Extractive Text Summarization Technique Using Fuzzy C-Means Clustering Algorithm

***Abstract*—Text summarization process has become one of the significant research areas for years owing to cope up with the astounding increase of virtual textual materials. Text summarization is the process to keep the relevant important information of the original text in a shorter version with the main ideas of the original text. There are two main classifications of text summarization process Extractive and Abstractive text summarization. Extractive summarization processes by using most important fragments of exiting words, phrases or sentences from the original document. A sentence based model using Fuzzy C-Means clustering has been proposed this research. Six best key features including a new feature Sentence Highlighter Feature have been introduced for the sentence scoring. Performance of the proposed FCM model is evaluated by ROUGE, which has been gauged with the precision, recall and f-measure. The result shows that this FCM model extractive techniques with a less outline repetition and profundity of data. Furthermore, it demonstrates more following and intelligent summary than past similar existing methodologies.**

***Index Terms*—Sentence Extraction; Clustering; Summarization**

I. Introduction

Human have the inherent ability to comprehend the implicit meaning of a text and summarize the precise information using own words from the most important aspects of the text. Every day we have to go through a large volume of textual data. But it has been quite difficult to make the best use of this progressively increasing amount of data. Search engines have been extracting snippets by using information retrieval systems for making our life easier but those snippets are becoming a larger size documents themselves [1]. Text summarization has been playing a vital role to solve this issue.

Aim of the text Summarization is to prune and filter a large amount of data into a shorter version keeping the most relevant and significant ideas of the original document [2] [3] [4]. A compact summary of a text allows a user to quickly have an overall idea about the text, indexing effectively and also helps to select relevant documents according to ones necessity [2]. Text summarization has been classified into several categories depending on various aspects [1] [2]. Based on input type, it can be categorized in Single Document and Multi Document text summarization. Depending on purpose, it is divided into Generic, Domain-Specific and Query Based summarization.Abstractive and Extractive are the two approaches of text summarization based on output type [3]. Abstractive text summarization is referred by many scientists as a human generated summary [5]. On the contrary,extractive approaches based on the extraction of sentences with the help of some renounced methods like sentence-based model, word-based model or graph based model. Sentence ranking method is based on some key features and sentence-based model iterates though all the sentences in the document to find out the main ideas. Most of the text summarization methods have used binary parameters. In this paper, we try to solve the problem and overcome this scenario by giving the attributes fuzzy quality. Fuzzy C-Means clustering algorithm allows data to be member of more than one cluster. They have a certain degree of membership each cluster between 0 and 1 [6]. The summary is organized as the sentences of the original texts are incorporated in the summary according to importance. Here, six of the best features among those included a new future ”Sentence Highlighter Score” has been introduced which can help to improve the performance of sentence ranking with Fuzzy C-Means clustering algorithm [6] [3] .

The rest of the paper is categorized as follows. Section II represents the related works done in the field of Text Summarization, Section III describes the key features and the Fuzzy C-Means Clustering model, Section IV presents the experiments and results and finally in Section V the conclusion has been portrayed including future ideas.

II. Related works

The very first text summarization technique was introduced by Luhn [5] back in 1958 using the the thematic feature, Term-Frequency. It has been almost 70 years, text summarization technique has been a research concern and researchers have introduced many genres to represent the best outcome of text documents [4][7].

Again in 1958, sentence location for assessing sentence importance introduced by Baxndale [8]. The examination was exploratory in nature and was roused by an undertaking to decrease the lopsided work required to process the topic of distributed writing. Practical evidences mentioning difficulties innate in the perfect summary concept was visible in the work of Rath et al. [9] in 1961.The examination demonstrates that both the human chose sentences and the program chosen sentences contrast essentially from irregularity. There is an extensive variety of individual contrasts between the human subjects, while the 5 strategies yield little contrasts in the sentences picked. On the other hand syntactic analysis [10]in text summarization introduced the notion of entity-level approaches. Furthermore, the use of Bayesian classifier [11] introduced a probabilistic approach for the sentence selection for summarization. Moreover, from the late 90s are bushy path and aggregate similarity [12] has been used for text summarization. Erkan proposed LexRank [13] algorithm in 2004 inspired by the graph based model.

Fuzzy sets [14] provide a solution in text summarization technique by denoting a parameter to measure the degree ambiguity in a context. In spite of a number of works done in fuzzy logics based text summarization, the Fuzzy C-Means (FCM) clustering is hardly explored in this area [3] [15] [16].

Graph based models have been one of the core inspirations behind the exploration in extractive [17] summarization approaches. In A four dimension Graph Model for Automatic Text Summarization [17] a graph model has been utilized to form extractive summary. In later works, sentence based extractive summarization methods have been proven to be more efficient and less time and space consuming than graph based and word based models [18].

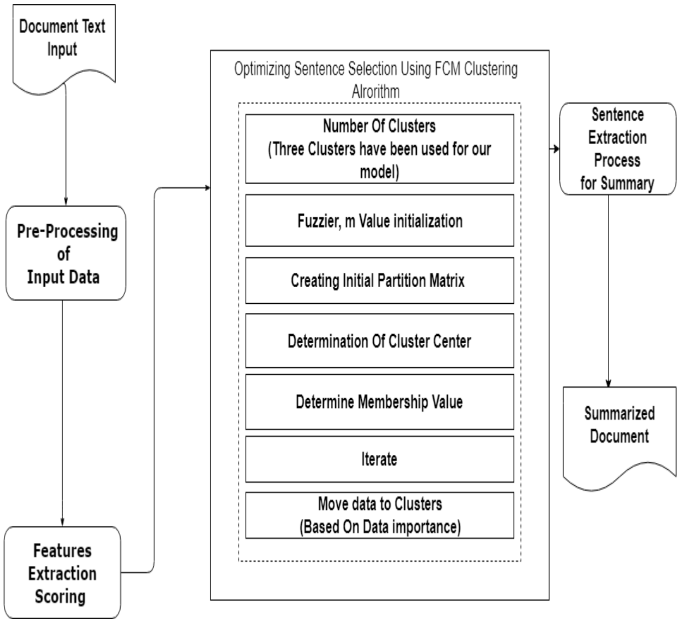
Zheng et al. [19] used applied relations of sentences for multidocument summarization. This idea was made out of three noteworthy components. They were idea bunching, sentence idea semantic connection, and synopsis age.

Ferreira et al. [1] planned a multidocument synopsis display dependent on linguistic and statistic treatment. This methodology separates the significant worry of set of reports to maintain a strategic distance from the issues of this sort of outline.

However, in this paper,using FCM algorithm for sentence extraction for generating summaries. The FCM algorithm uses fuzzy sets and fuzzy partition matrix to denote the membership of an element across multiple clusters [6].

III. FUZZY C-MEANS CLUSTERING MODEL

A. Block Diagram of Fuzzy C-Means Clustering Algorithm Model



Document Text Input
Pre-Processing of Input Data
Features Extraction Scoring
Optimizing Sentence Selection Using FCM Clustering Alroithm
Number Of Clusters (Three Clusters have been used for our model)
Fuzzier, m Value initialization
Creating Initial Partition Matrix
Determination Of Cluster Center
Determine Membership Value
Iterate
Move data to Clusters (Based On Data importance)
Sentence Extraction Process for Summary
Summarized Document

Fig. 1. Block Diagram of Fuzzy C-Means Clustering Algorithm Model

B. Data-set and Prepossessing

1) Data-set: CNN News article data-set texts [1] have been used for this experiment. The data-set has followed some pattern like bullet points, tittle of the article, date, numbers etc. As a result the data set can be easily highlighted for extracting main sentences.50 sample texts have been chosen and resemble due to have the performance comparison of the generated summary.

2) Prepossessing: The tasks in processing included reprocessing the input data using NLTK a library of python [18].Besides, the spitting the text into sentences and words, stop words elimination and POS tagging.

C. Feature extraction

1) TF-IDF Score: Term Frequency- Inverse Document Frequency feature this is the very first feature introduced by Luhn [5] back in 1958 for sentence extraction and gauge the uniqueness of a sentence [1][4] [5]. TF-IDF score has been measured using Equation (1) and equation (2) for normalizing the values of sentences,

$$TF - IDF(term) = frequency(term) \times log\frac{frequency(term)}{No.\ of\ Sentences} \tag{1}$$

For a sentence S_i,

$$TF{-}IDF(Score) = \frac{Sum\ of\ TF - IDF(term)\ in\ S_i}{Max\ sum\ of\ TF - IDF(term)\ in\ a\ Sentence} \tag{2}$$

2) Proper Noun Count Score(PNCS): A sentence containing high numbers of proper noun is considered more important than other sentences [1] [11]. Equation (3) shows the formula for calculating PNS (Proper Noun Score) of a sentence, For a sentence S_i,

$$PNCS(S_i) = \frac{No.\ of\ Nouns\ in\ S_i}{Max\ No.\ of\ Proper\ Nouns\ in\ a\ Sentence} \tag{3}$$

3) Numerical Value Score(NVS): Numerical value represents more significant sentence and contains a resourceful information of the text [15] For a sentence S_i Numerical Value Score(NVS),

$$NVS(S_i) = \frac{(No.\ of\ Numerical\ Data\ in\ S_i}{Length\ of\ S_i} \tag{4}$$

4) Sentence Length Score(SLS): It is used for filtering the least and maximum sentence size as they are considered not that important for sentence ranking [1]. For a sentence S_i,

$$SLS(S_i) = \frac{Length\ of\ S_i}{Mean\ Length} \tag{5}$$

5) Tittle Sentence Score(TSS): Sentences first and foremost characterize the subject of the archive though sentences at last finish up or outline the record. The positional estimation of a sentence is determined by doling out the most elevated score an incentive to the primary sentence and the last sentence of the report [17]. For a sentence S_i,

$$TSS(S_i) = \frac{Sum\ (Unique\ Val\ Score)}{Max\ (Sum\ (Score\ in\ a\ Sentence))} \tag{6}$$

6) Sentence Highlighter Score(SHS): The novelty of this proposed model is this feature Sentence Highlighter Score which includes the connection among sentences correlation is imperative for the outline as the sentence frequently alludes to the past or the following sentence. On the off chance that we consider just the connection of a sentence with the past sentence at that point sentences beginning with connectives. For example, this, those, moreover, however, such, although etc. related with significant data reserved sentences. Apart from this, highlighted bullet points, quotation, bold words portrays a significant meaning of that documents. Therefore, this feature includes the value of a sentence by adding the frequency of highlighter in a sentence compared to the maximum gain highlighter score of a sentence from the document.

D. Fuzzy C-Means Clustering

Fuzzy C-Means Clustering algorithm is a soft computing technique was introduced by Dunn [14] in 1973.Later on, improved by Bezdek [6] back in 1981.In 1965 Zadeh [14] used fuzzy sets in the FCM clustering.

1) Partition Matrix: The fuzzy C partition of a set S is represented by U, S .Where,

$$Partition\,Matrix,\,U = \left(\left(\mu_{ij}\right)\right)_{N\times C}$$

The partition matrix U must satisfy the following constraints,

- $0 \leqslant \mu_{ij} \leqslant 1$
- $\sum_{j=1}^{c} \mu_{ij} = 1, \;\; for\,all\,i = 1, 2\,.\,.\,.\,.\,.\,N$
- $0 < \sum_{j=1}^{c} \mu_{ij} \leqslant$ N , for all j=1,2 C

2) Objective Function: The FCM algorithm is focused on until any termination criterion is met, attractively minimize the value objective function J and denoted as,

$$J = \sum_{i=1}^{N} \sum_{j=1}^{C} {\mu_{ij}}^{m} \| \; x_i - c_j \; \|^2 \tag{7}$$

Where x_i is the data element and $\mathrm{c}_|$ is the cluster center.

3) Cluster Center: Cluster center c_j calculation formula is,

$$c_j = \frac{\sum_{i=1}^{N} \mu_{ij} \cdot x_i}{\sum_{i=1}^{N} \mu_{ij}} \tag{8}$$

4) Membership Value: The formula for updating the membership values, i_jf the partition matrix is,

$$U_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right) \overline{\frac{2}{m-1}}}$$

(9)

E. Initialization

1) Input Data: After the splitting sentences of the document thorough NLTK [18] input data is ready to be clustered and each sentences represented as a 6 dimensional vector.

2) Clusters: For the classification of the input document there is 3-clusters used for this experiment based on sentence ranking importance.

3) Fuzzier: Changing the fuzzier value m seven experiments were experimented and the values were 1, 1.2, 1.5, 2.5,3,4,5.

4) Initial Partition Matrix: Formula of initial partition matrix calculated,using the equation of partition matrix.

5) Termination Criterion: Two terminations criterion,

- Error limit, e=.0001, Max Iteration =1000

F. Iteration

1) Cluster centers denoting using equation (8)
2) Objective function denoting using equation (7)
3) Update partition matrix using equation (9)
4) Check termination Criterion and stop.
5) Else back to step 1.

G. Sentence Extraction For Summary

The General idea of choosing sentences from their membership value to the higher cluster. As FCM Algorithm breaks down the full membership of a sentence to single cluster to all the three clusters. The highest scorer sentence of a cluster clustered is regarded as most important and comparing to the centroid the most relevant sentences have been extracted.

IV. RESULTS AND EVALUATION

A. ROUGE Analysis

Evaluating summary is quite difficult task .A summary may have different types of summaries with various sentences but the idea is illustrated perfectly in all cases. So, one perfect summary of a document cannot be said [2]. As it compares n-gram statistics approach to measure the precision, recall and f-measure of a summarizer. The equations for measuring,

$$recall, r = \frac{(Human\;Generated\;Summary) \cap \;(Generated\;Summary)}{(Generated\;Summary)} \tag{10}$$

$$precision, p = \frac{(Human\;Generated\;Summary) \cap \;(Generated\;Summary)}{(Human\;Generated\;Summary)} \tag{11}$$

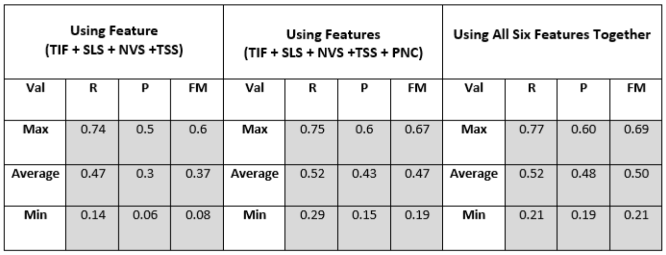
$$f - measure, f = \frac{(2\times\;r\times\;p)}{(r\;+\;p)} \tag{12}$$

1) Feature Based Comparison of the Model: The feature based comparison of the model has been experiments by separating the features and the result shows a promising output for the six features together.

The gauges,

R=recall, P=precision, Fm= f-measure

Table:1 Feature Based Comparison of the Model



Using Feature (TIF + SLS + NVS +TSS)				Using Features (TIF + SLS + NVS +TSS + PNC)				Using All Six Features Together			
Val	R	P	FM	Val	R	P	FM	Val	R	P	FM
Max	0.74	0.5	0.6	Max	0.75	0.6	0.67	Max	0.77	0.60	0.69
Average	0.47	0.3	0.37	Average	0.52	0.43	0.47	Average	0.52	0.48	0.50
Min	0.14	0.06	0.08	Min	0.29	0.15	0.19	Min	0.21	0.19	0.21

B. Similar Approaches Based Comparison on CNN Dataset:

K-means is popular unsupervised algorithm which creates clusters for un-leveled dataset. But the f-measure shows a lower result then our model. Minibatch K-Means is an upgraded version of K-means algorithm and this approach is faster than K-Means. Lastly, fuzzy logic which is another popular approach for text summarization and it shows a moderate result scoring 0.47. Our model scored 0.53 which is a promising f-measure and determines the accuracy of finding main ideas.

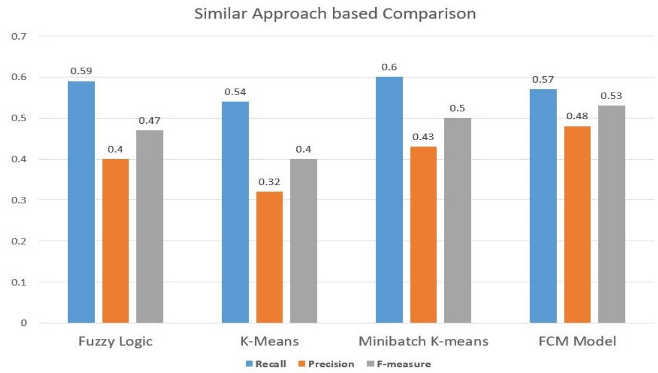


Fig. 2. Similar Approaches Based Comparison on CNN Dataset

C. Fuzzier Value Based Comparison

Changing the fuzzier value m seven experiments were experimented and the values were 1, 2, 2.5, 3, 3.5, 4, 5. The results shows different progress rate the best progress can be noticeable in the value of m=2.5.

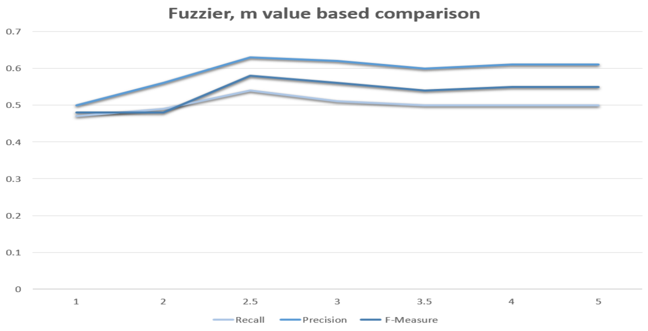


Fig. 3. Different Fuzzier Value Based Comparison

Apart from this,the f-measure comparison with other models like such as Baseline, MS-Word, GSM etc [16] which use different approaches(Based on Duc-2002 Dataset), the performance of fuzzy C-Means clustering model gives a better f-measure performance than other popular summarizing models also.

V. Conclusion

Text classification and summarization has been one of the core areas of NLP. Our Fuzzy C-Means clustering model which is based on the combination of six most significant sentence ranking features with a new feature ”Highlighter Sentence Score” has provided a new dimension to the extractive text summarization technique in significant way. We have taken the ranking features considering the fact that they are the most effective and useful for extractive text summarization and they have given our expected result in the study. We have implemented the technique on the CNN dataset to test the result to evident our claim of improving the technique. We also found out that the F-measure is highest for our chosen algorithm Fuzzy C-Means, compared to others. In future, the ideas can be extended and applied also for sentence ranking procedure in abstractive text summarization technique.

References

1 Ferreira, R., de Souza Cabral, L., Lins, R. D., e Silva, G. P., de Freitas, F. L. G., Cavalcanti, G. D. C., Lima, R., Simske, S. J., and Favaro, L., “Assessing sentence scoring techniques for extractive text summarization,” *Expert Syst. Appl.*, vol. 40, pp. 5755–5764, 2013.

2 Hovy, E., “Text summarization,” in *The Oxford Handbook of Computational Linguistics 2nd edition*, 2003.

3 Moratanch, N. and Chitrakala, S., “A survey on extractive text summarization,” in *Computer, Communication and Signal Processing (ICCCSP), 2017 International Conference on*. IEEE, 2017, pp. 1–6.

4 Das, D. and Martins, A. F., “A survey on automatic text summarization,” *Literature Survey for the Language and Statistics II course at CMU*, vol. 4, pp. 192–195, 2007.

5 Luhn, H. P., “The automatic creation of literature abstracts,” *IBM Journal of research and development*, vol. 2, no. 2, pp. 159–165, 1958.

6 Bezdek, J. C., Ehrlich, R., and Full, W., “Fcm: The fuzzy c-means clustering algorithm,” *Computers & Geosciences*, vol. 10, no. 2-3, pp. 191–203, 1984.

7 Manning, C., Jurafsky, D., and Liang, P., “The stanford natural language processing group,” 2010.

8 Baxendale, P. B., “Machine-made index for technical literaturean experiment,” *IBM Journal of Research and Development*, vol. 2, no. 4, pp. 354–361, 1958.

9 Rath, G., Resnick, A., and Savage, T., “The formation of abstracts by the selection of sentences. part i. sentence selection by men and machines,” *American Documentation*, vol. 12, no. 2, pp. 139–141, 1961.

10 Mani, I., *Advances in automatic text summarization*. MIT press, 1999.

11 Kupiec, J., Pedersen, J., and Chen, F., “A trainable document summarizer,” in *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1995, pp. 68–73.

12 Salton, G., Singhal, A., Mitra, M., and Buckley, C., “Automatic text structuring and summarization,” *Information processing & management*, vol. 33, no. 2, pp. 193–207, 1997.

13 Erkan, G. and Radev, D. R., “Lexrank: Graph-based lexical centrality as salience in text summarization,” *Journal of artificial intelligence research*, vol. 22, pp. 457–479, 2004.

14 Zadeh, L. A. *et al.*, “Fuzzy sets,” *Information and control*, vol. 8, no. 3, pp. 338–353, 1965.

15 Suanmali, L., Salim, N., and Binwahlan, M. S., “Feature-based sentence extraction using fuzzy inference rules,” in *2009 International Conference on Signal Processing Systems*. IEEE, 2009, pp. 511–515.

16 ——, “Fuzzy logic based method for improving text summarization,” *arXiv preprint arXiv:0906.4690*, 2009.

17 Ferreira, R., Freitas, F., de Souza Cabral, L., Lins, R. D., Lima, R., França, G., Simskez, S. J., and Favaro, L., “A four dimension graph model for automatic text summarization,” in *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on*, vol. 1. IEEE, 2013, pp. 389–396.

18 Dunn, J. C., “A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters,” 1973.

19 Zheng, H.-T., Gong, S.-Q., Guo, J.-M., and Wu, W.-Z., “Exploiting conceptual relations of sentences for multi-document summarization,” in *International Conference on Web-Age Information Management*. Springer, 2015, pp. 506–510.