ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

Automatic Text Summarization using Features Extraction and Fuzzy Logic Scoring

Riya Kamble[1], Saurabh Shah[2], Aalok Nerurkar[3], Kanhaiya Prasad[4], Reena Mahe[5]

B.E. Student, Dept. of I.T., Atharva College of Engineering, Malad, Mumbai, India[1]

B.E. Student, Dept. of I.T., Atharva College of Engineering, Malad, Mumbai, India[2]

B.E. Student, Dept. of I.T., Atharva College of Engineering, Malad, Mumbai, India[3]

B.E. Student, Dept. of I.T., Atharva College of Engineering, Malad, Mumbai, India[4]

Assistant Professor, Dept. of I.T., Atharva College of Engineering, Malad, Mumbai, India[5]

ABSTRACT: With the fast development of the quantity and complexity of archive sources on the internet, it has come to be increasingly more essential in imitation of providing a modern mechanism for user for finding specific facts in available documents. Text summarization has turned out to be an essential and well timed tool because of supporting and then decoding the tremendous volumes of text available into documents. “Text Summarization” is a method of bringing a lesser version of original text that contains the important information. It can be broadly differentiated into two types which are Extraction and Abstraction. This project focuses on the Fuzzy logic Extraction approach for text summarization and the semantic approach of text summarization using Latent Semantic Analysis.

KEYWORDS: Text summarization; Fuzzy logic Extraction; Latent Semantic Analysis; Semantic approach; Extraction; Abstraction

I. INTRODUCTION

Automatic summarization means a mechanically short output is addicted when an input is applied. We should remember that an input is a well structured document. For this even there are opening pre-processes such as Tokenization, Sentence Segmentation, Removing stop words and Word Stemming. An extractive summarization method is composed of choosing most important sentences, words, paragraphs etc. from the original record and concatenating them into shorter form. An Abstractive summarization is a grasp about the predominant ideas in a file and expresses those thoughts into an obvious simplistic language.

Text Summarization is a lively concern of research among every text regarding the IR and NLP communities. People can keep up with the world affairs by listening to news bites. People can go to the movies largely over the basis of critiques they’ve seen. People can base investment decisions on stock market updates. With summaries, People can make effective decisions in less time. The motivation right here is in conformity to construct a certain system which is computationally surroundings pleasant and then create summaries automatically.

Text summarization is able to stay categorized in two ways, as abstractive summarization and extractive summarization. Extractive summarization [6] is bendy but consumes much less time namely compared to abstractive summarization. In extractive summarization it considers the whole paragraph into a matrix form, and on the basis of some feature vectors all the indispensable or vital sentences are extracted.

II. RELATED WORK

This work [1] is done by Archana AB and Sunitha. C in 2013 and it deals with Text summarization which can be classified into two approaches: extraction and abstraction. This paper focuses on extraction technique. The purpose of text summarization concerning extraction strategy is sentence selection. One of the techniques in conformity is to gain the appropriate sentences assigned partial numerical measure of a sentence for the summary called sentence

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

weighting and then choose the best ones. A summary textual content is a derivative of a source text condensed through selection and/or generalization on necessary content. Query-focused summaries enable customers to find more applicable documents more accurately, with less need to seek advice for the full text of the document. Extractive summarization methods try to locate the most necessary topics of an input document and pick sentences that are related to these select concepts to create the summary. This paper is a Comparative discipline of four methods used for extractive summarization, namely, Neural Network, Graph Theoretic, Fuzzy based method and Cluster based method.

This work [2] is done by Josef Steinberger and Karel Ježek in 2009 and it deals with using latent semantic analysis in text summarization. It describes a generic text summarization method which utilizes the latent semantic analysis technique to perceive semantically necessary sentences. Then it proposes twin modern evaluation methods based on LSA, which measure context similarity between an authentic file and its summary. In the evaluation part we compare seven summarizers by a classical content-based evaluator and by the two new LSA evaluators. We also learn an influence regarding summary length on its quality from the angle of the three mentioned assessment methods. LSA is very sensitive on a stop list and a lemmatization process. Other weighing schemes and a normalization of a sentence vector on the SVD input is needed. Other evaluations are needed, especially on longer texts than the Reuters documents are.

This work [3] is done by Rafael Ferreira, Luciano de Souza Cabral, Rafael Dueire Lins, Gabriel Pereira e Silva in 2013 and deals with Text summarization which is the process of automatically creating a shorter version of one or more text documents. Essentially, text summarization techniques are classified as Extractive and Abstractive. Extractive strategies perform textual content summarization by choosing sentences of files according to partial criteria. Abstractive summaries strive to improve the coherence amongst sentences by disposing redundancies and clarifying the contest of sentences. In phrases regarding extractive summarization, sentence scoring is the technique most used for extractive textual content summarization. This paper describes and performs a quantitative and characteristic assessment concerning 15 algorithms for sentence scoring ready in the literature. Three special datasets (News, Blogs and Article contexts) have been evaluated. In addition, instructions to improve the sentence extraction results obtained are suggested. This paper provided the five best results obtained with different test sets, one would obtain a coincidence of four methods as being the best ones: TF/IDF, Word Frequency, Lexical Similarity and Sentence Length. The strategy ‘‘Text-Rank Score’’ was also chosen by as providing good results for two of the three data sets tested.

This work [5] is done by Róbert Móro and Mária Bieliková in 2012 and it deals with Automatic text summarization which aims to address the information overload problem by extracting the most important information from a document and which can help a reader to decide whether it is relevant or not. In this paper we advocate an approach of personalized textual content summarization which improves the conventional automated text summarization methods with the aid of accepting the differences within reader’s characteristics. It uses annotations added by readers as one of the sources of personalization. In this paper we have proposed a method of personalized summarization, which improves traditional summarization strategies by taking various user characteristics including context. The achievement lies in the proposal of the specific raters that take into account terms applicable for the domain or the stage of knowledge of an individual user and the technique of the raters’ aggregate which permits thinking about more than a few parameters or context of the summarization.

III. PROPOSED ALGORITHM

A. *Pre-processing:*

- Segmentation: It is a process of dividing a given document into sentences.
- Removal of Stop words: Stop words are frequently occurring words such as ‘a’ an’, the’ that provides less meaning and contains noise. The Stop words are predefined and stored in an array.
- Tokenization: The words are assigned tokens or weights according to the usage and importance.
- Word Stemming: converts every word into its root form by removing its prefix and suffix so that it can be used for comparison with other words.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

B. ***Features Extraction:***

The text document is represented by set, D= {S1, S2, - - - , Sk} where, Si signifies a sentence contained in the document D. The document is subjected to feature extraction. The important word and sentence features to be used are decided .This work uses features such as Title word, Sentence length, Sentence position, numerical data, Term weight, Sentence similarity, existence of Thematic words and proper Nouns . The flow of summarization is as shown in the [Fig.1].

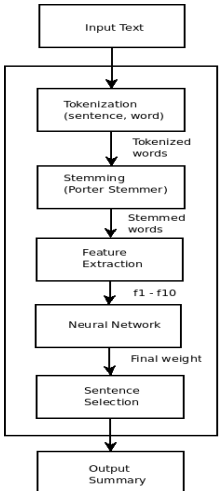

Fig 1. Flow for Summarization

C. ***Fuzzy Logic Scoring:***

Thus each sentence is associated with 8 feature vectors. Using all the 8 feature scores, the score for each sentence are derived using fuzzy logic method. The fuzzy logic method uses the fuzzy rules and triangular membership function. The fuzzy rules are in the form of IF-THEN. The triangular membership function fuzzifies each score into one of 3 values that is LOW, MEDIUM & HIGH. Then we apply fuzzy rules to determine whether sentence is unimportant, average or important. This is also known as defuzzification.

Sample of IF-THEN rules are described below:

IF (No Word In Title > 0.81) and (Sentence Length > 0.81) and (Term Freq > 0.81) and (Sentence Position > 0.81) and (Sentence Similarity > 0.81) and (No Proper Noun > 0.81) and (No Thematic Word > 0.81) and (Numerical Data > 0.81) THEN (Sentence is important).

Fuzzy Logic for Summarization is as shown in the [Fig.2].

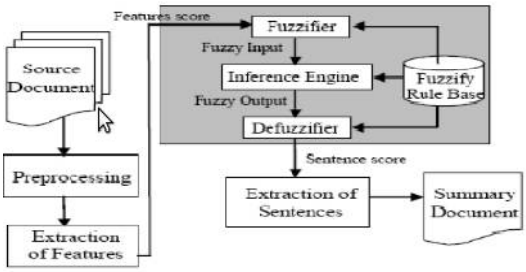

Fig 2. Fuzzy Logic for Summarization

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

D. ***Sentence Selection:***

All the sentences of a document are ranked in a descending order based on their scores. Top n sentences of highest score are extracted as document summary based on compression rate. Finally the sentences in summary are arranged in the order they occur in the original document.

E. ***Mathematical Approach:***

D = {s1 , s2 , s3 , s4,.....sn}
Where
- s is Sentences.
- D is Document.
If(W>T)
Where
- W is Weight.
- T is Threshold.
Sum(old) = {s1 , s2};
Sum(new) = {s1 , s3 , s4};
Sum(Final) ={ Sum(old) ^ Sum(new)} U { WP[Sum(old) U Sum(new)]};
Where ..... (W>T)
Sum(Final) = {s1} U {s4} ;
Sum(Final) = {s1 , s4};

IV. RESULTS

We have developed a Desktop application for the proposed system. In Desktop application, upon successful login by the user, the user will be presented with the main form [Fig.4] in which user will be allowed to 'Select' a sample document and an 'Upload' or 'View' option. After uploading [Fig.5] and selecting the document or viewing it, the user will be allowed to 'Summarize' the available document and can undergo removal of 'Stop Words'. If the user selects the 'Summarize' option, the summarized document will be displayed on the screen whereas removal of stop words would display a summarized document without stop words. After obtaining this, the user can click on 'Categorization' [Fig.9] to interpret the summary related to a particular domain. Lastly, the user can click on sentimental analysis to perceive a particular summary as positive or negative.

In Desktop Application, upon successful login, the admin will be presented with a screen [Fig.8] in which admin will be allowed to select between the displayed options. After selecting the 'Summarize' option, a screen will be displayed in which the admin will be able to see the summarized document available.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

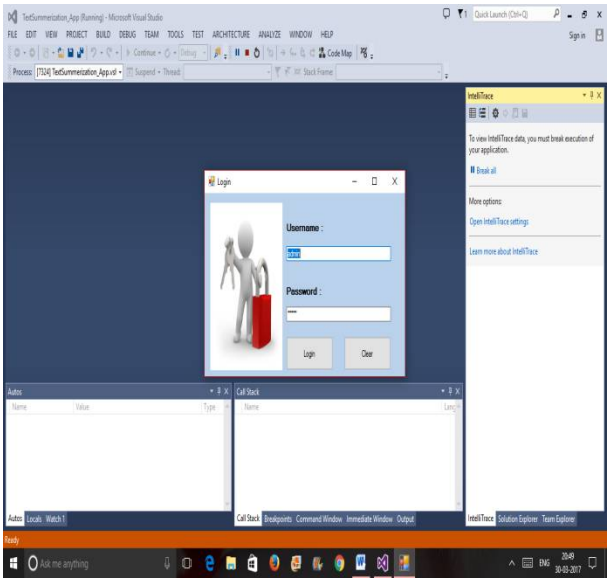

Fig 3. Login Page

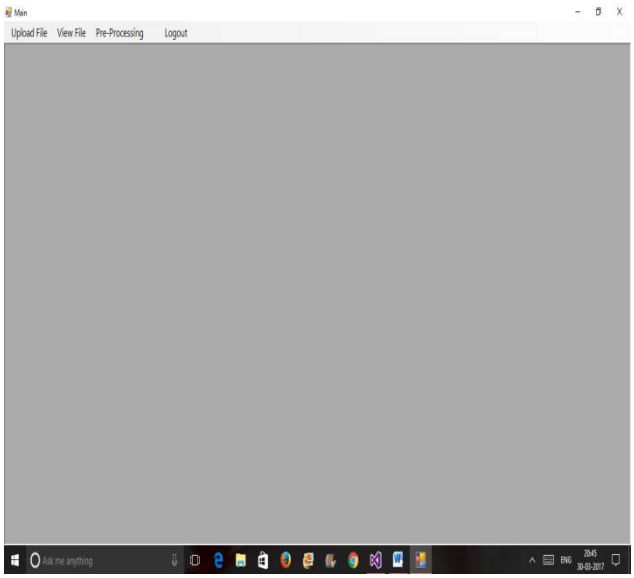

Fig 4. Main Form for Summarization

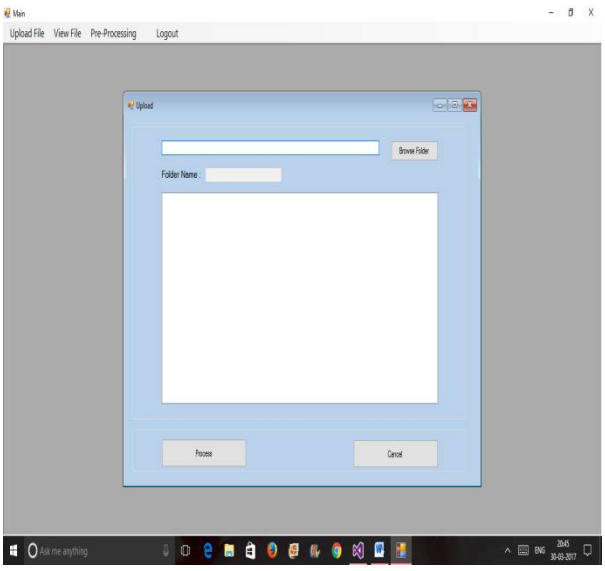

Fig 5. Upload Document

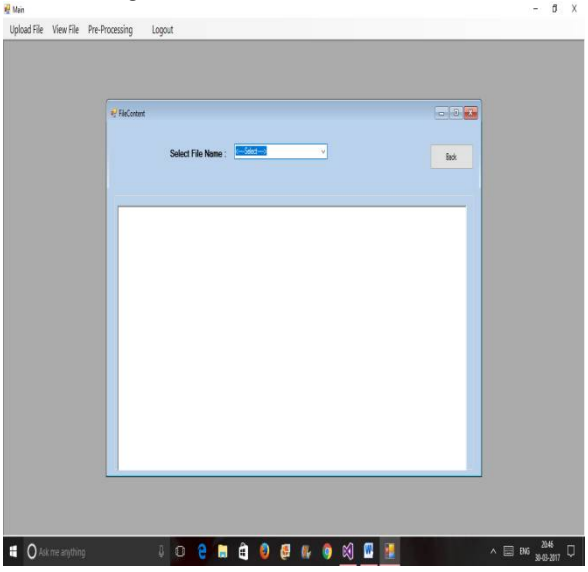

Fig 6. View Document

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

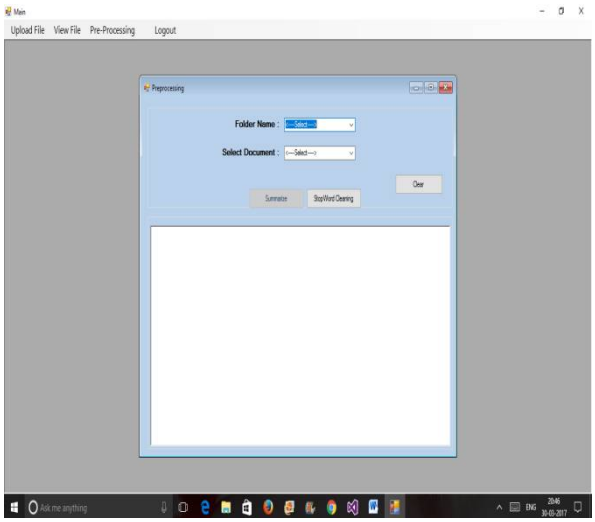

Fig 7. Pre-processing stage 1

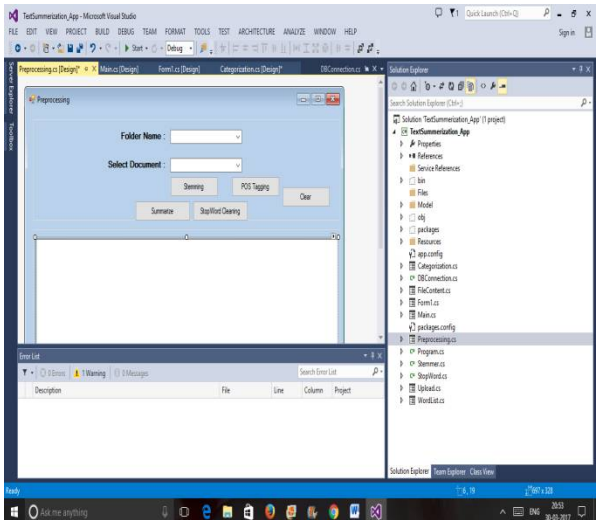

Fig 8. Pre-processing stage 2

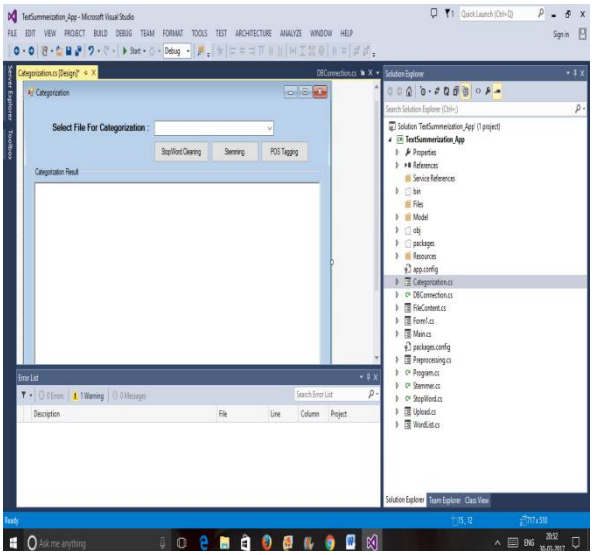

Fig 9. Categorization

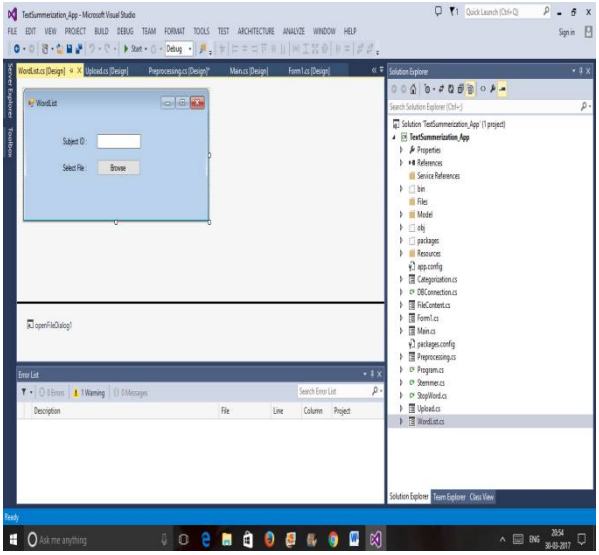

Fig 10. Sentimental Analysis

V. CONCLUSION AND FUTURE WORK

Automatic summarization is a complicated assignment that consists of quite a few sub-tasks. Every sub-task immediately affects the purpose to cause high virtue summaries. In extraction based summarization the necessary portion regarding the process is the identification of essentially applicable sentences of text. Use of fuzzy logic as a summarization sub-task elevated the virtue concerning summary by a greater amount. The results are clearly visible in the comparison graphs. Our algorithm shows better results as compared to the output produced by twin online summarizers. Thus our proposed technique improves the virtue of summary by incorporating the latent semantic analysis into the sentence function extracted fuzzy logic system to capture the semantic relations between ideas in the text. Future enhancement includes the categorization of useful information from data using different mining techniques. This system can also be used as a base for sentimental analysis.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

REFERENCES

1. Archana AB.1, Sunitha. C.2, "An Overview on Document Summarization Techniques", International Journal on Advanced Computer Theory and Engineering (IJACTE), Volume-1, Issue-2, 2013, ISSN (Print): 2319 ″U 2526.
2. Josef Steinberger.1, Karel Ježek.2 , "Using Latent Semantic Analysis in Text Summarization and Summary Evaluation", Department of Computer Science and Engineering, Univerzitní 22, CZ-306 14 Plzeň.
3. Rafael Ferreira.1, Luciano de Souza Cabral.2, Rafael Dueire Lins.3, Gabriel Pereira e Silva.4, Fred Freitas.5, George D.C. Cavalcanti.6, Luciano Favaro.7, "Assessing sentence scoring techniques for extractive text summarization ",Expert Systems with Applications 40 2013 Elsevier, 5755-5764.
4. Mrs.A.R.Kulkarni.1, Dr.Mrs.S.S.Apte.2 "A domain specific automatic text summarization using fuzzy logic ",International Journal of Computer Engineering and Technology (IJCET), Volume 4, Issue 4, July-August (2013) ISSN 0976- 6367(Print), ISSN 0976 - 6375(Online).
5. Róbert Móro.1, Mária Bieliková.2 "Personalized Text Summarization Based on Important Terms Identification ", 23rd International Workshop on Database and Expert Systems Applications , 2012 IEEE, 1529-4188.
6. Ladda Suanmali.1, Naomie Salim.2, Mohammed Salem Binwahlan.3 "Fuzzy Genetic Semantic Based Text Summarization ", Ninth International Conference on Dependable, Autonomic and Secure Computing , 2011 IEEE, 978-0-7695-4612-4.
7. ZHANG Pei-ying.1, LI Cun-he.2 , "Automatic text summarization based on sentences clustering and extraction ", 2009 IEEE, 978-1-4244-4520-2.
8. Farshad Kyoomarsi.1, Hamid Khosravi.2, Esfandiar Eslami.3, Pooya Khosravyan Dehkordy.4; "Optimizing Text Summarization Based on Fuzzy Logic ",Seventh IEEE/ACIS International Conference on Computer and Information Science, 2008, 978- 0-7695-3131-1.
9. Ladda Suanmali.1, Naomie Salim.2, Mohammed Salem Binwahla.3 ,"Feature-Based Sentence Extraction Using Fuzzy Inference rules ",International Conference on Signal Processing Systems , 2009 IEEE, 978-0-7695-3654-5.
10. Ladda Suanmali.1, Mohammed Salem Binwahlan.2, Naomie Salim.3 "Sentence Features Fusion for Text Summarization Using Fuzzy Logic ", Ninth International Conference on Hybrid Intelligent Systems, 2009 IEEE, 978-0-7695-3745-0.
11. Riya Kamble.1, Saurabh Shah.2, Aalok Nerurkar.3, Kanhaiya Prasad.4, Reena Mahe.5 “Automatic Text Summarization” International Journal of Engineering Research and Technology (IJERT) ICATE – 2017Conference Proceeding, Special Edition – 2017 ISSN: 2278-0181.

BIOGRAPHY

Riya Kamble is a B.E. student in the I.T. Department, Atharva College of Engineering, Mumbai University, Mumbai, Maharashtra, India.

Saurabh Shah is a B.E. student in the I.T. Department, Atharva College of Engineering, Mumbai University, Mumbai, Maharashtra, India.

Aalok Nerurkar is a B.E. student in the I.T. Department, Atharva College of Engineering, Mumbai University, Mumbai, Maharashtra, India.

Kanhaiya Prasad is a B.E. student in the I.T. Department, Atharva College of Engineering, Mumbai University, Mumbai, Maharashtra, India.

Reena Mahe is working as an Assistant Professor at Department of Information Technology, Atharva College of Engineering, Mumbai University, Mumbai, Maharashtra, India.