Text Summarization Using Fuzzy Classification and Normal Distribution

Sandip Patil[1] Anuja Patil[2] Genish Gajjar[3] Suvarna Satkar[4]
[1,2,3,4]Department of Computer Engineering
[1,2,3,4]G.H.Raisoni College of Engg & Mangement, Pune, India

Abstract— There has been extensive research that has been conducted in the field of Text Summarization, to decrease the time taken and increase its precision. The research has been going on for decades to solve this extremely complex problem with innumerable permutations and combinations with vast data and time constraints. This is one of the most crucial subjects for research as there is an immediate need for systems that can effectively summarizethe text in a stipulated amount of time. The most affected applications are Academic careers and courtrooms. The summary of documents is derived by utilizing Natural Language Processing and Machine Learning, with varying amount of accuracy. Therefore, to increase the accuracy of summaries, this article proposes a Fuzzy Classification model with Gaussian distribution to extract a semantically sound summary of a given input of multiple documents.

Key words: Fuzzy Classification, NLP, Feature Extraction, Gaussian Distribution

I. Introduction

Feature Extraction is an extensive field which is used to derive features from a list of elements in a dataset. It is one of the most important and useful processes that make up the idea of Machine Learning. feature extraction is one of the primary techniques that are used for the process of classification. Feature Extraction features as the initial steps that are used predominantly for the purpose of learning.

The method of feature Extraction is used so that relevant information can be extracted from the big pool of data. This is very necessary because there are a lot of useless points and elements in the data which makes it a problem for processing. Extracting the relevant data makes it an easier process for the algorithm to work.

As large datasets require a lot of processing to be done on it to make any productive use of the data. Feature Extraction processes the data for the algorithm beforehand so that it is efficient for the algorithm to work on the data productively. Useless data would take a long time and precious computational power of the system to compute and generate an output. But if the data is classified earlier with the help of feature extraction, this reduces the amount of time wasted in performing the algorithm on useless data.

Due to large computing resources that are being used in large scale applications of the machine learning systems, every useless data being processed, increases the overall time taken and moreover, the loss in resources that are being used to provided such high computational power. This computational might could be utilised more efficiently in a more valuable and useful applications. Therefore, feature extraction is one of the most essential components and it also enables us to achieve high levels of efficiency.

Normal distribution refers to the distribution that is one of the most essential of all distributions that are studied in the area of statistics. As it can be used for a variety of applications and natural phenomena such as heights, measurements error etc. as these values follow the pattern of a normal distribution. Normal distribution is also another name for Gaussian Distribution.

The Normal Distribution or Gaussian Distribution dictates or predicts how certain values of a variable can be distributed. The distribution acts like a probability function which guides the distribution curve. The Gaussian Distribution therefore, has a characteristic Bell-shaped curve. And thus, the normal distribution is also predominantly known as the Bell Curve. As it is very revered in statistics.

The normal distribution probability function when plotted on a graph yields a symmetric curve that corresponds to a bell shape. The bell signifies the flow or the distribution of the values along the curve of the graph. Where the central values corresponding to the peak of the bell shape are the observations that have the largest values and are clumped together. These are similar values that correspond to the larger set of values.

The values on the fringes of the curve correspond to the less frequent values. Unlike the values in the centre of the bell curve, the values on the fringes are less probable out of all the values depicted on the distribution. There are two sets of such values on the either end of the bell-shaped graph and are characterised with a low frequency of occurrence.

Fuzzy in general, would mean the indecisive or unclear values for a particular entity. This is referred to as vague or somewhat inconsistent values. This is true for most of the real world, as there is a very less chance to have a clear black and white in nature. As most things have a gradient and it generates some kind of inaccuracies which are quite evident.

But all computers and computing devices work or run on binary numbers. Which have a clear distinction between 1’s and 0’s, or true or false, this contradicts a lot of naturally occurring phenomenon which might have values in-between which are difficult to convey or process for a computing machine. Therefore, to have an apt representation of these elements, gave rise to the Fuzzy logic. Fuzzy logic is highly valued in environments which are natural and require a degree of precision with values between 1 and 0. Such applications are generally control systems that benefit from such human like observation of critical components and thus is really helpful in managing certain situations.

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] Now in recent years there is a large amount of data is available on different websites and on the different type’s electronic books on different researches so it's very hard to find the reliable and the accurate data. It is very difficult for humans to read large documents in very less amount of time. The text Summarization method is becoming one of the key methods to research nowadays. There are different approaches to rate important sentences and the different parameter of summarization is used.

P. Krishnaveni [2] With the increasing the growth of the internet, there is a large amount of data available to read, but it's hard to read the data so to shortcut to this text summary is an option available. From early 90's this topic as became the in the interesting topic for the researcher's but every time there is some flaws in the research there is a loss of some data which is important to the user. Automatic Text summarization takes a text file as an input and preserves data from the content and overall meaning.

E. Reategui [3] This paper proposed the method of text summarization in this method they extract the graph on the basis of the text summarization. This method of text summarization is very easy to understand for students.
This method based on the graph as graphic organizers this helps students to get the idea fast and on the basis of the graph they can write the text. This method was experienced on the 20 students of class the text file was given and they're told to write the summary of the text file it was to get the idea on the text. But this student later was told to write the summary on the basis of the graph and there successful to get the idea.

M. Afsharizadeh [4] There is a massive growth in the online information in the last 4 to 6 years. Humans face the problem to read and analysis data, thus the text summarization enables to access the summary of the text file in the given shortage of time. In this paper there are 11 different features are used to extract the best features from each sentence. These 11 features make this paper more effective and most improved summary is generated.In this paper, the ROUGE criteria are used and it is successful

J. Xiao-Yu [5] In this paper the author has published two approaches of text summarization which improves the quality and the effective summary of the given document. The first approach is being directly used for feature selection and categorization and the second approach is used the select and weight the feature of the given document by using the KNN algorithm. By using these two approaches the result of the summary gives all the important feature of the given text ND it also increases the speed of the performance of text categorization.

Z. Pei-Ying [6] In recent years there has been an interesting topic for the researchers to research on the Automatic Text Summarization it plays an important role in information retrieval and the text classification. In this paper, it contains the three steps for text summarization first steps are the to divide this sentence into clusters according to the semantic distance of the given document and the second step is to calculate the accumulated distance based on the multi-feature combination method and then they choose some sentences by using the extraction rule.

V. Dalal [7] Due to the World Wide Web has provoked the information storm. It becomes very hard for readers to read this lengthy document and to understand this document meaning. In this paper, they have used the text mining method and also the text summarization method. Special attention is provided to Bio-inspired method here for text summarization. There is one of the approaches, i.e. is an obstructive approach that analyzes the text and then generates the summary. Herr abstractive approach and Bio-inspired approach is combined to get the best and effective Text Summarization.

H. Huang [8] There is two types of data gathering structure such as structured and the other one unstructured.In structured data gathering, the data are gathered by social sensing such as temperature sensing, GPS devices where the data is gathered only in the form of numerical form and then after it is converted into the statistical way and then it is used to summarize the dataset. And the second one unstructured where the data is received in the text form and the image. In this paper they have managed with both information gathering structures such as structured and unstructured it can be used in military intelligence.

J. Zenkert [9] For computers, however, it is very hard to understand the natural language, it requires very high complexity to understand the language relationship between the text information.The information on the internet it is very hard to identify the relationship between the text information so that they can deal with the summarization they must identify the relationship and the contact of each line and the important concept of the document so they can summarize the text. So, the Multidimensional knowledge representation is one of the most interactive part text mining, which used in this paper.

C. Wang [10] There is a very high-speed development in the text summarization technique due to the growth of the internet services there is a very large data or the text is available totheir vast evaluation in the field of text summarization. In this paper the new technique is developed it is called as HowNet it gives the same meaning of the text with short summary the sequences and the and the meaning is much better than the previous results of the text summarization. The result of this paper relates to humans very effectively

A. Pal [11] In recent years there has been tremendous growth in the field of text mining and text summarization. There are few methods such as tagged rules, the format of writing on paper and the position of text and there are a few more. The methodology of this paper is one approach of the text summarization i.e. unsupervised learning methodology. In this paper, the first approach is to find the weights of all the sentences and then they're sorted in decreasing order and then percentage given for each sentence for summarization. The result gives 50% of text summarization of a given text and 25% of the original text.

S. Chakraborti [12] Text summarization is nothing but the extracting the data from the given document and the given text file. Later then the researchers started focusing on the different technique to summarize the text in the business rule and helped the managers to make growth in their organization. The methodology such as Text Analysis, Text Summarization, Multi-Document Summarization, Topic Detection, Clustering, Decision Support Systems, and Competitor Intelligence is used in this paper to get the exact and the correct information of the text.

A. Bagalkotkar [13] Automatic Text summarization plays an important role in reducing human efforts. In this paper, the specific web document is given as an input file by using the statistical NLP techniques for generating the text summary. Text summary mostly depends on the number of terms and the number of words in sentences and by using a number of words in a sentence. It is nothing but extracting the effective data from the given document. Text summarization is one of the major topics of the research nowadays

III. Proposed Methodology

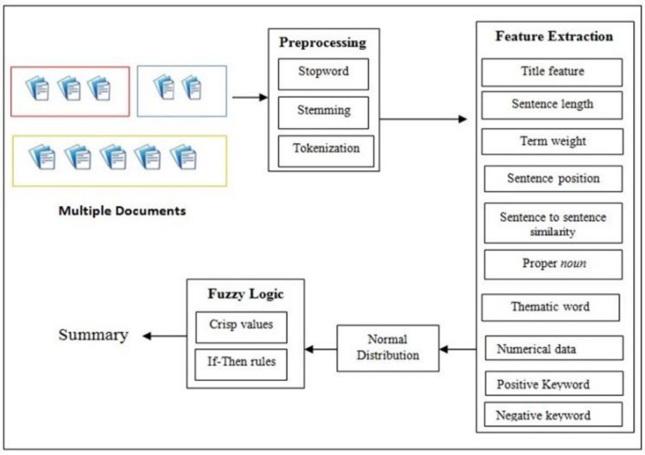

Fig. 1: Overview of the Multi Document Summary Extraction

The presented technique for the extraction of the summary from a document has been elaborated below and depicted in a pictorial format in the figure.

A. Step 1: Document Input –

The system can handle various formats of text-based data files, such as .doc, .pdf, etc. therefore, the input file is uploaded in the system as a folder. This folder contains all the documents that are supposed to be summarised. As this system is capable of extracting summaries from multiple documents at a time, the input is designed to access a folder by default. The various extensions such as .pdf (pdf API), .doc (Apache POI API) are read using their respective APIs. The documents are scanned and the content is concatenated to a string to provide easier processing.

B. Step 2: Pre-processing –

Most of the time, there are unwanted and meaningless words in a sentence which actually serve no purpose for increasing the understanding of the system effectively. Therefore, these words are eliminated from the concatenated string to make the string lighter and less complex for the system. To actually identify the sentences, the string is split at a period and the sentences are then stored in a separate list, before performing the pre-processing. The various steps in pre-processing are given below.

1) Special Symbol Removal –

Spaces and periods are the only useful and discerning symbols that can be used meaningfully, therefore, the rest of the symbols are eliminated in this process.

2) Tokenization –

After symbols have been removed the string is subjected to a split along the period. This is done to separate the sentences in the text before, being sent for the stop word removal and stemming procedures.

3) Stopword Removal –

The various conjunctions used to tie up words and sentences, such as from, to, like, of, etc. are eliminated from the string as they do not serve any meaningful purpose for the summarization technique. This also has an added benefit of reducing the size of the string thereby reducing the time taken for extraction of the summary.

4) Stemming –

Due to the fact that most of the words have different variations according to the time, setting, gender, singular and plural forms of the same word. This increases the complexity of the system as a whole, therefore, the in the string are stemmed, or reduced to their root form for easier and straightforward processing and a lightweight string.

C. Step 3: Feature Extraction –

Feature Extraction is one of the most useful steps of this system as the pre-processed text is utilized to extract the features from all the sentences. These features are then stored in the form of a list. The intricacies of the steps are as follows.

1) Title Sentence:

The very first sentence of the text is usually one of the most important parts of the text as it introduces the document. This is formed as the basis for the rest of the text and it is compared and the relevant features are extracted from Equation 1.

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

2) Sentence Length:

One of the key points that are a defining feature of the text is the length of the text. The longer the text, the more amount of information it in comparison with the other shorter sentences. Therefore, sentence length is one of the most important features of the text, which can be derived by using equation 2.

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

3) Term Weight –

The term weight refers to the feature that detects the importance of the words in a given sentence which can be extracted by equation 3.

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

4) Sentence Position –

The placing of a sentence in a text reveals a lot of information about the text and the importance of the sentence as well. The quality of the sentence also depends upon its placement in the text. Therefore, to analyze the importance of the first five sentences of a text are given non zero values and the consequent values are given the score as 0.

5) Sentence Similarity –

Some of the sentences in the text can be similar to one another, and this metric would provide valuable insight. Equation 4 can help identify similar sentences.

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

6) Proper Noun –

A workbook dictionary is utilized to identify the proper nouns in a sentence. Which works like this, if the word in the sentence cannot be found in the dictionary, it is assumed as a proper noun. The equation for the same is given in equation 5.

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

7) Thematic Words –

Meaningful words from a sentence are calculated and treated as a feature, that can be calculated with the help of equation 6.

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

8) Numerical Data –

The statistical strength of the sentence is determined with the help of Equation 7.

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

9) Positive and Negative Data:

Bag of Words technique has been utilized extensively for understanding the semantic motto of the document. As the positive and negative data is a measure of how good or bad the words in a document are respectively. This can be easily calculated by equation 8 and 9.

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

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

D. Step 4 – Gaussian Distribution –

A single factor representation the 10 features of the sentences that have been extracted; is referred to as the optimized factor of the sentences. The presented system then utilizes this optimized factor to generate the mean and standard deviations with the help of Equation 10 and 11. The next step utilizes equation 12 to derive the quality ranges of the optimized factors. The quality ranges depict the most distributed sentences and are, therefore, displayed along with the result to the users.

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

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

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

E. Step 5 – Fuzzy Classification –

The presented system utilizes the Fuzzy classification parameters for the optimized factors that have been calculated in the previous step and classified in accordance to two the Fuzzy crisp values, which are defined as VERY HIGH, HIGH, MEDIUM, LOW AND VERY LOW.

To classify the sentences in the given range, IF-THEN rules are applied based on the fuzzy crisp values. The next step is to segregate the sentences according to the summary level selected by the user. Gaussian Distribution is used to optimize the summary process.

IV. Result and Discussions

A Windows-based machine is utilized for the deployment of the proposed Multi-Document Text Summarization methodology. Java is the programming language used for the coding purposes on a NetBeans IDE. The machine utilized for this process is powered by an Intel Core i5 Central Processing Unit with 6 GB RAM used as the primary memory.

The percentage of error between the extracted sentences for the summary and actual sentences can be measure using the attribute called Mean Absolute Error (MAE). This can be represented using the following equation. And Table 1 represents some of the results that are obtained during the experiment and that is also plotted in figure 2.

$$\mathrm{MAE} = \frac{(\sum_{i=1}^{n} |\, xi - yi \,|)}{n}\_\_\_\_\_\_\_\_\_\_(13)$$

Where,

xi - Number of Actual Summary Sentences

yi - Number of Obtained Summary Sentences

n- Number of Trails

No of Input Sentences	Actual Summay Sentences (xi)	Obtained Summary Senteces (yi)	xi-yi
10	4	4	0
20	6	5	1
30	11	9	2
40	15	13	2
50	17	14	3
60	18	17	1
70	20	18	2
80	21	18	3
90	21	17	4
100	24	20	4
		MAE	2.2

Table 1: Experiment Results for MAE

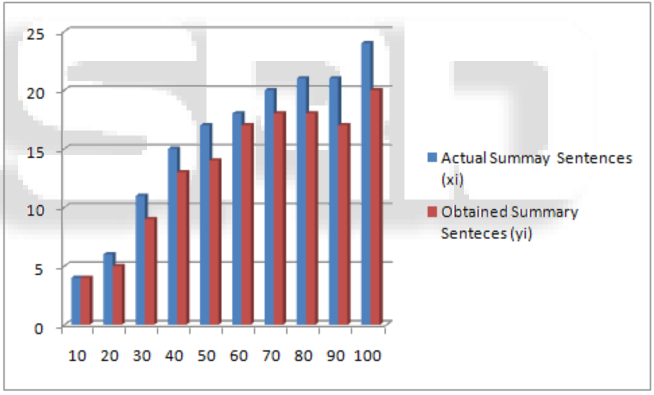

Fig. 2: MAE Measurement

By observing the plot in figure 2, it is come to know that the experiment yields the MAE of 2.2. This MAE for Text summarization process is really good in the very first attempt of implementing NLP protocols using the Fuzzy Classiccation Model and Gaussian Distribution system.

V. Conclusion and Future Scope

There has to be more research done in the field of text summarization as the narration of the sentences governs the extraction of the sentences. The proposed methodology for the Multi-Document Text Summarisation has been deployed successfully, the proposed method has handled various categories of texts and has successfully extracted a huge variety of features from the text as depicted in the previous section. The Quality of the text summarization method is enriched by the introduction of the Gaussian Distribution and eventually Fuzzy Classification. The proposed methodology has been confirmed to drastically improve the performance and quality of the text summary, especially when working with the unstructured data.

The future scope of this methodology revolves around more research into this field of multi-document text summarization techniques. There should be a lot more focus on aspects such as support for large documents and unstructured text amounting to thousands of lines. The methodology can be further simplified into the form of an API that can be used extensively by students and other users seeking a summary.

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 Scoringand 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] .Zenkert, A. Klahold and M. Fathi, "Towards Extractive Text Summarization usingMultidimensional 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