iJRASET
International Journal For Research in
Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 5 **Issue:** V **Month of publication:** May 2017

**DOI:**

Volume 5 Issue V, May 2017
ISSN: 2321-9653

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

# Analysis of Social Data Opinion through Public User Raw Information

Shilna K S[1], T Sivakumar[2], Miji K Raju[3]
[1]PG Scholar, [2]M.E. (CSE), Professor, [3] PG Scholar, Maharaja Institute of Technology, Coimbatore
Computer Science and Engineering, Anna University, Coimbatore

***Abstract: In this paper, we move one step further to interpret sentiment variations. To the best of our knowledge, our study is the proposed work that tries to analyze and interpret the public sentiment variations in micro blogging services. We propose a sentimental data analysis model using Neural Networks. Positive, Negative and moderate feed backs will be calculated here. We observed that emerging topics (named foreground topics) within the sentiment variation periods are highly related to the genuine reasons behind the variations. These foreground topics can give potential interpretations of the sentiment variations.***
***Keywords: Sentiment variations, Generative models, Micro blogging services, neural networks.***

## I. INTRODUCTION

A social network is a public structure build up of a group of public leads (such as individuals or organizations) and a group of the lively ties among these leads. The social network viewpoint contributes a group of methods for analyzing the structure of whole public entities as well as a mixture of theories illuminating the patterns observed in these structures. The study about these structures consists of social network examination to discover limited and universal patterns, trace significant entities, and inspect system dynamics. Thousands and millions of users share their opinions as well as suggestions on Social Networks, making it a precious stage for tracking and analyzing public sentiment. Such tracking and analysis can offer vital information for decision making in different domains. Therefore it has concerned interest in both academic world and industry. Earlier study largely focused on modelling and tracking public sentiment. Data mining involves mining of data and information related to software engineering, draw out some knowledge from it and make use of this knowledge to develop the software engineering process. Software engineering consists of many tasks from specification, design, development, monitoring at runtime, etc. Each such task is itself divided into a large number of sub tasks. For example, a developer frequently switches between different tasks, such as operating code, scanning documentation, composing code, debugging, etc. The major inspiration of our study is Sensitivity analysis, which is the review of how the ambiguity in the outcome of an arithmetical model or scheme (numerical or otherwise) can be apportioned to different sources of ambiguity in its inputs. A similar practice is uncertainty or ambiguity analysis, which has a larger focus on ambiguity assessment and circulation of ambiguity or uncertainty. If possible, ambiguity and sensitivity analysis have to be run in team. This area of research in Sentiment Analysis will describe mechanical tools which are competent to mine skewed information from content or texts in normal speech, such as suggestions, feelings, opinions and sentiments, so as to create structured and actionable knowledge to be used by either a decision support system or a decision maker. The discussion also goes through the logical and intellectual practices underling societal system communications. The research explores both semantic and machine learning models and methods that address context-dependent and dynamic text in social networks, showing how social network streams pose numerous challenges due to their large-scale, short, noisy, context- dependent and dynamic nature.

*A. System Details*

1) *Existing System:* The existing system of the prediction is chart. Here the charts will be in the usual design to understand the data. Classification is involved with allotting objects to classes on the basis of dimensions made on these objects. There are two main aspects to classification: discrimination and clustering, or supervised and unsupervised learning. In unsupervised learning which includes cluster analysis, class discovery and unsupervised pattern recognition, the classes are indefinite earlier and require to be recognized from the data. On the contrary, in supervised learning which mainly focuses on discriminate analysis, class prediction, and supervised pattern recognition, the classes are predefined and the mission is to recognize the basis for the classification from a set of labelled objects. To organize upcoming observations this information can be used. The current editorial focuses on cluster analysis (unsupervised learning) , but draws on ideas from supervised learning to deal with the

crisis.

In cluster analysis, the data are supposed to be sampled from a combination distribution with K components related to the K clusters to be recovered. Let (X1, ..., Xp) indicate a arbitrary 1 ×p vector of descriptive variables or type, and let Y∈ {1, ..., K} indicate the indefinite component or cluster label. Specified a sample of X values, the aim is to assess the number of clusters K and to approximate, for each observation, its cluster label Y. Suppose we have data X = (xij) on p descriptive variables (for example, genes) for n observations (for example, tumor mRNA samples), where xij denotes the realization of variable Xj for observation i and xi = (xi1,...,.xip) denotes the data vector for observation i, i = 1,...,n, j = 1,...,p. We consider clustering procedures that partition the learning set $\mathcal{L}$ = {x1,...,xn} into K clusters of observations that are 'related' to each other, where K is a user prespecified integer. More specifically, the clustering $\mathcal{P}(\cdot;\mathcal{L})$ assigns class labels $\mathcal{P}$ (Xi;$\mathcal{L}$) $= \hat{y}_i$ to each observation, where $\hat{y}_i \in \{1,\ldots,K\}$. Clustering procedures generally operate on a matrix of pair wise dissimilarities (or similarities) between the observations to be clustered, such as the Euclidean or Manhattan distance matrices. A partitioning of the learning set can be produced directly by partitioning clustering methods (for example, k-means, partitioning around medoid (PAM), self-organizing maps (SOM)) or by hierarchical clustering methods, by 'cutting' the dendrogram to obtain K 'branches' or clusters. Important issues, which will only be addressed briefly in this article, include: the selection of observational units, the selection of variables for defining the groupings, the transformation and standardization of variables, the choice of a similarity or dissimilarity measure, and the choice of a clustering method. Our foremost apprehension here is to assess the number of clusters K.

When a clustering algorithm is applied to a set of observations, a partition of the data is returned whether or not the data show a true clustering structure, that is, whether or not K = This reality causes no inconvenience if clustering is made to attain a realistic combination of the specified group of objects, as for directorial or revelation purposes (for instance, hierarchical clustering for presenting huge gene-expression data matrices. However, if interest lies primarily in the recognition of an unknown classification of the data, an artificial clustering is not satisfactory, and clusters resulting from the algorithm must be investigated for their relevance and reproducibility. By explanatory and pictorial investigative practices or by relying on probabilistic models and appropriate arithmetical importance tests this tasks can be carried out.

We dispute here that authenticating the outcomes of a clustering process can be done successfully by concentrating on prediction correctness. Once new classes are identified and class labels are assigned to the observations, the next step is often to build a classifier for predicting the class of future observations. The reproducibility or predictability of cluster assignments becomes very important in this context, and therefore provides a motivation for using ideas from supervised learning in an unsupervised setting. Resembling methods such as bagging and boosting have been applied successfully in the field of supervised learning to improve prediction accuracy. Here we recommend a unique resampling technique, Clest, which blend notions from discriminator and cluster analysis for assessing the number of clusters in a dataset. Although the proposed re-sampling methods are applicable to general clustering problems and procedures, particular attention is given to the clustering of tumors on the basis of gene-expression data using the partitioning around methods (PAM) procedure

2) *Existing Algorithm:* Mapping Function**:** The main strategy of mapping the words and documents to the space is to first compute the word embeddings and then derive the document embeddings based on the word embeddings by considering the word occurrences. Linear projection is believed to convert the novel trait representation of words to their piercing arrangement. In particular, adk projection matrix PA published to map words in domain A to a k-dimensional embedding space R k, while a dh projection matrix PB issued to map words in domain B to the same embedding space. Given in total M+MA words in domain A including the M pivots appearing in both domains and MA non-pivot words only appearing in domain A, we let n e z I oM+MA. Denote their corresponding word embeddings stored in an (M+MA)k embedding matrix e ZA computed by the linear projection mapping given as

$$\widetilde{\mathbf{Z}}_A^T = \left[\mathbf{P}_A^T\mathbf{U}_A^T,\ \mathbf{P}_A^T\mathbf{A}^T\right].$$

Similarly, ne z(B )ioM+MBi=1denotes the embeddings forward s in domain B, which results in an(M+MB)k embedding matrix e ZB computed by

$$\widetilde{\mathbf{Z}}_B^T = \left[\mathbf{P}_B^T\mathbf{U}_B^T,\ \mathbf{P}_B^T\mathbf{B}^T\right].$$

Volume 5 Issue V, May 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

3) *Proposed System:* Users of decision support systems often see data in the form of data cubes. This data cube is used to symbolize data along some measure of interest. However a "cube" can be 2-dimensional, 3-dimensional, or higher-dimensional. Each dimension stands for certain attribute in the database. The cells in the data cube stand for the measure of interest. For instance, they might hold a count for the number of times that attribute arrangement appears in the database, or the least amount, utmost, total or middling value of some attribute. Queries can be functioned on the cube to retrieve decision support information.

In case a database that contains transaction information relating company sales of a part to a customer at a store location. A data cube built from m attributes can be stored as an m-dimensional array. Each element of the array contains the measure value, such as count. The array itself can be represented as a 1-dimensional array. For instance, a 2-dimensional array of dimension x x y can be stored as a 1-dimensional array of size x*y, where element (i,j) in the 2-D array is stored in position (y*i+j) in the 1-D array. The drawback of storing the cube openly as an array is that the majority data cubes are thin, so the array will enclose numerous void elements (zero values). Turn up or narration of the data cube can be made by transiting upwards throughout a model hierarchy. A model hierarchy maps a group of low level concepts to higher level, extra universal concepts. It can be used to sum up information in the data cube. As the values are pooled, cardinalities reduce in size and the cube gets minor. Universalizing can be considered of as estimating some of the abstract total cells that hold ANYs, and accumulating those in help of the inventive original cells.

4) *Proposed Algorithm Implementation:* Declaration

DS = data set
V = Vocabulary (Extracted from dataset)
C = Categorization :sen count
N = Occurrence
A(Pj | Ni ) = Array Deceleration: Pj denotes positive and Ni denoted negative
R = Result

*a)* Let start the process : class index : System.Web.UI.Page, Open database access to receive the dataset: OleDbConnection con;
*b)* Declare Array : ArrayListreccomment = new ArrayList(); Received data will be stored in an array;
*c)* String.IsNullOrEmpty (GridView1.Rows[i].Cells[j].ToString())) Confirming the data grid is empty to receive the data;
*d)* Now LET dataset will be DS;
*e)* Updating positive word : posi.Add((string)red[0].ToString()); Positive word will be stored as the separate string collection : cmd1.ExecuteReader();
*f)* Updating Negative word : nega.Add((string)red[0].ToString()); Negative word will be stored as the separate string collection : cmd2.ExecuteReader();
*g)* Grid updating: GridView2.Rows[i].Cells[j].Text == posi and nega;
*h)* Merging as a neural Networks = A(Pj | Ni ) array deceleration for Pj and Ni;
*i)* OleDbDataReader rd1 = cmd1.ExecuteReader();
*j)* postedby = rd[0].ToString();
pdate = rd[1].ToString();
ptime = rd[2].ToString();
shareto = rd[3].ToString();
post = rd[4].ToString();
memname = rd[5].ToString();
comment = rd[6].ToString();
memdate = rd[7].ToString();
*k)* Display the DS grid value in the data reader : GridView4.Visible = false; cmd1 = new OleDbCommand(query, con);
*l)* Data analysis process: Let V = Vocabulary: toterr.Add(rd1[0].ToString()): Spliting word as vocabulary
*m)* Also Split word will be stored as the C = Categorization : string[] a = toterr [i1] .ToString () .To Upper().Split(' ');
*n)* Now Analysis will produce the result: Compare Srting[DS,V,C] with A(Pj | Ni )
*o)* loadarr();
pcount = posi.Count;

Volume 5 Issue V, May 2017
ISSN: 2321-9653


ncount = nega.Count;
scount = sen.Count;

*p)* (sen[i].ToString() == posi[j].ToString());
*q)* (sen[i].ToString() == nega[j].ToString());
*r)* Display Result R : input sen = r count
*s)* rcount = rank.Count; for ranking: (sen[i].ToString() == rank[j].ToString())
stat = rd[8].ToString();
*t)* Display ranking result : string[] a = toterr[i1].ToString().ToUpper().Split(' ');
*u)* Retrieve data from the DB
*v)* Display results in charts

## II. RESULT AND DISCUSSION

Initially all the data will be considered as the input data and processing data. But as per proposed method we need to preprocess the data for a fine tuned result.

### *A. Compares the Result with the Existing System*

Even though the existing system deals with huge data, the obtaining results are very less. No different types of results are produced in the existing system. According to the sentimental analysis, various sentiments are available like

*1)* Positive
*2)* Negative
*3)* Moderate

In the existing system, the system categorizes only the major positive and negative categories only. It did not look up in depth for the data analysis. The major categories will not give a refined result and clarity information from the data set. Also in the existing system major techniques are not implemented.

As per the existing system:

TABLE I
CATEGORIES IN THE EXISTING SYSTEM

| Positive Category | Negative Category |
|---|---|
| +This product is so good | -I hate that product after delivery |
| +I am Happy to get such a good product | -I won't buy this product any more |

While comparing with the existing system, the proposed system has been improved a lot by adding more techniques. The techniques are discussed in the abstract and objectives.

*4) Findings*

TABLE II
FINDINGS BASED ON COMPARISON

| Data set | Positive | Negative | Moderate |
|---|---|---|---|
| This product is more excellent and good | Yes | No | No |
| I am more disappointed in this product | No | Yes | No |
| I am more disappointed in this product. I won't buy this anymore | No | Yes | No |
| I am ok with this product | Yes | No | ok |
| This is not worthy for 1000 rupees | No | Not | No |
| After buying the mobile phone, there are many scratches in the panel. | No | No | No sentimental words found. Scratches are the common word. |

www.ijraset.com Volume 5 Issue V, May 2017
IC Value: 45.98 ISSN: 2321-9653
**International Journal for Research in Applied Science & Engineering Technology (IJRASET)**

The below mentioned graph shows the unprocessed data,

| | |
|---|---|
| Total Number of Data | 199 |
| Number of Positive Data | 100 |
| Number of Negative Data | 39 |
| Number of Moderate Data | 60 |

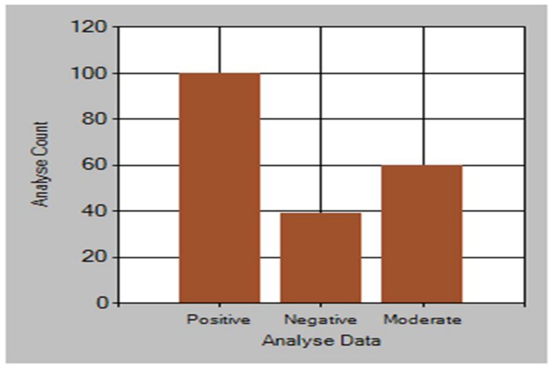



Fig.1 Before Preprocessing

| | |
|---|---|
| Total Number of Data | 140 |
| Number of Positive Data | 59 |
| Number of Negative Data | 31 |
| Number of Moderate Data | 50 |

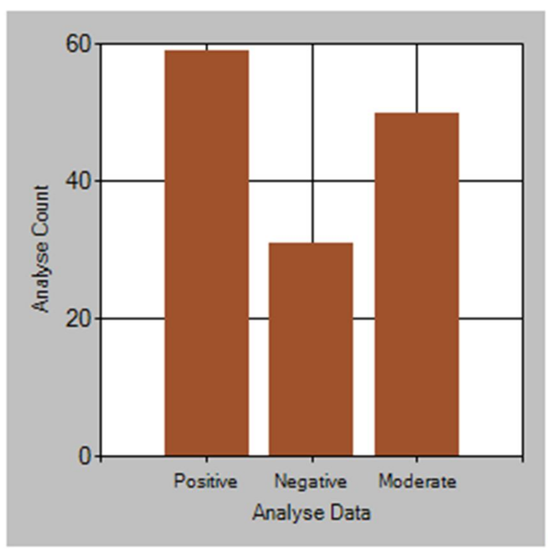



Fig.2 After data analysis

The above mentioned figure shows the actual result after the preprocessing. The preprocessing has been done using clustering and classification methods. It removes all the repeated users and repeated comments. After the removal of data, the analysis will we executed again in the same execution process.

Volume 5 Issue V, May 2017
ISSN: 2321-9653
**International Journal for Research in Applied Science & Engineering Technology (IJRASET)**

*5) Comparison Method*

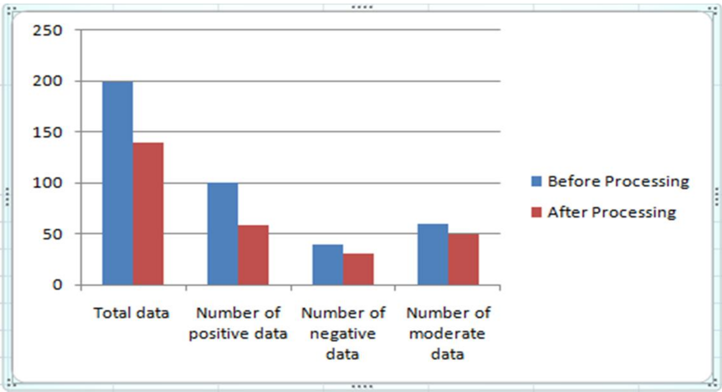



Fig.3 Comparison Chart

*6) Performance and Time*

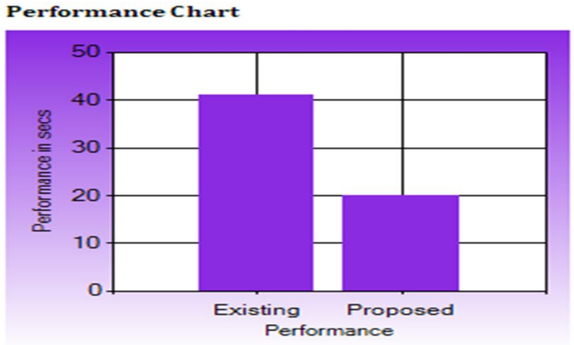



TABLE III
PERFORMANCE IN SECONDS

| Performance | In seconds |
|---|---|
| Existing System | 41 |
| Proposed system | 20 |

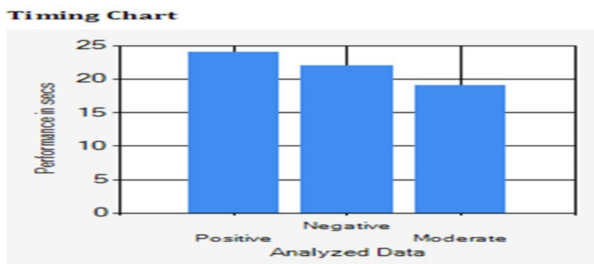



TABLE IV
TIMING IN SECONDS

| Data Processing | In seconds |
|---|---|
| Positive | 24 |
| Negative | 23 |
| Moderate | 18 |

Volume 5 Issue V, May 2017
ISSN: 2321-9653

**International Journal for Research in Applied Science & Engineering Technology (IJRASET)**

Hence the time taken for the execution and also performance has been formulated in the above mentioned charts. The proposed system takes less number of times. This method will calculates the query execution time and result gathering time between two processes. The timing process will calculates all the sentimental process like, positive performance, Negative performance and moderate performance. Each and every process will be calculated according to the query execution process. The mentioned chart is the time taken for three different sentiments.

## III. PRACTICAL RESULTS

The practical working to verify the sentiments in social data is successful and the screenshots are given below:

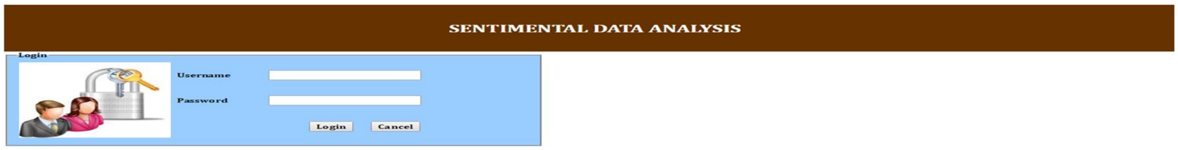



Fig.4 Login Page

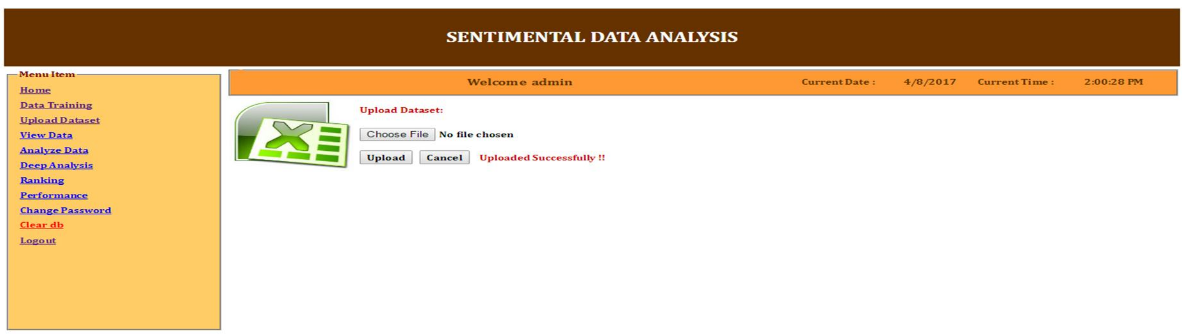



Fig.5 Data Uploaded Page

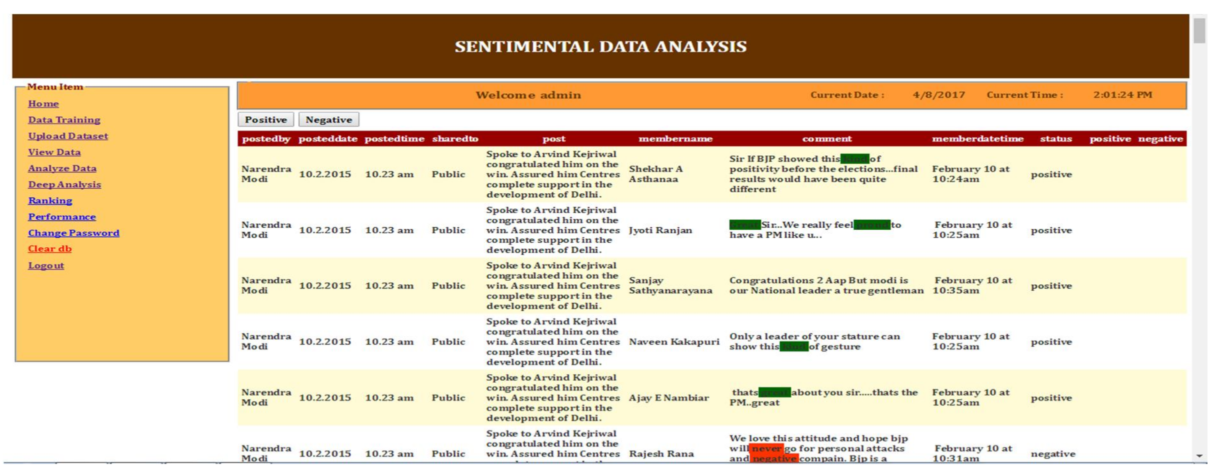



Fig.6 Deep Analysis Page

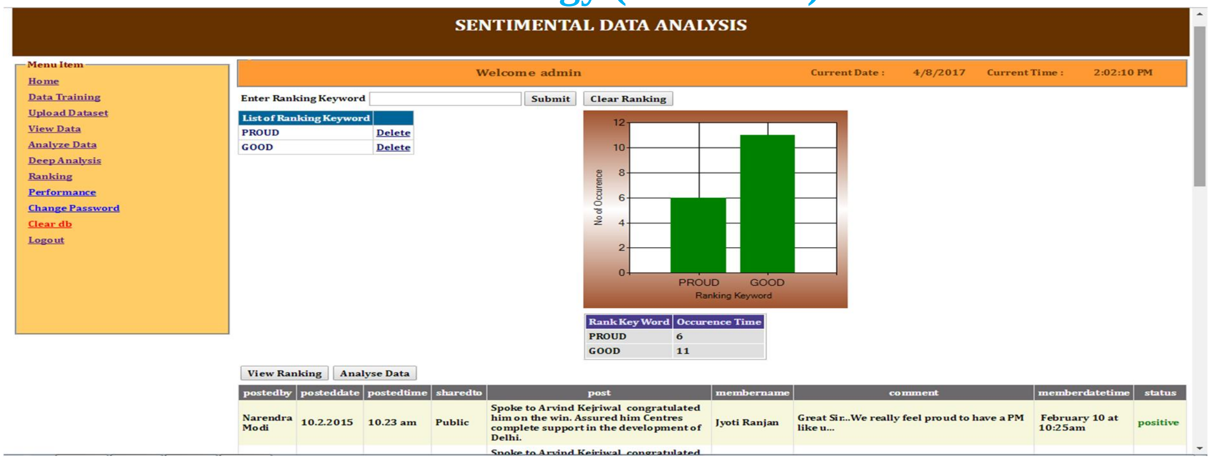



Fig.7 View Ranking Page

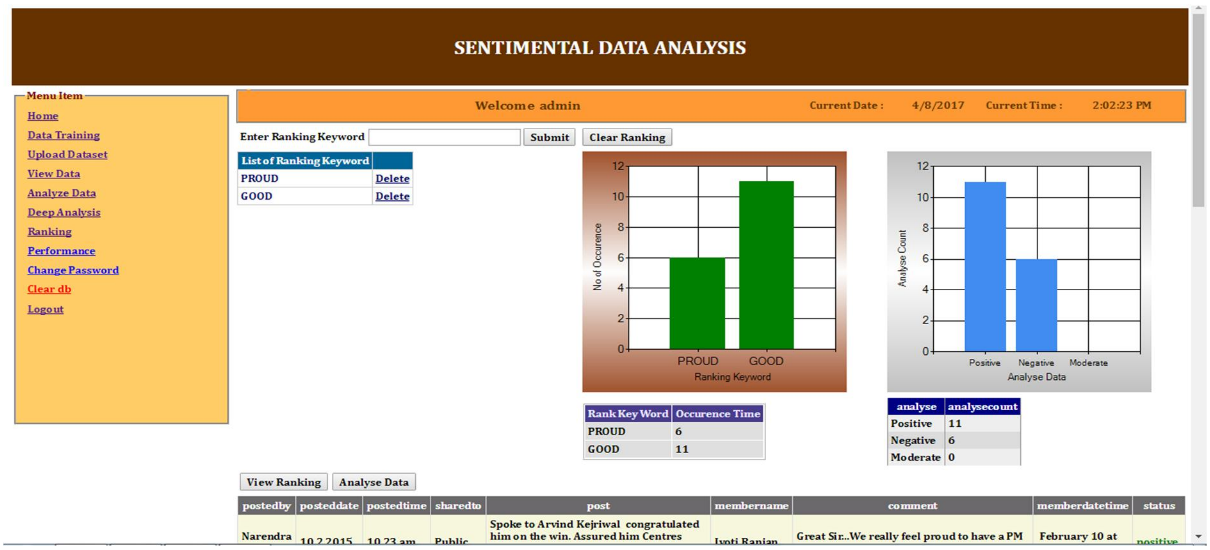



Fig.8 Analyzing Count Page

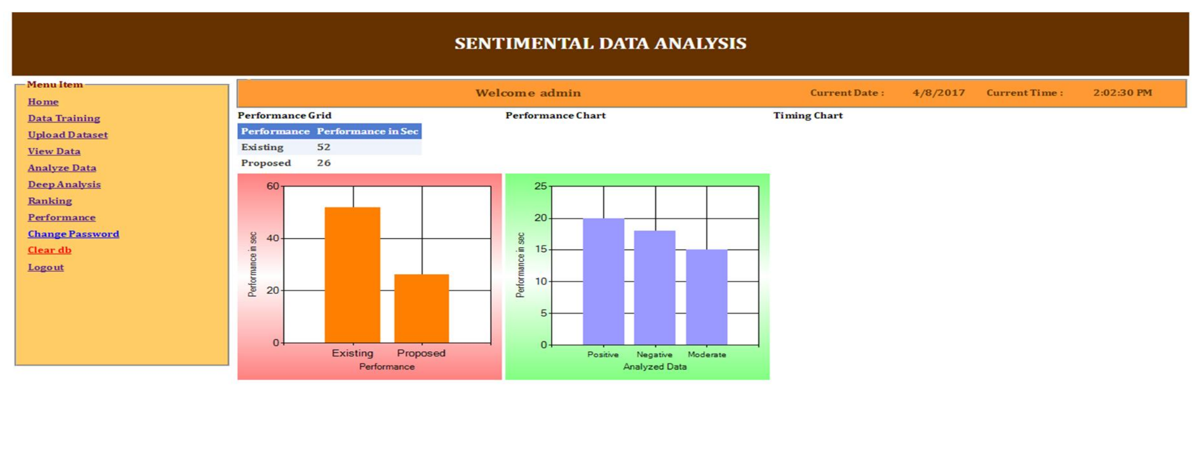



Fig.9 Performance and Timing Page

## IV. CONCLUSIONS

Inductive learning algorithms have been recommended as alternatives to knowledge attainment for expert systems. However, the relevance of machine learning algorithms frequently includes a lot of supplementary tasks to be implemented as well as algorithm

Volume 5 Issue V, May 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

execution itself. It is important to help the domain expert manipulate his or her data so they are suitable for a specific algorithm, and subsequently to assess the algorithm results. These activities are often called pre-processing and post processing. The future enhancement discusses issues related to the application of the text categorization algorithm, an important representative of the inductive learning family. A prototype workbench which has been developed to provide an integrated approach to the application of text categorization is presented. The architecture explanation and the possible application of the system are warranted. Finally, future directions and further enhancements of the workbench are discussed.

*A.* Can implement for web based application
*B.* Handshakes with inductive learning algorithm
*C.* Improvisations can be done in the performance evaluation
*D.* Prediction can be done for all kind of diseases
*E.* In case of huge range of data set, data load balancing can be done

## V. FUTURE WORK

Classification is a crucial process to manage data, fetch information suitably and suddenly. Text categorization algorithm depends entirely on the accuracy of the training data set for building its decision trees. The text categorization algorithm learns by supervision. Due to this text categorization algorithm, it cannot be successfully classify documents in the web. The data in the web is unpredictable, volatile and most of it lacks Meta data. The way forward for information retrieval in the web, in the future opinion would be to advocate the creation of a semantic web where algorithms which are unsupervised and reinforcement learners are used to classify and retrieve data. Thus the thesis explains the trends, threads and process of the text categorization algorithm which was implemented for finding the sensitive data analysis**.**

## REFERENCES

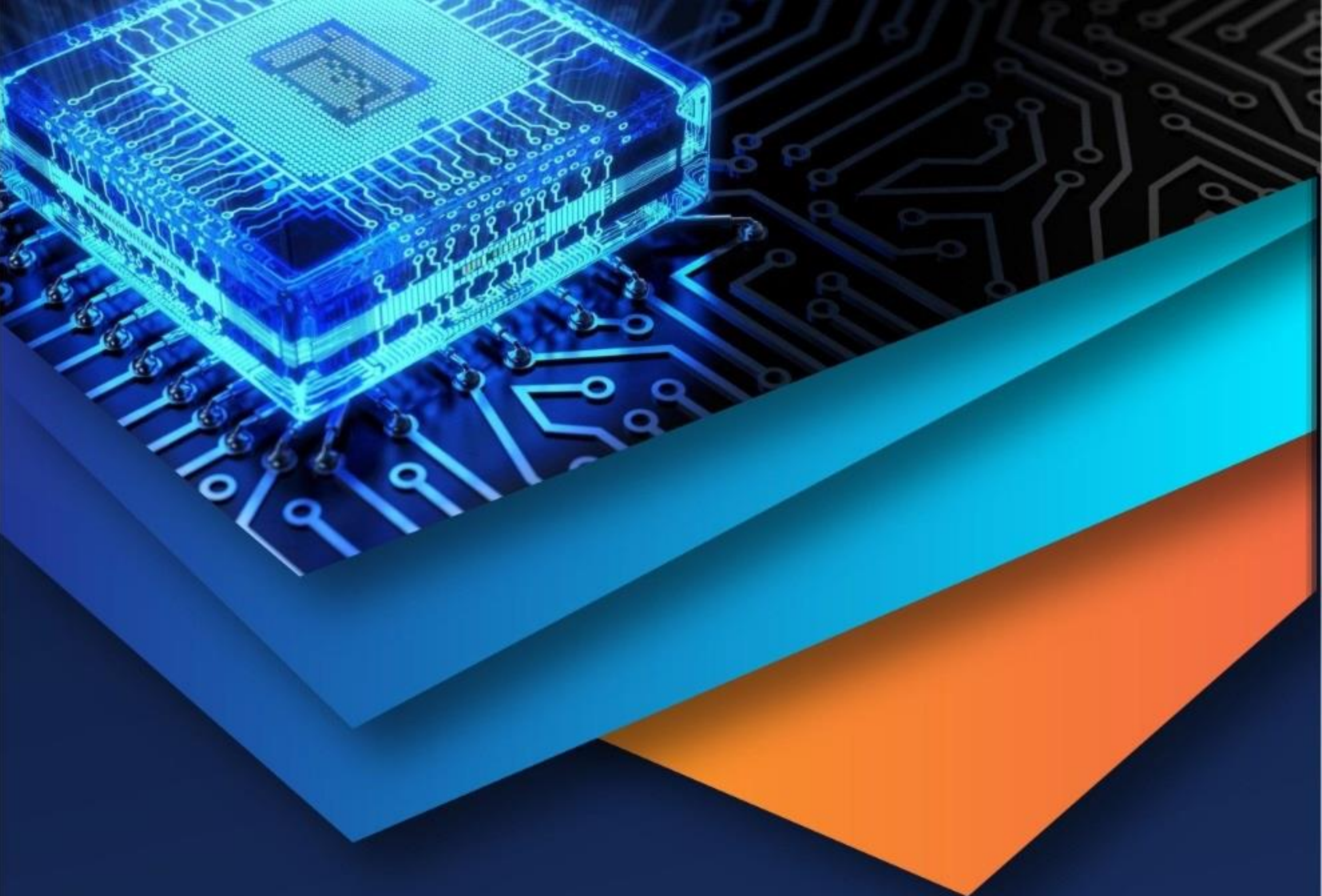
[1] Becker.H, Gravano.L and Naaman.M, (2010) ‘Learning similarity metrics for event identification in social media,’ in Proc. 3rd ACM WSDM, Macau, China.
[2] Blei.D.Mand Jordan.M.I, (Jan,2003) ‘Latent dirichlet allocation,’ J. Mach. Learn. Res., vol. 3, pp. 993–1022 TAN ET AL:INTERPRETING THE PUBLIC SENTIMENT VARIATIONS ON TWITTER 1169.
[3] J. Bollen, Mao.H and Pepe.A, (2011) ‘Modeling public mood and emo-tion: Twitter sentiment and socio-economic phenomena,’ inProc. 5th Int. AAAI Conf.Weblogs Social Media, Barcelona, Spain.
[4] Bollenand.J, Mao.H andZeng.X,(Mar,2011) ‘Twitter mood predicts the stock market,’ J. Comput. Sci., vol. 2, no. 1, pp. 1–8.
[5] Chakrabarti.D and Punera.K,(2011) ‘Event summarization using tweets,’ in Proc. 5th Int. AAAI Conf. Weblogs Social Media, Barcelona, Spain, 2011.
[6] Bhayani.R, Go.A and Huang.L, (2009) ‘Twitter sentiment classification using distant supervision,’ CS224N Project Rep., Stanford: 1–12.
[7] Griffiths.T.L and Steyvers.M (Apr,2004) ‘Finding scientific topics,’ inProc. Nat. Acad. Sci. USA, vol. 101, (Suppl. 1), pp. 5228–5235.
[8] Hall.D, Jurafsky.D and Manning.C.D,(2008) ‘Studying the history of ideas using topic models,’ inProc. Conf. EMNLP,Stroudsburg, PA, USA, 2008, pp. 363–371.
[9] Hong.Z, Mei.X and Tao.D, (2012) ‘Dual-force metric learning for robust distracter-resistant tracker,’ in Proc. ECCV, Florence, Italy.
[10] Hu.M and Liu.B, (2004) ‘Mining and summarizing customer reviews,’ in Proc.10th ACM SIGKDD, Washington, DC, USA.
[11] Hu.Y, John.A, Seligmann D.D andWang.F, (2012) ‘Et-lda: Joint topic modeling for aligning events and their twitter feedback,’ inProc. 26th AAAI Conf. Artif. Intell., Vancouver, BC, Canada.
[12] Jiang.L, Liu.X, Yu.M, Zhao.T and Zhou.M,(2011) ‘Target-dependent twitter sentiment classification,’in Proc. 49th HLT, Portland, or USA.
[13] T. Minka and J. Lafferty, “Expectation-propagation for the generative aspect model,” inProc. 18th Conf. UAI, San Francisco, CA, USA, 2002.[18] G. Mishne and N. Glance, “Predicting movie sales from blogger sentiment,” inProc. AAAI-CAAW, Stanford, CA, USA, 2006.
[14] B. O’Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, “From tweets to polls: Linking text sentiment to public opinion time series,” inProc. 4th Int. AAAI Conf. Weblogs Social Media, Washington, DC, USA, 2010.
[15] B. Pang and L. Lee, “Opinion mining and sentiment analysis,” Found. Trends Inform. Retrieval, vol. 2, no. (1–2), pp. 1–135, 2008

## AUTHOR DETAILS

Shilna K.S, B-TECH (CSE), ME (CSE), is pursing Master of Engineering from Maharaja Institute of Technology. She passed B-tech in the year 2013 and has intense curiosity in the fields like pattern recognition, data mining and machine learning. She had done projects in sentiment analysis and also using GSM technology. Having 2 years industrial experience in India and G.C.C

Prof. T. Sivakumar, ME (CSE), He had 10 years of experience in teaching industry. His areas of interest are Data mining and Data source. Participated in various international and national level conferences, seminars/webinars and workshops

Miji.K.Raju, B-TECH (CSE), ME (CSE), is pursing Master of Engineering from Maharaja Institute of Technology. She passed B-tech in the year 2013.She had done projects in Data mining and GSM Technology. 2 years of Teaching experience in India.




10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429


# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY