Multi Document Text Summaization using Machine Leaning

1. *Vishal N. Hemnani, Computer Dept, KJCOEMR, Pune*
2. *Monisha K. Prasad, Computer Dept, KJCOEMR, Pune*
3. *Sanemahdi A.S Aland, Computer Dept, KJCOEMR, Pune*
4. *Bhagyashri Vyas, Computer Dept, KJCOEMR, Pune*

Abstract

Summarization process is always needs more precision and time to yield the best results. This is due to the vastness of the data, complexities in the narration of the documents and constrained time boundaries. So the task of Document summary extraction is always full of complexities. So this is the key point in many sections of society where work is lagging behind like court rooms, Academic carrier and many more due to unavailability of the document summary on time. Natural language processing and Machine learning always plays a vital role in providing the summary of the documents, but accuracy is always a big question. So this research article concentrates on extraction of the semantic Summary for the input of multiple documents. This paper introduces using of the Gaussian distribution model and Fuzzy classification along with the natural language processing technique to yield well semantic summary for the given input of the multi documents.

Keywords: *Fuzzy Classification, NLP, Feature Extraction, Gaussian Distribution.*

I. INTRODUCTION

In today's world, there has been an unprecedented increase in the number of books are published material. News, for instance, can be several pages long for a single day. As in today's world, with the increased pace of living, with people working and traveling for most of the time. Most of them can't be bothered to read as most of them are averse to the long texts and multiple pages worth of reading, even if they want to, they won't be able to get the time for it.

This is not as good as the long length of the article and the busy schedule of the person work hand-in-hand and distance the person away. This is the reality for the majority of the people as they are denied the luxury of reading from their busy schedules. This is unacceptable and there should be a way to make the text more concise and get the point across to the reader efficiently in a stipulated time.

Therefore, we need to devise a method for summarising the text in smaller units, so that it can be read by people pressed for time in a short duration. This is very good for people who have long work hours and do not get the time to get up to date with various current issues or a book they have been trying to complete. The summarised text needs to be concise and should not omit anything and contain all the important information from the parent text.

The process of retrieving important and relevant information from a large set of values. This is a very important feature in the process of data mining, where large sets of data are used to extract the data that is relevant to the application that is concerned. Feature extraction is one of the most essential components that can enable a far superior insight about a process that is being analyzed.

Feature extraction is being widely used in applications concerning machine learning. it is really useful in the area of image processing as it can extract the useful areas from an image and reduce the computational requirement of the system. As utilizing the whole image for the processing can be very computationally taxing and would take up increasingly more resources on something that is not even useful. Feature Extraction is useful because most of the time, too much information is not useful and reduces the effectivity of the algorithm. Relevant data even though less in number is highly useful for the algorithm as it does not require more resources to be allocated for data that cannot be put into productive use. As often the datasets are large and require a lot of processing to be useful.

The feature extraction is one of the most essential aspects of Machine learning as it saves a lot of time and resources so that it can be utilized in an organized and efficient manner to derive an immense amount of productivity from the particular dataset. Normal Distribution is one of the most essential tools in the area of Statistical Mathematics. It can be used in a variety of

applications ranging from the distribution of weights and their corresponding heights in a group of students or it can be also used for the representation of measurement error in various fields. These values are considered as they correspond to the pattern of normal distribution.

The normal distribution is also known as Gaussian Distribution. The name Gaussian Distribution is named after its creator Carl Friedrich Gauss, a pioneer in the area of mathematics and statistics and his contribution to the world of statistics in the early 1800s. The distribution indicates the distribution of values of a certain variable corresponding to another variable. The distribution has been widely used since its inception.

The normal distribution is predominantly a probability function that corresponds to the distribution curve. This distribution curve is represented as a bell-shaped curve that is a very characteristic curve and has a very revered status in the statistical community. The Gaussian curve as it is popularly known is symmetrical and representative of the values and their frequencies in the graph.

The graph is a characteristic bell-shaped and its values are corresponding values. The values in the center corresponding to the peak of the bell shape are the most frequent values, as depicted with the rise in the graph as the values and the amount of time they have been represented. The values on the fringes are the small vales and have been depicted less frequently. There are two fringes on the curve, one on each side of the curve forming the bell shape. Fuzzy is a concept that relies on the indecisiveness or the reduced clarity in the decisions or values of a particular entity. The Fuzzy values are most of the time vague or very inconsistent in nature as they are bound to be. The values in fuzzy logic are meant to be inconclusive and do not take or represent a particular clarity in their representation. The fuzzy values are the representation of the real world, where there is a very small probability of something being completely clear.

The real world is not decisive and it is a very bleak possibility of encountering a complete black or complete white value in nature. As everything in nature is not completely true or false and lands up somewhere in the middle. But most computers are binary, which means they only accept inputs that are either 1 or 0, true or false. There is a clear distinction between the two which is why fuzzy values cannot be accepted by a computer. As most of the computation and complex calculations are handled by the computer, it is imperative that it is able to comprehend the inconsistency of the binary numbers in representing nature. Therefore, is unable to understand the importance of the gradient that allows for the system to acknowledge values other than 1s and 0s. the Fuzzy logic enables the computers to recognize inputs having values in between.

The fuzzy logic helps the computers understand values that are not completely true but also not completely false, but somewhere in the middle. This enables a much more responsive and sensitive application of various elements into the computer. This is useful in the sense of providing the computer a way to interact with the world a lot more comprehensively and understand it rather precisely. Fuzzy logic is really useful in applications which require natural responses, some-what human-like responses, rather than the values 1 and 0. Control systems require a very precise input and handling techniques that rely on the inconsistencies in the values residing in the middle of the 1s, and 0s. these applications are highly critical and fuzzy logic can help automate them for a much efficient approach.

This research paper dedicates section 2 for analysis of past work as literature survey, section 3 deeply elaborates the proposed technique and whereas section 4 evaluates the performance of the system and finally section 5 concludes the paper with traces of future enhancement.

II. LITERATURE SURVEY

S. Rahimi [1] expresses that due to technological advancements and the age of the internet, there has been an increasing amount of data being generated. The data has been generated through different types of devices and a lot of people can publish their data online. Therefore, it is really difficult to find reliable information online. As it is very difficult to manually read all the documents therefore, automatic text summarization is one of the most important technologies for the credibility of information.

P. Krishnaveni [2] states that there has been a huge growth in the size of the internet with different devices being connected and generating a large amount of data. Some of this data is in the form of text, very large texts. As most of these texts cannot be evaluated by a human being fast enough. Therefore, the authors present a technique for the automatic summarization of text, the automatic summarization of the text has been an interesting topic since the 90s, but couldn't be perfected due to its many flaws and have been fixed in this iteration.

E. Reategui explores the area of text summarization with the help of graphs as the primary data structure. This is one of the easiest techniques to summarize any texts and can be taught to young students easily. this technique utilizes the graph data structure to extract the summary of the given text. The technique was trained by 20 students that were asked to write the

summary without the use of graphs and then the students were made to write the summary with the assistance of the graphs. The students showed remarkable improvement with the graph assisted technique. [3]

M. Afsharizadeh [4] explores the rapid increase in the amount of data being created online and its uses as humans are now facing huge problems in analyzingand reading this data. there is a large influx of information today and this is increasing day-by-day. This is problematic because we have not yet developed ways to effectively analyze this information. Therefore, the authors present an innovative text summarization technique which utilizes 11 different feature extraction methods that can summarize a sentence very effectively.

J. Xiao-Yu [5] introduces two different technique for the purpose of text summarization which is helpful in increasing the quality and efficiency of the text summarization technique. The first method is based on the feature selection and segregation, which is useful for the selection of important words in the sentence and segregates it together for a better understanding of the technique. The second method utilizes the KNN algorithm for the selection based on the weights given to the sentences. These two techniques can be utilized together to increase accuracy and improvement in the performance of the text summarization approaches.

Z. Pei-Ying states that there has been an enormous increase in interest in the area of text-summarization as a lot of researchers have been doing active research on this topic. Automatic text summarization is one of the most useful technique for a quality summary of a document and plays an important role in the information retrieval of large documents. The authors have proposed a technique for the summarization of the text through 3 steps, the first one involves the segregation of sentences into clusters based on the semantic distance of the document, and then the distance is calculated based on the multi-feature combination method and lastly, the summary is extracted using the extraction tools. [6]

V. Dalal [7] professes that there has been an information explosion following the creation and success of the internet. As there has been a steady increase in the information on the internet, it has become highly difficult to read and analyzea largenumber of texts. Therefore, to provide a solution for this, the authors present a text mining technique for the summarization of text. The proposed method manages an obstructive approach that analyses the text and produces the summary. The proposed methodology has been significantly better than the other approaches.

H. Huang [8] comments on the different methods for gathering or retrieving data, as there are two types of gathering structures, namely, structured and unstructured. The first structured method is the process through which data is gathered using social sensing techniques which includes temperature sensing and GPS devices that generate data in the numerical form and subsequently converted to its statistical form. The second technique for a gathering of the unstructured data in the form of the text of images. The authors have efficiently utilized both the techniques for the purpose of automatic text summarization for military purposes.

J. Zenkert explains that the normal text-based information involving the use of natural language is very difficult for the computer to analyze and process effectively due to the very high complexity of the human language and understanding the relationship between text and information. Therefore, the information on the internet has been very difficult to summarize as the relationship between the text and the information it is conveying is extremely difficult for the computer to understand. To ameliorate this effect, the authors need to identify the relationship between each sentence and the subsequent words that convey the actual idea of the document for the summarization. [9] .

C. Wang [10] demonstrates that there has been a lot of development in the area of automatic text summarization techniques as there has been an exponential growth of information on the internet and has been readily available everywhere. Therefore, the authors have presented HowNet, which is a text summarization technique that extracts all the relevant information from the text and forms a short summary of the large text. The technique has been verified to be highly effective and provides an efficient summary.

A. Pal [11] states that there has been unprecedented growth in the area of text mining and summarization due to the large influx of information on the internet. There have been a number of techniques for the summarization, such as the tagged rules, the format and position of text,etc. have been used for this purpose. The authors present a technique for summarization with the use of unsupervised learning. The first step in this technique is to extract and assign the weights to all the sentences. The next step the weighted sentences are then arranged in descending order and each of them is given a percentage. The technique has been highly effective in summarization of large texts.

S. Chakraborti explains the process and idea that summarization is a technique for the extraction of the relevant and useful information from a large document or a text file. There has been an exponential growth in the area of automatic text summarization as it is very beneficial for businesses and has helped countless managers in accelerating the growth of their company. To extract the relevant information the authors in this paper have utilised a plethora of methods, such as topic detection, clustering competitor intelligence and decision support systems. [12]

A. Bagalkotkar [13] expresses that the process of automatic text summarization plays an integral part in the reduction of human efforts. The authors in this paper have designed an approach for the automatic text summarization for a web document with the help of statistical NLP techniques that generate a text summary. The text summary is totally dependent on the number of words and the number of terms in the sentences. This is used in the extraction of useful data effectively from the input document. The proposed technique has been verified by the researchers to be useful and highly effective.

III PROPOSED METHODOLOGY

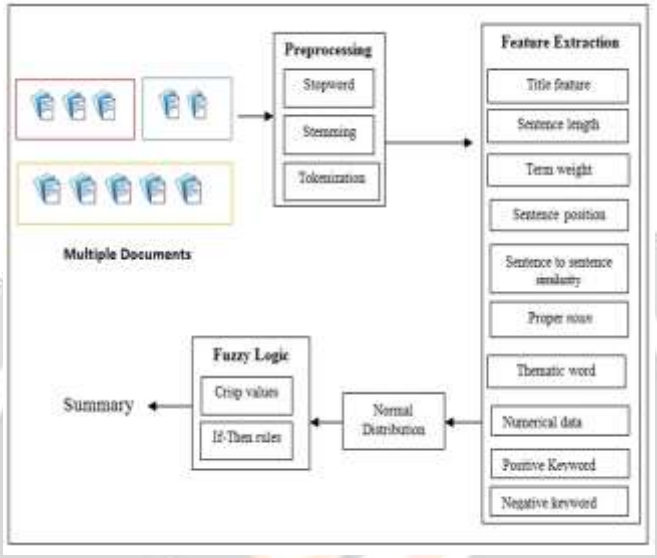

Figure 1: Overview of the Multi Document Summary Extraction

The proposed system for document summary extraction is narrated with the below mentioned steps. And also pictorially it is shown in figure 1.

Step 1: Document Input - This is the initial step of the system where multiple text documents of the extensions like pdf, txt and doc are put in a folder . This folder is given as the input to the system, once these documents are given as the input all the documents are read in a string and they concatenate to form a single string. Pdf files are read using he itext pdf API and doc extension file are read using the Apache POI API.

Step 2: Preprocessing - The concatenated Single string from the previous step is subjected to preprocessing. By doing this the string is getting rid of the unwanted text that really does not add any meaning to the text and also it makes the string more lightweight, which will be the added advantage to reduce the complexity of the process. Before performing the preprocessing step, the single document string is split on ". " Character to retrieve the sentences and then these sentences are stored in a list. The preprocessing includes the four steps as mentioned below

Special Symbol removal - Here in this step all the special characters from the strings are shred off except the space character and ". " .

Tokenization - Here special symbol removed string is split on the " . " to get the string into a list. Then each of this string is subjected to Stopword removal and Stemming process.

Stopword Removal- This is the step of removing all the conjunction words from the sentences like and, of, the, from, to etc.. By Removing these words the semantics of the string remains intact.

Stemming- Stemming technique brings down any word to its base form and this makes the word more light weight and meaning of the word is remain unchanged.

Step 3- Feature Extraction - This is the crucial step of the proposed system where each of the preprocessed Sentence is subjected to extract its feature to store in a list. The process is explained below with all the futures.

Title Sentence : The Title sentence is the very first sentence that eventually holds the introductory part of the document. So all other remaining sentences are compared with the title sentences to extract this feature according to given below equation 1.

$$\mathrm{T}f = \frac{\text{Frequency of Title sentence words in the Sentence}}{Sentence\ Length}$$

Sentence Length: The length of the sentence is always contains more information than the other sentences. So it is always important to consider the length as one of the feature and it can be extracted using the following equation 2.

$$\mathrm{SL}f = \frac{\text{Sentence Length}}{Biggest\ Sentence\ Length}$$

Term Weight - This feature in a sentence represents the number of important words in it. So it can be extracted with the below shown equation 3.

$$\mathrm{TW}f = \frac{\text{Frequency of Top 10 words in a sentence of the Document}}{Sentence\ Length}$$

Sentence Position - The position of the sentence in a given document always plays an important role. As the position reveals the quality of the content that a sentence is having. To measure this first five sentences of the document is given a score as 1, 0.8, 0.6 , 0.4 ,0.2 and all other sentences after that are given a score of 0.

Sentence Similiarity - The sentence similarity of a sentence always tells about the correlation between the other sentences. And this can be evaluated with the following equation 4.

$$\mathrm{SS}f = \frac{\sum_{i=1}^{n} \text{Similarity with other sentences}}{Maximum\ sentence\ Similarity}$$

Prope Noun - Here Proper noun in a sentence is evaluated using a dictionary of Workbook that contain around 1,20,000 words of all alphabets. If any Word is not present in the dictionary, then it is referred as the Proper noun. And the proper noun in a sentence can be measured by the following equation of 5.

$$\mathrm{PN}f = \frac{\text{Frequency of proper noun in the Sentence}}{Sentence\ Length}$$

Thematic Words - This feature in a sentence represents the most meaningful words in it. This can be expressed by the following equation 6.

$$\mathrm{Th}f = \frac{\text{Frequency of Top 20 words in a sent of the Document}}{Sentence\ Length}$$

Numerical Data: Numerical data show the statistical strength of the sentence. And it can be evaluated by the below shown equation 7.

$$\mathrm{N}f = \frac{\text{Frequency of Numerical data in the Sentence}}{Sentence\ Length}$$

Positve and negative Data: Positive and negative data shows the amount of good and bad keywords in the document respectively. Which helps to understand the motto of the document semantically. This is evaluated using the bag of words technique. And this is represented using the equation 8 and 9.

$$\mathrm{PD}f = \frac{\text{Frequency of Postive Words in the Sentence}}{Sentence\ Length}$$

$$\mathrm{ND}f = \frac{\text{Frequency of Negative Words in the Sentence}}{Sentence\ Length}$$

Step 4-Gaussin Distribution - Here all the 10 features of sentences are summed up to convert into a single factor to call as the optimized factor of sentences. Now Proposed model evaluates the mean and standard deviation of optimized factors of all the sentences based on the equation 10 and 11. Once this Mean and standard deviation is evaluated, then the quality ranges of this optimized factors evaluated using equation 12. Based on this range most distributed sentences are identified and kept aside to provide along with the summary result.

$$\mu = \frac{(\sum_{i=1}^{n} xi)}{n} \_\_\_\_\_(10)$$

$$\delta = \sqrt{\frac{1}{N}} \sum_{i=1}^{n} (xi - \mu)^2 \_\_\_(11)$$

$$Fr = (\mu - \delta)\ TO\ (\mu + \delta) \_\_\_\_(12)$$

Step 4-Fuzzy Classification - Here in this section of proposed methodology the optimized factors of the sentence are classified according to the 5 Fuzzy crisp values ranging from the lowest values to the highest. The crisp values are labelled as VERY LOW, LOW, MEDIUM, HIGH and VERY HIGH.

Based on these fuzzy crisp values IF THEN rules are applied to segregate all the sentences in these ranges. Then, according to the user defination for the summary level the sentences are extracted above the lowest level from the original sentence list. Sentences which are selected through Gaussian distribution are also added into this to optimize the Text summary.

IV RESULT AND DISCUSSIONS

The Proposed system of Multi document text summarization is developed in Windows based machine by using Java as the programming language. To develop the application Netbeans is used as the standard IDE. The laptop which is used for this purpose is equipped with Core i5 Processor and 6GB of Primary memory. To prove the effectiveness of the proposed model in text summarization process some experiments are conducted as described below.

Root mean square Error (RMSE) is measured to estimate the error rate of any system. Here in this experiment RMS is used to measure the Error rate between the Expected Summary and Obtained Summary of the input text documents. This can be represented by the following equation.

$$\mathrm{RMSE}_{fo} = [\sum_{i=1}^{N} (z_{fi} - z_{oi})^2 / N]^{1/2}$$

Where
$\sum$ - Summation
$(Zfi - Z_{oi})^2$ - Differences Squared for the Smma
N - Number of samples or Trails

Input Sentenecs	No of Expected Sentences in Summary	No of Obtained Sentences in Summary	MSE
32	8	6	4
47	13	11	4
84	22	18	16
97	29	19	100
108	32	23	81

Table 1: Mean Square Eror Reading

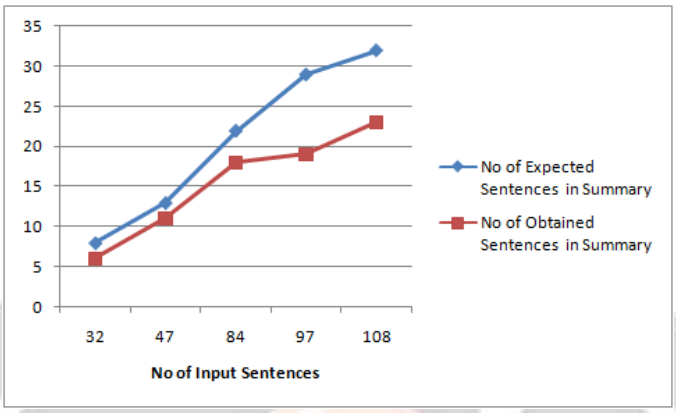

Figure 2: No Expected of Sentences Expected Summary V/s No of Snetences in Obtained Sumamry

The Table 1 indicates the reading of MSE (Mean Square Error), And the average of this mean square error is around 41, This yields the RMSE of around 6.4. The Obtaned RMSE is considered as good in Natural language Processing. Because the proposed model deals with the complete unstructured data in textual format to provide the best possible text summarization.

V CONCLUSION AND FUTURES SCOPE

The concept of text summarization needs more attention as the extraction of the sentences are purely depend on the narration of the sentences. Proposed model handled all the sentences with the vast category of the features as explained in the past section. The using of Gaussian distribution eventually catalyzes the quality of the summary along with the Fuzzy Classification. Evaluation of the proposed model indicates that that deployed application yields around RMSE of 6.4 that is actually a better result while dealing with unstructured text data.

In the future this process can be enhanced to work for huge documents which contains thousands of lines using the Distributed computing. And also this system can be developed as the ready made API or a web service that can help a student or other summary seekers.

REFERENCES

[1] S. Rahimi, A. Mozhdehi and M. Abdolahi, “An Overview on Extractive Text Summarization”, IEEE International Conference on Knowledge-Based Engineering and Innovation, 2017.

[2] P. Krishnaveni and Dr. S. R. Balasundaram, “Automatic Text Summarization by Local Scoring and Ranking for Improving Coherence”, Proceedings of the IEEE 2017 International Conference on Computing Methodologies and Communication (ICCMC), 2017.

[3] E. Reategui, M. Klemann and M. David Finco, “Using a Text Mining Tool to Support Text Summarization”, 12th IEEE International Conference on Advanced Learning Technologies, 2012.

[4] M. Afsharizadeh, Hossein Ebrahimpour-Komleh and Ayoub Bagheri, “Query oriented Text Summarization using Sentence Extraction Technique”, 4th International Conference on Web Research (ICWR), 2018.

[5] J. Xiao-Yu, F. Xiao-Zhong, W. Zhi-Fei, and J. Ke-Liang, "Improving the Performance of Text Categorization using Automatic Summarization", International Conference on Computer Modeling and Simulation, 2009.

[6] Z. Pei-Ying, "Automatic text summarization based on sentences clustering and extraction", 2nd IEEE International Conference on Computer Science and Information Technology, 2009.

[7] V. Dalal and L. Malik, "A Survey of Extractive and Abstractive Automatic Text Summarization Techniques", International Conference on Emerging Trends in Engineering and Technology, 2013.

[8] H. Huang, S. Anzaroot, Heng Ji, H. Khac Le, D. Wang, and T. Abdelzaher, "Free-form Text Summarization in Social Sensing", ACM/IEEE 11th International Conference on Information Processing in Sensor Networks (IPSN), 2012.

[9] J. Zenkert, A. Klahold and M. Fathi, "Towards Extractive Text Summarization using Multidimensional Knowledge Representation", IEEE International Conference on Electro/Information Technology (EIT), 2018.

[10] C. Wang, L. Long, L. Li, "HowNet Based Evaluation for Chinese Text Summarization", International Conference on Natural Language Processing and Knowledge Engineering, 2008.

[11] A. Ranjan Pal and D. Saha, "An Approach to Automatic Text Summarization using WordNet", IEEE International Advance Computing Conference (IACC), 2014.

[12] S. Chakraborti and S. Dey, "Multi-Document Text Summarization for Competitor Intelligence: A Methodology", International Symposium on Computational and Business Intelligence, 2014.

[13] A. Bagalkotkar, A. Khandelwal, S. Pandey and S. Kamath S, "A Novel Technique for Efficient Text Document Summarization as a Service", Third International Conference on Advances in Computing and Communications, 2013.

*****