. Accuracy of group businessman is 24% while accuracy of another group common man is 53.81%.

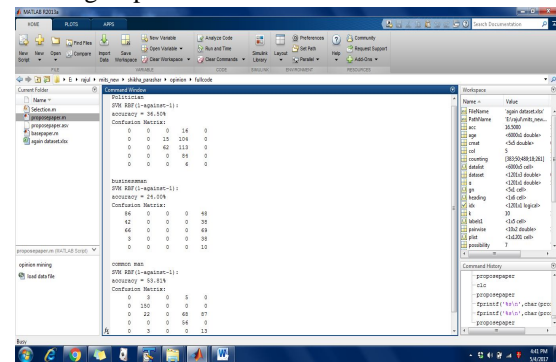


Fig.6. confusion matrix of each group of citizen

Above screenshot represents the resultant matrix of different citizen categories like politician, common man, business man. We are representing these matrix in table form that given below [table4, 5]. Similarly this matrix can also apply on other citizen categories to calculate accuracy of classification.

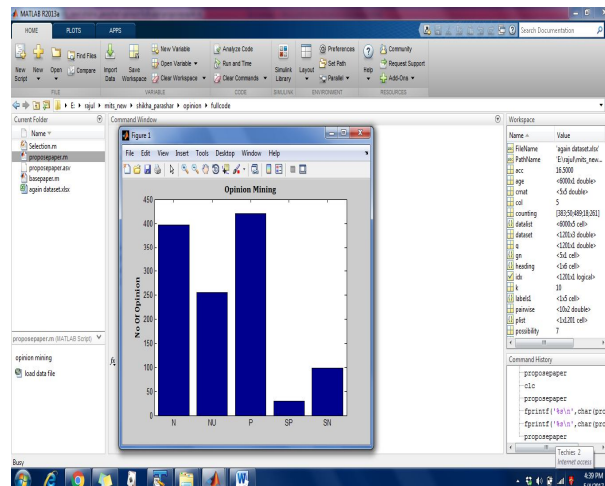


Fig.7. Graphical representation of businessman opinions

Fig.7.Represents a specific group (businessman) opinion in five different opinion classes (positive, negative, strong positive, strong negative and neutral). In this group ratio of positive and negative opinion is almost equal but neutral opinion is also high so resultant accuracy is low.

Businessman opinion- SVM RBF (1-against-1)

86	0	0	0	48
42	0	0	0	38
66	0	0	0	69
3	0	0	0	38
0	0	0	0	10

Table4.opinion classification of businessman

Accuracy = 24.00% The above matrix is used to calculate the result of group businessman opinion. Each row and column shows different classes (positive, negative, neutral, strong positive and strong negative opinion.) of group businessman sentiments.

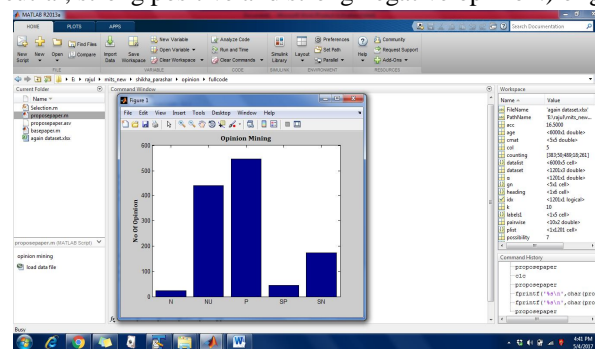


Fig.8. Graphical representation of common man opinions

Fig.8.Represents another group (common man) of opinion in five different opinion classes (positive, negative, strong positive, strong negative and neutral). In this group positive opinion is very high and number of strong positive opinion is also more than group of business man. So the resultant accuracy is comparatively very high.

Common man opinion-SVM RBF (1-against-1)

0	3	0	5	0
0	150	0	0	0
0	22	0	68	87
0	0	0	56	0
0	3	0	0	13

Table5.opinion classification of common man

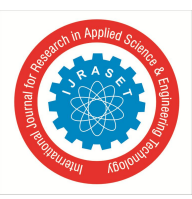
Accuracy = 53.81% The above matrix shows the result of group common man opinion. Each row and column shows different classes of group common man sentiments and all values shows number of positive, negative, neutral, strong positive and strong negative opinion.

VIII. CONCLUSION

In this proposed approach data analysis approach is presented in which we extract group of opinions from different micro blogging web sites like twitter, political blogs that are used to make decision for any object. SVM apply with RBF function to increase result accuracy than previous opinion classification algorithms. In our presented approach each group of citizen shows different opinion on any government decision or policies. Because it is important that if government wants to make any decision for specific group of citizen then that group of citizens opinion are very important for decision making. Aim of presented approach is to make strong decision for development of citizen and government empowerment.

REFERENCES

- [1] Nidhi R. Sharma , Prof. Vidya D. Chitre "Opinion Mining, Analysis and its Challenges" International Journal of Innovations & Advancement in Computer Science IJIACS, Volume 3, Issue 1 April 2014.



- [2] Richa Sharma, Shweta Nigam and Rekha Jain “Supervised Opinion Mining Techniques: A Survey” International Journal of Information and Computation Technology. ISSN 0974-2239 Volume 3, Number 8 (2013).
- [3] Jeevanandam Jotheeswaran, Dr. S. Koteeswaran Sentiment Analysis: “A Survey of Current Research and Techniques” International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 3, Issue 5, May 2015.
- [4] Alexander Pak, Patrick Paroubek “Twitter as a Corpus for Sentiment Analysis and Opinion Mining”2009.
- [5] Lavanya. T, Miraclin Joyce Pamila. J. C, Veningston. K “Online Review Analytics using Word Alignment Model on Twitter Data”2016 IEEE.
- [6] BalakrishnanGokulakrishnan, Pavalanathan Priyanthan, ThiruchittampalamRagavan, Nadarajah Prasath, AShehan Perera Opinion Mining and Sentiment Analysis on Twitter Data Stream “”20 12 IEEE.
- [7] MD. Azza F. Yatim, Yulistiyan Wardhana, Ahmad Kamal “A Corpus-Based Lexicon Building in Indonesian Political Context Through Indonesian Online News Media” 978-1-5090-4629-4/16/\$31.00@ 2016 IEEE.
- [8] Shokoufeh Salem Minab, Mehrdad Jalali and Mohammad Hossein Moattar “Online Analysis of Sentiment on Twitter” 2015 IEEE.
- [9] O’ nder C, oban, Baris, O’ zyer and Gu’ls, ah Tu’mu’klu’ O’ zyer “A Comparison of Similarity Metrics for Sentiment Analysis on Turkish Twitter Feeds” 2015 IEEE.
- [10] Malhar Anjaria, Ram Mahana Reddy Guddeti “Influence Factor Based Opinion Mining of Twitter Data Using Supervised Learning” 2014 IEEE.



10.22214/IJRASET



45.98



IMPACT FACTOR:
7.129



IMPACT FACTOR:
7.429



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24*7 Support on Whatsapp)