International Journal of Advance Research in Science and Engineering
Volume No.06, Issue No. 09, September 2017
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Email Summarization using Latent Dirichlet Allocation (LDA)

Kuldeep Kaur[1], Asst. Prof. Anantdeep Kaur[2]

[1,2]Department of Computer Engineering, Punjabi University, Patiala, (India)

ABSTRACT

Summarizing email conversations are challenging due to the characteristics of emails, especially the conversational nature. Most of the existing methods dealing with email conversations use the email thread to represent the email conversation structure, which is not accurate in many cases. We are presenting an approach which will help to remove flaws of Clue Word Summarization (CWS) algorithm. In CWS the emails are used to mark with only one category whereas according to LDA one email may not be categorized into one category only. The email may have more than one categorization. We have designed system for categorization of emails using NetBeans IDE. The accuracy of thealgorithm is calculated by testing both the algorithm with no. of content inputs. The Accuracy of CWS is 75% whereas the accuracy of LDA is 95%. The processing time of CWS is 1800 milliseconds whereas the processing time of LDA is 1500 milliseconds.

Keyword: Clue Word Summarizer, Emails Summarization, Latent Dirichlet Allocation, and Natural Language Processing.

I. INTRODUCTION

Before the invention of the Internet and the creation of the Web, the vast majority of human conversations were in spoken form, with the only notable, but extremely limited, exception being epistolary exchanges. Some important spoken conversations, such as criminal trials and political debates (e.g.,Hansard, the transcripts of parliamentary debates), have been transcribed for centuries, but the rest of what humans have been saying to each other, throughout their history, to solve problems, make decisions and more generally to interact socially, has been lost.

This situation has dramatically changed in the last two decades. At an accelerating pace, people are having conversations by writing in a growing number of social media, including emails, blogs, chats and texting on mobile phones. At the same time, the recent, rapid progress in speech recognition technology is enabling the development of computer systems that can automatically transcribe any spoken conversation.

The net result of this ongoing revolution is that an ever-increasing portion of human conversations can be stored as text in computer memory and processed by applying Natural Language Processing (NLP) techniques (originally developed for written monologues - e.g., newspapers, books). This ability opens up a large space of extremely useful applications, in which critical information can be mined from conversations, and summaries of those conversations can be effectively generated.

International Journal of Advance Research in Science and Engineering
Volume No.06, Issue No. 09, September 2017
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

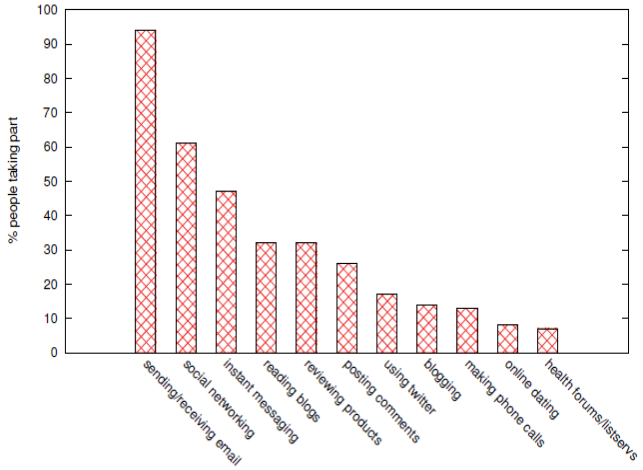

Fig.1: Popularity of various online conversational activities.

This is true for both organizations and individuals. For instance, managers can find the information exchanged in conversations within a company to be extremely valuable for decision auditing. If a decision turns out to be ill-advised, mining and summarizing the relevant conversations may help in determining responsibility and accountability. Similarly, conversations that led to favorable decisions could be mined and summarized to identify effective communication patterns and sources within the company. On a more personal level, an informative summary of a conversation could play at least two critical roles. On the one hand, the summary could greatly support a new participant to get up to speed and join an already existing, possibly long, conversation (e.g., blog comments). On the other hand, a summary could help someone to quickly prepare for a follow-up discussion of a conversation she was already part of, but which occurred too long ago for her to remember the details. Furthermore, the ability to summarize conversations will also be crucial in our increasingly mobile world, as a long incoming message or extensive ongoing conversations could be much more easily inspected on a small screen in a concise, summarized form.

1.1 Summarization

Automated summarization is the method of lowering a textual content file with PC software with the intention to create a summary that keeps the most vital points of the authentic document. Technologies that may make a coherent summary keep in mind variables inclusive of the period, writing style and syntax. Automated statistics summarization is part of machine mastering and records mining. The primary concept of summarization is to discover a consultant subset of the facts, which includes the facts of the entire set. Summarization technology is utilized in a big quantity of sectors in the industry these days. An instance of the use of summarization generation is search engines like Google and Yahoo such as Google. Different examples consist of report summarization, photograph collection summarization, and video summarization. Document summarization attempts to routinely create a consultant précis or abstract of the whole document, with the aid of finding the maximum informative sentences. Similarly, in photograph summarization, the device unearths the maximum representative and crucial (or salient) photos. Similarly, in patron motion pictures one could want to get rid of the uninteresting or repetitive scenes and extract out miles shorter and concise model of the video. That is additionally critical, say for surveillance films, wherein one might need to extract most effective essential occasions inside the recorded video, due to the fact that maximum part of the video can be uninteresting with

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

not anything happening. As the trouble of statistics overload grows, and as the quantity of information will increase, the interest in automatic summarization is also increasing.

Usually, there are processes to computerized summarization: extraction and abstraction. Extractive methods work through choosing a subset of current phrases, terms, or sentences inside the authentic text to shape the summary. In contrast, abstractive methods construct an internal semantic illustration after which uses herbal language generation techniques to create a summary that is what a human might generate. Such a summary might comprise phrases not explicitly gift in the original. Studies into abstractive techniques are an increasingly essential and active research region, however, because of complexity constraints, studies up to now have targeted usually on extractive techniques. In a few utility domain names, extractive summarization makes more experience. Examples of these consist of photo collection summarization and video summarization.

1.2 Extraction-based summarization

On this summarization assignment, the automated gadget extracts items from the entire series, without enhancing the objects themselves. Examples of this include key phrase extraction, where the aim is to pick out character phrases or terms to "tag" a report, and document summarization, in which the intention is to choose whole sentences (without editing them) to create a short paragraph précis. Similarly, in image collection summarization, the machine extracts photos from the collection without enhancing the photographs themselves.

1.3 Abstraction-based summarization

Extraction techniques merely copy the records deemed most critical by means of the device to the précis (for instance, key clauses, sentences or paragraphs), at the same time as abstraction includes paraphrasing sections of the source file. In preferred, abstraction can condense a text more strongly than extraction, however, the applications that may do that are tougher to develop as they require the use of natural language technology era, which itself is a growing discipline.

Whilst a few work has been achieved in abstractive summarization (developing an abstract synopsis like that of a human), most of the people of summarization systems are extractive (selecting a subset of sentences to the region in a précis) [9].

1.4 Applications of Summarization

There are widely two sorts of extractive summarization duties depending on what the summarization application specializes in. The first is common summarization, which focuses on acquiring a frequent summary or summary of the collection (whether or not documents, or units of pix, or videos, news testimonies etc.). The second is question applicable summarization, now and again known as question-based summarization, which summarizes objects precise to a query. Summarizations structures are capable of creating each question applicable textual content summaries and common machine-generated summaries depending on what the user desires.

An instance of a summarization hassle is reported summarization, which tries to automatically produce an abstract from a given record. Once in a while, one might be inquisitive about generating a summary from a single supply record, even as others can use a couple of supply files (as an instance, a cluster of articles at the same subject matter). This trouble is referred to as multi-document summarization. The related software is summarizing news articles. Believe a system, which robotically pulls collectively news articles on a given topic (from the web), and concisely represents the latest information as a summary.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Photograph collection summarization is some other application instance of computerized summarization. It is composed in deciding on a consultant set of photos from a bigger set of photos [10]. A précis on this context is beneficial to expose the most consultant pics of consequences in a photo collection exploration gadget. Video summarization is an associated area, wherein the device automatically creates a trailer of a protracted video. This also has applications in customer or personal movies, in which one would possibly need to, skip the dull or repetitive actions. In addition, in surveillance films, one would need to extract essential and suspicious hobby, even as ignoring all the uninteresting and redundant frames captured.

1.5 Clue word

A clue word in node (fragment) F is a word which also appears in a semantically similar form in a parent or a child node of F in the fragment quotation graph. Applying stemming to the identification of clue words, using Porter's stemming algorithm to compute the stem of each word, and use the stems to judge the reoccurrence.

Fragments (a) and (b) are two adjacent nodes with (b) as the parent node of (a).

Here observing 3 major kinds of reoccurrence : the same root (stem) with different forms, e.g., "settle" vs. "settlement" and "discuss" vs. "discussed" as in the example above. Synonyms/antonyms or words with similar/contrary meaning, e.g., "talk" vs. "discuss" and "peace" vs. "war". Words that have a looser semantic link, e.g., "deadline" with "Friday morning".

Algorithm CWS : Algorithm Clue Word Summarizer (CWS) uses clue words as the main feature for summarization. The assumption is that if those words reoccur between parent and child nodes, they are more likely to be relevant and important to the conversation.

II.LITERATURE REVIEW

TaiwoAyodele et.al in [1] presented the design and implementation of a system to group and summarizes email messages. The system exploits the subject and content of email messages to classify emails based on users' activities and auto generates summaries of each incoming messages. Their framework solves the problem of email overload, congestion, difficulties in prioritizing and successfully processing of contents of new incoming messages and difficulties in finding previously archived messages in the mail box by providing a system that groups emails based on users' activities, and providing summaries of emails.

Taiwo Ayodele et.al in [2] presented Intelligent email summarization system (IESS), encourages users to have summative messages by simplifying the content of the email. Using unsupervised machine learning techniques in combination with automated word and phrases modeler to intelligently provide a précis summary of each email messages is developed to reduce the burden of email users.

Ayman El-Kilany et.al in [3] investigated the problem of extractive single document summarization. They proposed an unsupervised summarization method that is based on extracting and scoring keywords in a document and using them to find the sentences that best represent its content. Keywords are extracted and scored using clustering and dependency graphs of sentences. They tested their method using different corpora including news, events and email corpora. They evaluated their method in the context of news summarization and email summarization tasks and compare the results with previously published ones.

Taiwo Ayodele et.al in [4] presented the design and implementation of a new system to predict whether an email received require a reply, group emails and summarize email messages. The system uses not only subjects and

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

headers fields but also the content of email messages to classify emails based on users' activities and generate summaries of each incoming message with unsupervised learning approach. Their framework tackles the problem of email overload, congestion, difficulties in prioritizing and difficulties in finding previously archived messages in the mail box.

Nidhika Yadav et.al in [5] presented a computationally efficient technique based on sentiments of key words in the text for the purpose of summarization. Sentiment analysis is already being used in various domains for analysis of large scale text data interpretation and opinion mining. The present work shows that sentiment analysis can also be used efficiently for the purpose of text summarization. They have tested their results on the standard DUC2002 datasets, and compared their results with different summarization approaches, viz. Random indexing based, LSA based, Graph based and Weighted graph based methods for different percentages of summarization. The proposed scheme is found to be efficient, in particular for 50% summarization.

Archana N. Gulati et.al in [6] proposed a novel technique for the multi document, extractive text summarization. Also considering the common language in India being Hindi, a summarizer for the same language is built. News articles on sports and politics from online Hindi newspapers were used as input to the system. Fuzzy inference engine was used for the extraction process using eleven important features of the text. The system achieves an average precision of 73% over multiple Hindi documents.

N. Moratanch et.al in [7] ascertained a comprehensive review of extractive text summarization process methods. In this paper, the various techniques, populous benchmarking data sets, and challenges of extractive summarization have been reviewed. This paper interprets extractive text summarization methods with a less redundant summary, highly adhesive, coherent and depth information.

Manisha Gupta et.al in [8] presented a novel approach for text summarization of Hindi text document based on some linguistic rules. Dead wood words and phrases are also removed from the original document to generate the lesser number of words from the original text. The proposed system is tested on various Hindi inputs and accuracy of the system in form of a number of lines extracted from original text containing important information of the original text document.

III. PROBLEM FORMULATION

With the ever increasing popularity of emails, it is very common nowadays that people discuss specific issues, events or tasks among a group of people by emails (Fisher and Moody, 2002). Those discussions can be viewed as conversations via emails and are valuable for the user as a personal information repository (Ducheneaut and Bellotti, 2001). In the base paper, they adopted three cohesion metrics, clue words, semantic similarity and cosine similarity, to measure the weight of the edges. The Generalized Clue Word Summarizer (CWS) and Page-Rank are applied to the graph to produce summaries. Moreover, they study how to include subjective opinions to help identify important sentences for summarization. By the use of CWS accuracy of the system is not so good which lies between 40 to 60 percent. The need of algorithm is felt in which the accuracy could be maximum from the previous algorithm, therefore, we propose a new algorithm using latent dirichlet allocation for Summarizing Emails.

International Journal of Advance Research in Science and Engineering
Volume No.06, Issue No. 09, September 2017
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

IV. PROPOSED RESEARCH METHODOLOGY

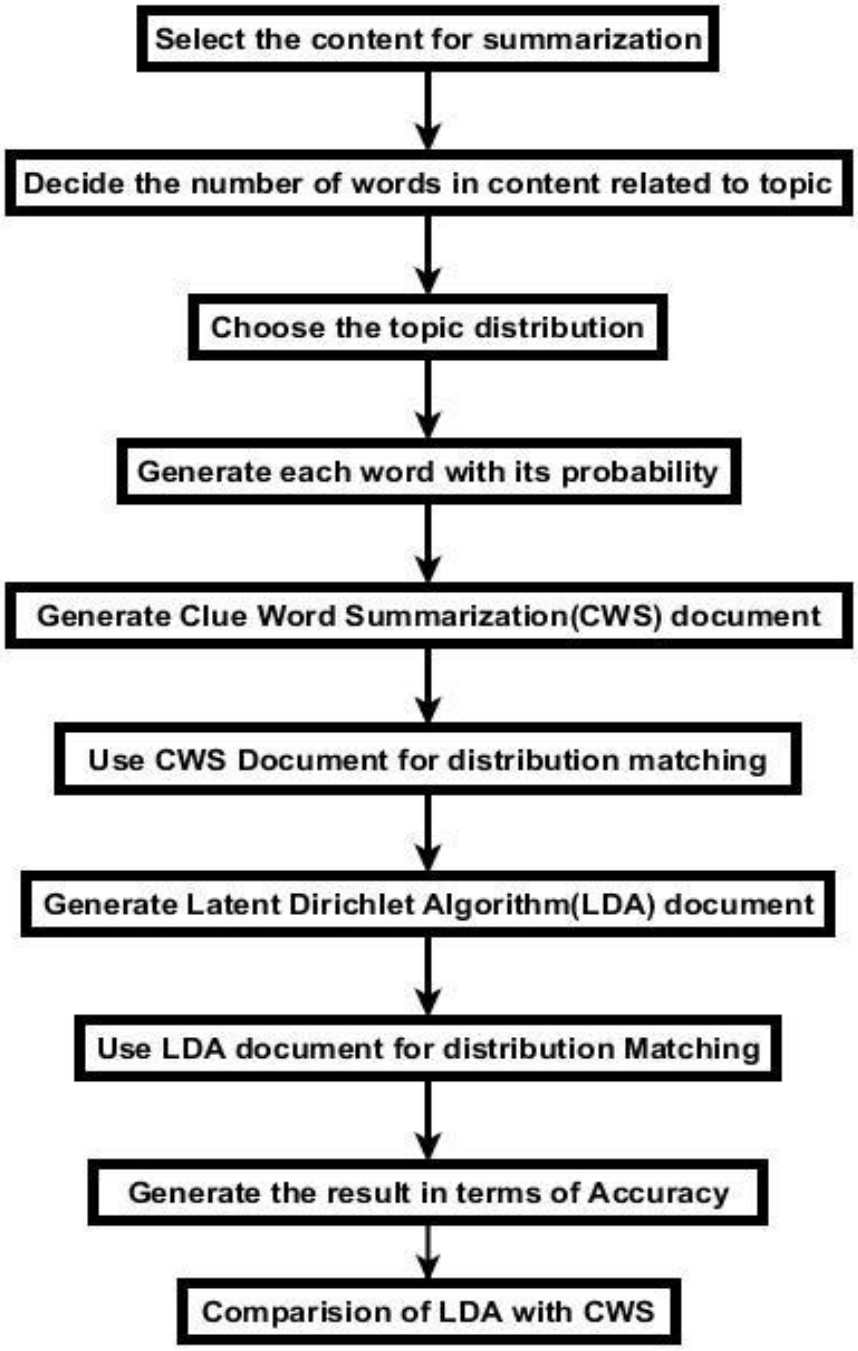

Fig. 2: Proposed Research Methodology

V. SIMULATION AND EXPERIMENTAL

We have proposed Latent Dirichlet allocation (LDA) for email summarization, which will help to remove flaws of Clue Word Summarization (CWS) algorithm. In CWS the emails are used to mark with only one category whereas according to LDA one email may not be categorized into one category only. The email may have more than one categorization. We have designed system for categorization of emails using NetBeans IDE.

International Journal of Advance Research in Science and Engineering
Volume No.06, Issue No. 09, September 2017
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

```
Scanner sc= new Scanner(System.in);
System.out.println("Enter Text: ");
String inputText = sc.nextLine();

categoriesValues = new ArrayList<String>();
categoriesValues.add("Orange");
categoriesValues.add("Apple");
categoriesValues.add("Mango");
categoriesMap.put("Fruits", categoriesValues);

categoriesValues = new ArrayList<String>();
categoriesValues.add("Lion");
categoriesValues.add("Tiger");
categoriesValues.add("Elephant");
categoriesMap.put("Animals", categoriesValues);
```

Fig. 3: Parts of Code developed using NetBeans IDE

The basic interface of NetBeans and Code developed is shown in figure 3. Here we code for detection of two categories of emails named as Animals and Fruits which are having quantities like Lion, Tiger, Elephant, Mango, Apple, and Orange. All these quantities are case sensitive.

```
Enter Text:
```

Fig. 4: Output Window

The output window pops below when we run the code is shown in figure 4. The output window shows message "Enter Text:" Where we have to enter text for categorization of Email Contents.

```
Enter Text:
In most case the the common fruits are Apple and Orange.
Paragraph is of following category:
Fruits
BUILD SUCCESSFUL (total time: 33 seconds)
```

Fig. 5: Output Window for fruits Category

The output window for fruits category is shown in figure 5. When we run the code and Enters Text: "In most case the common fruits are Apple and Orange". The Result shows that "Paragraph is of following Category: Fruits". At the end of output results, it also shows the time taken for providing output results. These Results are for CWS Algorithm.

```
Enter Text:
Lion and Tiger are the most dangerous animals.
Paragraph is of following category:
Animals
BUILD SUCCESSFUL (total time: 9 seconds)
```

Fig. 6: Output Window for animals Category

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

The output window for animals category is shown in figure 6. When we run the code and Enters Text: "Lion and Tiger are the most dangerous animals". The Result shows that "Paragraph is of following Category: Animals". At the end of output results, it also shows the time taken for providing output results. These Results are for CWS Algorithm.

```
Enter Text:
Lion is the king in the animals and likewise the Apple is the king of fruits.
Paragraph is of following categories:
Animals
Fruits
BUILD SUCCESSFUL (total time: 7 seconds)
```

Fig. 7: Output Window for animals and fruits Category

The output window for animals and fruits category is shown in figure 7. When we run the code and Enters Text: "Lion is the king of the animals and likewise the Apple is the king of fruits". The Result shows that "Paragraph is of following Category: Animals Fruits". At the end of output results, it also shows the time taken for providing output results. These Results are for LDA Algorithm.

```
Enter Text:
Animals play a very important role in the human life.
Paragraph doesnot belong to any category.
BUILD SUCCESSFUL (total time: 1 minute 40 seconds)
```

Fig. 8: Output Window for no Category

The output window for no category is shown in figure 8. When we run the code and Enters Text: "Animals play a very important role in the human life". The Result shows that "Paragraph does not belong to any Category". At the end of output results, it also shows the time taken for providing output results. There are no clue words which are meant for categorization, therefore, this content does not belong to any Category.

VI. EXPERIMENT RESULT

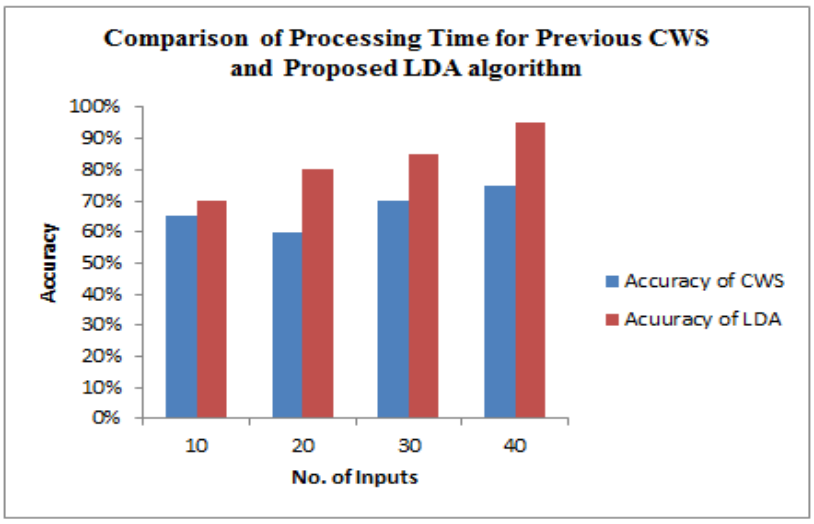

Fig. 9: Accuracy of Algorithm

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

The accuracy of previous algorithm CWS and Proposed Algorithm LDA is shown in figure 9. The accuracy of the algorithm is calculated by testing both the algorithm with no. of content inputs. The Accuracy of CWS is 75% whereas the accuracy of LDA is 95%.

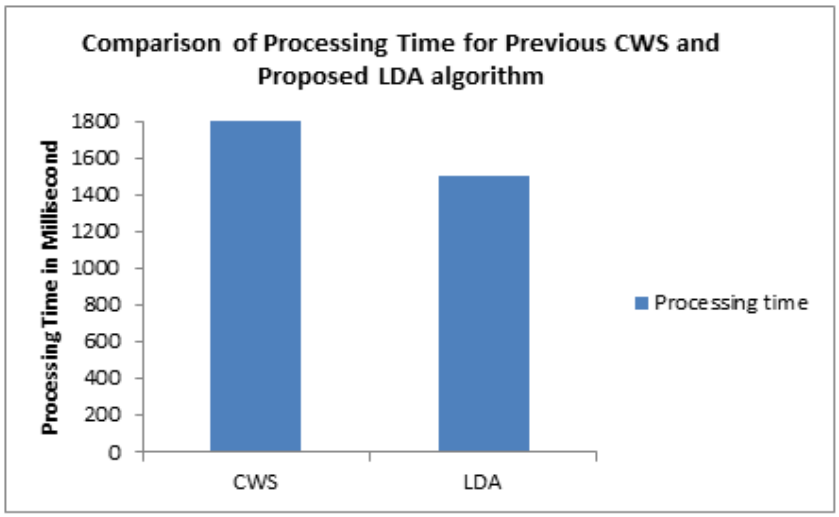

Fig. 10: Processing Time of Algorithm

The processing time of previous algorithm CWS and Proposed Algorithm LDA is shown in figure 10. The processing time of the algorithm is calculated by testing both the algorithm with no. of content inputs. The processing timeof CWS is 1800 milliseconds whereas the processing timeof LDA is 1500 milliseconds.

VII. CONCLUSION AND FUTURE SCOPE

Summarizing email conversations is challenging due to the characteristics of emails, especially the conversational nature. Most of the existing methods dealing with email conversations use the email thread to represent the email conversation structure, which is not accurate in many cases [11]. We are presenting an approach which will help to remove flaws of Clue Word Summarization (CWS) algorithm. In CWS the emails are used to mark with only one category whereas according to LDA one email may not be categorized into one category only. The email may have more than one categorization. We have designed system for categorization of emails using NetBeans IDE. The accuracy of the algorithm is calculated by testing both the algorithm with no. of content inputs. The Accuracy of CWS is 75% whereas the accuracy of LDA is 95%.The processing timeof CWS is 1800 milliseconds whereas the processing time of LDA is 1500 milliseconds.

REFERENCES

[1] T. Ayodele, S. Zhou and R. Khusainov, "Email Grouping and Summarization: An Unsupervised Learning Technique," 2009 WRI World Congress on Computer Science and Information Engineering, Los Angeles, CA, 2009, pp. 575-579.

[2] T. Ayodele, S. Zhou and R. Khusainov, "Intelligent email summarization system (IESS)," 2010 International Conference on Information Society, London, 2010, pp. 330-335.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

[3] A. El-Kilany and I. Saleh, "Unsupervised document summarization using clusters of dependency graph nodes," 2012 12th International Conference on Intelligent Systems Design and Applications (ISDA), Kochi, 2012, pp. 557-561.

[4] T. Ayodele and S. Zhou, "Applying Machine learning Algorithms for Email Management," 2008 Third International Conference on Pervasive Computing and Applications, Alexandria, 2008, pp. 339-344.

[5] N. Yadav and N. Chatterjee, "Text Summarization Using Sentiment Analysis for DUC Data," 2016 International Conference on Information Technology (ICIT), Bhubaneswar, 2016, pp. 229-234.

[6] A. N. Gulati and S. D. Sawarkar, "A novel technique for multi document Hindi text summarization," 2017 International Conference on Nascent Technologies in Engineering (ICNTE), Navi Mumbai, 2017, pp. 1-6.

[7] N. Moratanch and S. Chitrakala, "A survey on extractive text summarization," 2017 International Conference on Computer, Communication and Signal Processing (ICCCSP), Chennai, 2017, pp. 1-6.

[8] M. Gupta and N. K. Garg, "Text Summarization of Hindi Documents Using Rule Based Approach," 2016 International Conference on Micro-Electronics and Telecommunication Engineering (ICMETE), Ghaziabad, 2016, pp. 366-370.

[9] En.wikipedia.org. (2017). Automatic summarization. [online] Available at: https://en.wikipedia.org/wiki/Automatic_summarization [Accessed 13 Jun. 2017].

[10] Jorge E. Camargo and Fabio A. González. A Multi-class Kernel Alignment Method for Image Collection Summarization. In Proceedings of the 14th Iberoamerican Conference on Pattern Recognition: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications (CIARP '09), Eduardo Bayro-Corrochano and Jan-Olof Eklundh (Eds.). Springer-Verlag, Berlin, Heidelberg, 545-552. doi:10.1007/978-3-642-10268-4_64.

[11] H. Aaron, Y. Jen-Yuan, "Email thread reassembly using similarity matching". In Proceedings of the Third Conference on Email and Anti- Spam (CEAS), 2006.