Opinion Mining From Blogosphere for Analysis of Social Networking

Amruta S. Dulange, R.B.Kulkarni, Supriya S. Ambarkar

Abstract*—Blogs are most common medium over web where user posts their opinion. It is considered to be a web space of the users where they share their views, beliefs and other philosophy. The blogs are generally categorized of two types: Itemized blogs, where the user posts his views and opinions against a web news or news item and personal blogs where users posts random topics of their interest under the header of their choice. As more and more number of users publish their data over the web, it becomes significant that Meanings are extracted from blog and they are indexed properly for information retrieval. In this work we develop a crawler to read data from RSS feeds of blogs and save them locally. Finally we apply data mining technique to index the blogs for easy searching and information extraction.

***Index Terms*— Web Mining, Blog, Blog Search, Blog Mining**

I. INTRODUCTION

The mining of opinions in textual materials such as Weblogs adds another dimension to technologies that facilitate search and analysis of particular entity. This paper focuses on identifying readers/customers viewpoint about a subject, rather than simply identifying the subject itself.

Analyzing text regarding its opinions can be extremely valuable to a legal researcher who is looking for a perspective on a legal issue or even information about a product or a service. Organizations may also benefit from automatic opinion mining by obtaining a timely picture of how their products or services, or more generally their names, are perceived by their customers. The Web is an expanding environment where customers go to find or submit opinions that may be ripe for mining.

II. BACKGROUNG

Web mining is a specific area of Data Mining, and is defined as the process of discovering knowledge, such as patterns and relations, from Web data. Web mining is generally divided into three main areas: *content mining*, *structure mining* and *usage mining*. Each one of these areas are associated to the three predominant types of data found in the Web[19].

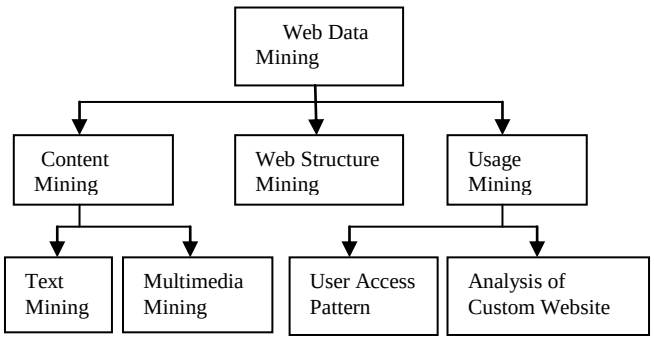

Fig:1 Classification of Web data mining

Web content mining from the document described its contents or takes a course of interesting knowledge, is a web-based content of the elements of the target Web mining. These elements have targeted text and hypertext data as well as graphics, images, and other multimedia data; both from the database of structured data, it also uses XML tags of HTML or semi-structured data and unstructured text of the free. Mining is the structure of the Web page from the hyperlink found in its structure and its relationship with each other. Through to find hidden in a page after the link structure of the model will be able to take advantage of this model on the Web page re-classification, can also be used to find similar sites. Based on the hyperlink topology, Web mining structure can be classified pages summed up the page and site structure, such as the generation of similarity between the Web site, the relationship between the Web site. Web Usage Mining is the user "visit marks" to obtain valuable information on the Web log data and data mining[21]. These data include: client, server-side data and data-side proxy. Web Usage Mining can be divided into general and special access to track the path of track. The former is used KDD (Knowledge Discovery in Database) to visit the general understanding of the technical patterns and trends, such as Web log mining; the latter is an analysis of each and every time the user visits the model, on the basis of these sites will automatically Mode Built structures, such as adaptive site.

Web use records of the excavation is aimed at forecasting the on-line users, compared with the actual site and look forward to the use of the difference, according to the user's interest to adjust structure of the site.

Blogs are likely to represent a single individual or a group. Blog information may differ from other kinds of text. For example, blogs are not always likely to be as well edited, as newspaper or magazine text.

Manuscript received September, 2013.

Miss.Amruta S. Dulange, BE CSE,ME CSE IInd Year Walchand Institute of Technology, Solapur, India.

Dr.Raj B.Kulkarni, Associate Professor in CSE Department, Walchand Institute of Technology, Solapur, India.

Mrs.Supriya S. Ambarkar, Assistant Professor in CSE Department, Walchand Institute of Technology,Solapur, India.

In contrast to other forms of text, blogs may use incomplete sentences and phrases. Individual blogs can have a huge impact and bloggers can gain substantial unsavory reputation. Blogs have also been used as the basis of generating information to predict sales.

Tomoyuki Nanno and his colleagues present architecture for collecting and monitoring blogs in Japanese [1]. Other researchers have studied how marketers use text mining techniques to analyze customers' opinions and reviews on the Web [2]. However, researchers created these systems for blog collection or Web page text mining only, so you can't apply them directly to blog mining applications.

Knowledge discovery in blogs is different from knowledge discovery in areas such as databases or Web documents due to blogs' unique characteristics, which introduce additional mining challenges. Although researchers have investigated several techniques to address different aspects of blog discovery, no comparisons among key knowledge discovery techniques for blogs exist [3]. This article examines three prominent techniques that are frequently applied to discovery in blogs — clustering, matrix decomposition, and ranking. The authors Geetika T. Lakshmanan & Martin A. Oberhofer compare them in terms of effectiveness in combating present challenges and their ability to accomplish challenging tasks required for effective blog mining. [3]

Gruhl et al. [9] apparently were among the first to examine the use of information from blogs as a basis to predict sales when they analyzed Amazon book sales. They found that peaks to discussions (chatter) in blogs were likely to be followed by sales peaks. Mishne and Glance [16] extended that work by noting that the nature of the chatter (positive, negative, etc.) also helped in the prediction of sales.

III. BLOGS, BLOG SEARCH AND BLOG MINING

Blogs are websites that provide content often generated by individuals. An analysis of "icerocket.com" provides many examples of the types of topics that are found in blogs: Technical, financial, political, entertainment, and news. Virtually, anyone can blog. There are few filters in place to limit blogs or what is in them. As a result, there are millions of blogs. However, since the information is coming from so many different sources, at so many different times, there may be real information, not previously realized or recognized, that is embedded in the blogs.

Blogs are likely to represent a single individual or a group. Blog information may differ from other kinds of text. For example, blogs are not always likely to be as well edited, as newspaper or magazine text. In contrast to other forms of text, blogs may use incomplete sentences and phrases. Individual blogs can have a huge impact and bloggers can gain substantial unsavory reputation.

A. Blog search

A number of search engines, including Technorati, Yahoo and Google, provide the ability to search blogs for specific concepts or issues. These search engines allow users to easily find blogs that contain pre-specified chunks of opinionated text. For example, "AMAZING" would allow finding all of the pages with the appropriate set of opinion-oriented text. Such search engines can be employed by other software to generate information and insights[22].

B. Blog mining

Blog mining is the process of searching and analyzing blogs in order to generate additional insights that might otherwise not be found by examining a single blog[23].If blogs contain information and knowledge, whether tacit or explicit, by analyzing and "mining" the information in them, we can begin to make assertions, particularly in those settings where we are able to pull together information and knowledge from multiple different blogs. Blog mining tries to create an overall understanding of information from the disparate sources. Marketing researchers and companies have long been interested in capturing information and knowledge about the opinions of buyers or potential buyers of their products. However, interviewing people about their opinions is time consuming and costly, and there is concern if the individual is telling the truth or telling the marketer what they want to hear. In contrast, blogs provide a readily available and opinion-based content media that provides sentiment about a range of issues. As a result, being able to use those blogs for gathering opinion information potentially can provide a low cost source of information about those opinions and sentiment, regarding particular issues and concerns, gathered in real time.

IV. OPINION CAPTURED IN BLOGS

Every Blog site provides a forum for bloggers to present opinion, ratings, sentiment and information about a range of issues. Although bloggers may use pictures, links, Dilbert cartoons, and videos, in this paper we are primarily concerned with the expression of opinion using text only.

A. Opinions and sentiment

Humans generally are able to distinguish between positive and negative opinions, although the case of sarcasm can make it difficult. However, it is difficult for humans to distinguish between neutral positions and opinion bearing positions. As a result, the goal for a computer program is to determine a similar set of perspectives, but it is more likely that such programs will be able to determine positive or negative positions.

B. Sample opinion word dictionary

+ve verbs	-ve verbs	Positive adjectives	Negative adjectives
Love, like	Hate, dislike	Good,best,better, happy, fantastic, extraordinary, successful,glad,desirable,worthy,remarkable, funnylovely,perfect, nice,impressive, decent,beautiful, entertaining, fascinating,brilliant, gorgeous,amazing, splendid,distinctive, desirable,excellent, great,awesome, fabulous	Bad, awful, suck, worse, worst, poor annoying, and stupid

Table 1 : Sample of opinion word dictionary

V. METHODOLOGY

The working of the work and the methodology can be described through following figure.

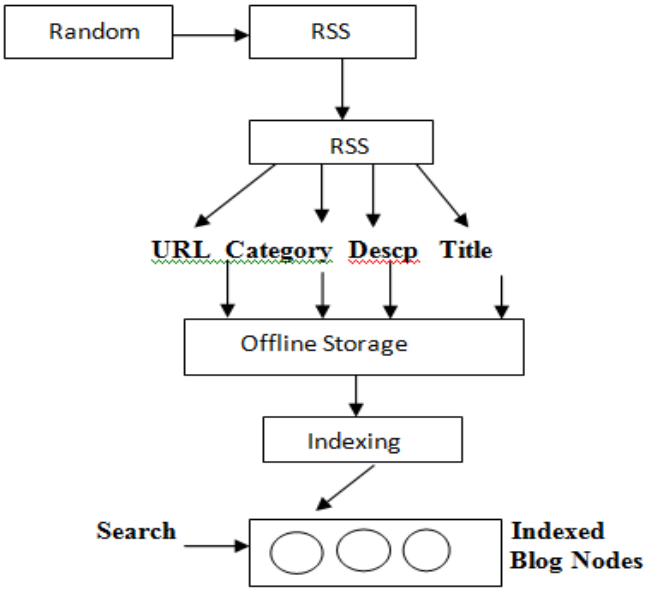

Fig2: Methodology block diagram

In above Fig 2 ,a blog reader is essentially a XML parser that reads the xml feed from the blogs. These feeds provide information about the topics and the URL of the blogs. Here the feed data is first stored as a plain text locally in the system and then this plain text is given as input to the indexer. Indexer indexes the files based Fig: Opinion Mining from Similar Blogs Block diagram on information and using maximum likelihood based indexing. Indexed data is stored in the database for fast querying and searching.

VI. APPROACH

A. *Indexing the collection*

The feeds are pages in RSS/Atom format. Each RSS feed represents a single channel, with metadata for title, URL, description, generator, language and a list of items. Each item contains elements such as title, URL for the content, URL for the comments, description, date, creator and category. Atom feeds use slightly different naming but contain similar metadata and items.

B. *Identifying content*

A topic can be considered as a representative of the bloggers interest in real world events. A topic can vary from the commercial launch of a product like the Apple's new iPod to variations in political policies. But unfortunately blog pages are messed up with all sort of extra information besides the blog post and the reader's comments: pages often include lists of similar related pages, annotated lists of previous posts, other reviews, navigation bars, side bars, advertising, etc. If we indexed the page as a normal HTML page, all the text in these parts would end up in the index, leading to results with poor relevance.To avoid this here we deal specially with blogs generated by programs which follow well defined markup rules allowing the post's content to be identified.
Here numbers of blogs are aggregated from few blog sites which are well-liked in India like WordPress, Technorati, Yahoo and Times of India. Each generator creates pages with a specific markup style. For instance WordPress encloses the posts within a element, and the post head within a <H2 class=post-head> element. Comments are enclosed in a <DIV class="commentlist"> element. Blogger instead uses div's post, postbody and to enclose comments. For these most used content generators, we created a list of elements to be included.

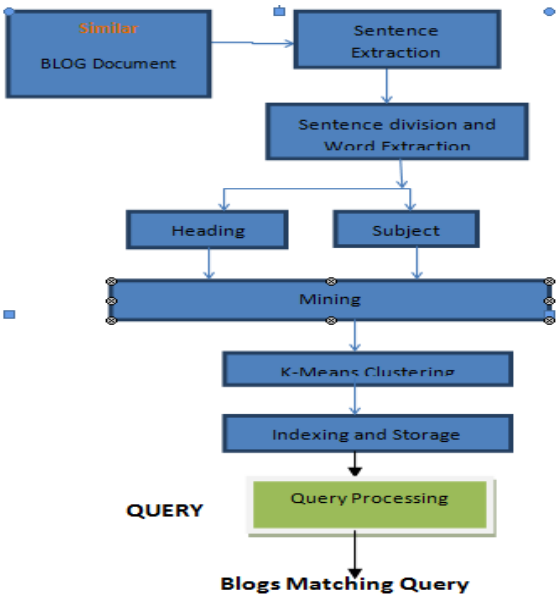

Fig 3:Opinion Mining from Similar Blogs Block diagram

C. *Opinion Mining*

Once we get HTML structure of a blog of our interest we can extract the comments by filtering it from the rest of code. Suppose 'A' is our topic of Interest then there may be 'n' number of blogs written on it on multiple different sites. By using the GUI provided for comment extraction, we can extract comments from each individual blog on 'A' and these are finally grouped under title 'A'[15]. These comment set grouped into cluster which contains satisfactory rang of comments extracted from desperate sources. As we are examining number of blogs on same topic we may get multiple clusters which are stored locally for further procedure.

Fig 4 : Opinions are aggregated from multiple blog sites and are clustered.

D. *Search Strategy*

At the beginning of search operation aggregated index of

the entire document within the cluster is formed. Such aggregated indexed data is now forwarded for searching. As the blog data is updated in our repository, the tool can periodically build an index using the Lucene search and indexing library. This feature can be used similar to a traditional search engine for the data sources currently indexed by the system[20].

The search engine now will perform look for opinions which matches the keyword entered through interface provided for searching.

For example to search for type:

```
[OPINIONATED: 'United States']
```

This will return all documents where opinionated word occurs within the document. By using subjective word classifier we can easily distinguish between positive and negative comments or we can enter any keyword which reflects the mood of comments.

e.g. If we enter the word agree, good, like or anything which may be considered as positive or in favor will be matched & result will be displayed for it.

VII.RESULTS AND ANALYSIS

Following GUI shows the Interface for extracting data/opinion from Blog. We can select any blog site from listed sources and Fetch XML Content from it. Once a link is selected the main story and the blog with its comments appears. After clicking Get, the Page Source of fetched blog will get stored locally on the system as a text file which is our training sample for mining of opinions. Mining procedure can be conducted by extracting user post from the text file and by simply running our mining evaluator which indexes each node we can address each blog with the result showing positive /negative perception of users/customers on it.

Fig 5 a) Interface for extracting data/opinion from Blog.

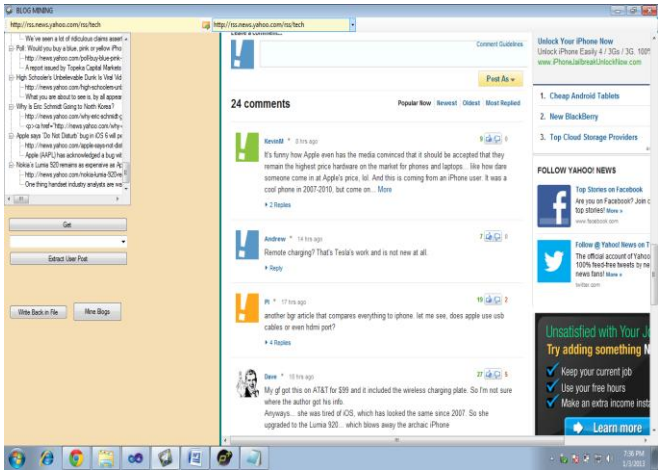

Fig 5 b) Interface for extracting data/opinion from Blog.

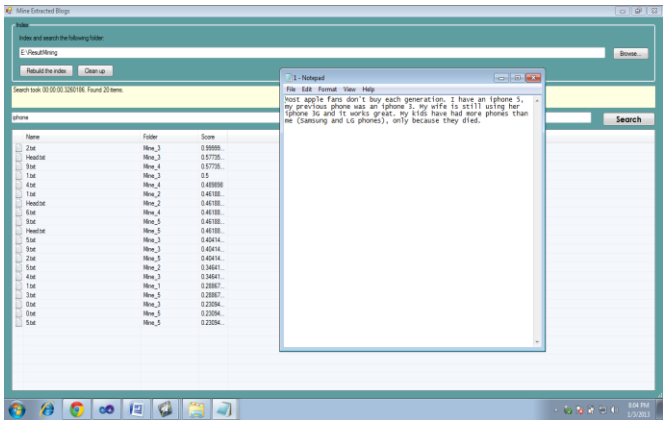

Fig 5 c). Search Returns fast result showing the score, source file location & keyword in a sample file.

This evaluator is executed by considering documents (opinion) on below stated subjects from yahoo, Technorati and times of India's blog sites. And for each subject we have referred number of blogs and extracted hundreds of opinionated text for analyzing polarity of opinions. . Following table 2 is the blog sources which we have considered for evaluation; similarly we can use the same evaluator to work with hundreds of blog posts from disparate sources.

Blackberry 10	**On Galaxy S IV Launch**

• BlackBerry says an unnamed partner has ordered 1 million BlackBerry 10 phones.- *By Brad Reed* • BlackBerry shares surge on huge order for new devices.-*by Euan Rocha* • BlackBerry promises to update its Android emulator to run Jelly Bean apps.-*by Dan Graziano* • Nokia Lumia 920 vs Sony Xperia Z vs BB Z10 vs HTC Butterfly.-*by Ravi Sharma & Neeraj Saxena,* • BlackBerry Z10 sell-outs continue in new markets, but sales said to be slowing in Canada, U.K.-*by Zach Epstein*	• Review: Tech in Galaxy S 4 doesn't come together. -*By Peter Svensson* • How Samsung's Galaxy S4 matches up vs iPhone5. - *Reuters* • Samsung set to launch new iPhone challenger.- By Peter Svensson • Why Apple Should Worry About the Samsung.- *By Rebecca Greenfield* • Galaxy S IV. Samsung announces the Galaxy S 4: Eight cores, 13 megapixels, one gorgeous HD display. -*By Zach Epstein* • Galaxy S IV could feature a 'Hyper Bright Display'.- *By Dan Graziano* • Live from Samsung's Galaxy S 4 unveiling.-by *By Dan Graziano*
Micromax	Windows 8
• 30 smartphones in a year? Slow down, Micromax. -*By Javed*	• Windows 8 still used less than Vista despite slowly gaining traction in 2013. *By Dan Graziano* • Windows 8 is no Vista, but still considered polarizing. *By Brad Reed* • WINDOWS 8: GOOD, BAD AND UGLY -*BY JAVED ANWER* .

Table 2: Blogs referred for mining of opinions.

According to this evaluation our runs obtained the following scores as shown in table 3. Positive column give you an idea about the percentage of people who liked the product and have focused on its positive features. Similarly next column for negative shows that people that has not accepted it and shared their reason of refusal. As people always talk about products affordability and compare the cost and features with the other one and they may declare it as an expensive product. Non opinionated results are presented with the neutral column.

Subject	Positive opinion	Negative Opinion	Over Priced	Neutral
Blackberry 10	**19%**	**11%**	**18%**	**44%**
Nokia Lumia				
Windows 8	**24%**	**37%**	**-**	**39%**
Nexus 7				
HTC Butterfly	**46%**	**12%**	**24%**	**8%**
Miramax Smartphone				

Table 3. Score obtained after examining result of search.

VI. CONCLUSION

Blog Mining is an important aspect of web mining and the web data analysis. As most of the modern day news are debated online and the user opinions are presented online, it becomes important for developing tools which can not only extract correlated blogs but also gets an overview of independent and in turn generalized overview of the blogs. Many algorithms are proposed in this direction. Most of these papers are organized to detect the categories in the blogs only and do not present a comprehensive overview of the entire technique of fetching the RSS blog data and analyze them on the fly. In this work we developed an entire lifecycle of fetching and analyzing the blogs for mining information. The technique is based on similarity of the blog with its subject matter and the presence of opinion in such correlated blogs. The result shows a significant similarity with human perception. The technique can be further improved by incorporating machine learning technique with the current algorithm for better learning of the opinions in the blogs.

REFERENCES

[1] T. Nanno et al., "Automatically Collecting, Monitoring, and Mining Japanese Weblogs," Proc.13th Int'l Conf. WWW, (WWW 2004), ACM Press, 2004, 320–321W.-K. Chen, *Linear Networks and Systems* (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.

[2] N. Glance et al., "Analyzing Online Discussion for Marketing Intelligence," Proc. 14th Int'l Conf. WWW (WWW 2005), ACM Press, 2005, pp. 1172–1173.

[3] Geetika T. Lakshmanan IBM T.J. Watson Research Center Martin A. Oberhofer IBM Software Group, Germany Knowledge Discovery in the Blogosphere Approaches and Challenges

[4] M. Kobayashi, K. Takeda, "Information retrieval on the web". ACM Computing Surveys (ACM Press) 32 (2): 144–173, 2000

[5] S. Thies, Content-Interaktionsbeziehungen im Internet. Ausgestaltungund Erfolg, 1st ed., Gabler, 2005

[6] C. Marlow, "Audience, structure and authority in the weblog community," in Proceedings of the International Communication Association Conference, 2004

[7] N. Agarwal, H. Liu, L. Tang, and P. S. Yu, "Identifying the influential bloggers in a community," in WSDM '08: Proceedings of the international conference on Web search and web data mining. New York, NY, USA: ACM, 2008, pp. 207–218.

[8] M. Chau and J. XU, "Mining communities and their relationships in blogs: A study of online hate groups," International Journal of Human- Computer Studies, vol. 65, no. 1, pp. 57–70, January 2007.

[9] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins, "On the bursty evolution of blogspace," World Wide Web, vol. 8, no. 2, pp. 159–178, 2005.

[10] M. Goetz, J. Leskovec, M. Mcglohon, and C. Faloutsos, "Modeling blog dynamics," in International Conference on Weblogs and Social Media, May 2009.

[11] S. C. Herring, I. Kouper, J. C. Paolillo, L. A. Scheidt, M. Tyworth P. Welsch, E. Wright, and N. Yu, "Conversations in the blogosphere: An analysis "from the bottom up"," in Proceedings of the 38th Annual Hawaii International Conference on System Sciences. IEEE Computer Society, 2005, p. 107.2.

[12] P. Wortmann, "Topic-based blog article search for trend detection," Technical University of Kaiserslautern, Project Thesis, 2009. [Online]. Available: http://www.dfki.uni-kl.de/_obradovic/downloadpa-wortmann.pdf

[13] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," Stanford University, Technical Report, 1998. [Online Available: http://explorer.csse.uwa.edu.au/reference/browse paper.php?pid=233281827

[14] S. Wasserman, K. Faust, and D. Iacobucci, Social Network Analysis Methods and Applications (Structural Analysis in the Social Sciences). Cambridge University Press, 1994.

[15] Bruno Ohana and Brendan Tierney," Sentiment Classification of Reviews Using SentiWordNet", Dublin Institute of Technology, School of Computing Kevin St. Dublin 8, Ireland

[16] G. Mishne, N. Glance, Predicting Movie Sales from Blogger Sentiment, American Association for Artificial Intelligence, 2005.
[17] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins, "On the bursty evolution of blogspace," World Wide Web, vol. 8, no. 2, pp. 159–178, 2005.
[18] Darko Obradovi´c, Stephan Baumann and Andreas Dengel, "A Social Network Analysis and Mining Methodology for the Monitoring of Specific Domains in the Blogosphere", International conference on Advances in Social Network Analysis and Mining 2010.
[19] Barbara Poblete (2009), "Query-Based Data Mining for the Web" TESI DOCTORAL UPF, University Pompeu Fabra.
[20] Bo Pang, Lillian Lee, "Opinion Mining and Sentiment Analysis" Foundations and Trends in Information Retrieval Vol. 2, Nos. 1-2 (2008).
[21] A. Senthil Kumar & N. Palanisamy ,"CHALLENGES FOR WEB MINING" Proceedings of the 2008 International Conference on Computing, Communication and Networking 2008 IEEE.
[22] Shamanth Kumar, "Towards Building a Social Computing Tool for Social Scientists", Computer Science and Engineering Arizona State University Tempe, Arizona 85281.
[23] Daniel E. O'Leary, "Blog mining-review and extensions: From each according to his opinion", Marshall School of Business, University of Southern California, Los Angeles, CA 90089-0441, United States.

First Author Miss.Amruta S. Dulange,.BE CSE,ME CSE IInd Year Walchand Institute of Technology,Solapur.

Second Author Dr.Raj B.Kulkarni BE CSE, ME CSE, PhD in Web Restructuring and web mining , worked as a Associate Professor in CSE Department, Walchand Institute of Technology,Solapur, 8 papers published in International journals , 20 papers published in National and International Conference Proceedings. CSI member , Worked as a Akash Workshop coordinator held by IIT,Bombay. Attended so many workshop at national and international level.

Third Author Mrs.Supriya S. Ambarkar , BE CSE ,MTech CSE, working as a Assistant Professor in CSE Department, Walchand Institute of Technology,Solapur, 4 papers published in International journals , 10 papers published in National and International Conference Proceedings , IEI Member. Attended a 2 workshops arranged by IIT,Bombay.