AUTOMATIC TEXT SUMMARIZATION USING FUZZY C–MEANS CLUSTERING

SUBMITTED BY

A. M. Muntasir Rahman – 14101139
Nasif Noor Saleheen – 14301003
Shakil Ashraful Anam – 14301088
Department of Computer Science and Engineering

SUPERVISOR

Hossain Arif
Assistant Professor
Department of Computer Science and Engineering

DECLARATION

We, hereby declare that this thesis is based on results we have found ourselves. Materials of work from researchers conducted by others are mentioned in references.

Signature of Supervisor **Signature of Authors**

Hossain Arif
Assistant Professor
Department of Computer Science and Engineering
BRAC University

A. M. Muntasir Rahman – 14101139
Nasif Noor Saleheen – 14301003
Shakil Ashraful Anam – 14301088

ABSTRACT

Automatic text summarization process has been significantly explored throughout the years to cope with the staggering increase of virtual data. Text summarization process is commonly divided into two areas-Extractive and Abstractive. Abstractive summarization processes generate unique sentences that are different from the sentences in original document keeping the same theme, whereas Extractive summarization processes largely depend on sentence extraction techniques- implementing graph models or sentence-based models. In this paper, a sentence-based model has been proposed where the sentence ranking procedure adopts fuzzy C-Means (FCM) clustering, an unsupervised classification method, for sentence extraction purpose. The sentence scoring task relies on five key features, including Topic Sentence which is the first novelty of the proposed model. Furthermore, C-Means clustering is a soft-computing technique that is usually used for pattern recognition tasks but can be improved significantly by hard clustering the membership of the elements which has not been regarded in similar processes in any of the previous works, adding to the novelty of the presented model. Standard summary evaluation techniques have been used to gauge the precision, recall and f-measure of the proposed FCM model and have been compared with different summarizers from different perspectives. Summarizers having different dataset and approaches such as, bushy path, GSM, baseline, TextRank have been compared to the proposed model using ROUGE method. The outcome shows that the FCM model surpasses the previous approaches by a significant margin.

ACKNOWLEDGMENT

We would like to express our utmost gratitude and appreciation to our supervisor Hossain Arif for his attention and time. We would also like to thank him for giving us the opportunity to work on this topic and assisting us throughout the process.

TABLE OF CONTENTS

LIST OF EQUATIONS ---------- i
LIST OF TABLES ---------- ii
LIST OF FIGURES ---------- iii
CHAPTER 1: OVERVIEW ---------- 1
1.1 INTRODUCTION ---------- 1
1.2 RELATED WORK ---------- 4
CHAPTER 2: FUZZY C–MEANS CLUSTERING MODEL ---------- 8
2.1 DATASET & PREPROCESSING ---------- 8
2.1.1 DATASET ---------- 8
2.1.2 PREPROCESSING ---------- 8
2.2 FEATURE EXTRACTION ---------- 9
2.2.1 TERM FREQUENCY-INVERSE SENTENCE FREQUENCY---------- 9
2.2.1 PROPER NOUN DENSITY ---------- 10
2.2.3 SENTENCE LENGTH ---------- 11
2.2.4 NUMERICAL VALUE ---------- 12
2.2.5 TOPIC SENTENCE SCORE ---------- 12
2.3 FUZZY C–MEANS CLUSTERING ---------- 14
2.3.1 PARTITION MATRIX ---------- 15
2.3.2 OBJECTIVE FUNCTION ---------- 17
2.3.3 CLUSTER CENTER ---------- 17
2.3.4 MEMBERSHIP VALUE ---------- 17
2.4 INITIALIZATION ---------- 18
2.4.1 INPUT DATA ---------- 18
2.4.2 CLUSTERS ---------- 18
2.4.3 FUZZIFIER ---------- 18
2.4.4 TERMINATION CRITERIONS ---------- 19
2.4.5 INITIAL PARTITION MATRIX ---------- 19
2.5 ITERATION ---------- 20
2.6 SENTENCE EXTRACTION ---------- 20
CHAPTER 3: EVALUATION & RESULT ---------- 22
CHAPTER 4: CONCLUSION ---------- 31
REFERENCES ---------- iv

LIST OF EQUATIONS

(1) TF – ISF ---------- 10
(2) TF – ISF – Score ---------- 10
(3) PND – Score ---------- 11
(4) Sentence – Length – Score ---------- 11
(5) Numerical – Value – Score ---------- 12
(6) Topic – Sentence – Score ---------- 13
(7) Objective Function ---------- 17
(8) Cluster Center ---------- 17
(9) Membership Value ---------- 17
(10) Initial Partition Matrix ---------- 19
(11) Recall ---------- 22
(12) Precision ---------- 22
(13) F–Measure ---------- 22

LIST OF TABLES

TABLE I FEATURE BASED COMPARISON OF THE FCMS ---------- 23
TABLE II SAMPLE RESULT – 1---------- 28
TABLE III SAMPLE RESULT – 2 ---------- 29
TABLE IV SAMPLE RESULT – 3 ---------- 30

LIST OF FIGURES

Fig. 1 COMPARISON BASED ON DIFFERENT FUZZIFIER VALUES ---------- 24
Fig. 2 CNN DATASET BASED COMPARISON OF SIMILAR MODELS ---------- 25
Fig. 3 F-MEASURE COMPARISON WITH DIFFERENT MODELS ---------- 26
Fig. 4 CNN DATASET BASED COMPARISON BETWEEN SENTENCE AND GRAPH BASED MODEL ------ 27

CHAPTER 1
OVERVIEW

1.1 **INTRODUCTION**

The Internet has without any doubt made our lives much easier, but making the best use of the progressively increasing amount of data is proving to be harder by each passing day. Information retrieval systems used by many search engines have been doing the job of extracting snippets of data from text documents, making the searching process much easier for us. But, we have reached a point where even those retrieved information require pruning as the snippets are no longer remaining snippets and are becoming huge documents themselves [1]. To solve this issue, Text Summarization (TS) process has become a very important tool in the sector of Natural Language Processing (NLP).

Text Summarization aims to prune and filter huge documents to shorter versions where the most significant ideas and information of the original document can be found [2] [3]. Undoubtedly, summarization of a text requires extensive analysis of the text and demands a lot of effort and time, which simply cannot be provided to the huge amount of data stored in the Internet. Automatic text summarization aims to reduce human effort when it comes to thoroughly going through a significantly large document. Automatic Text Summarization can be categorized into two major processes – Abstractive Text Summarization and Extractive Text Summarization [4] [5].

Abstractive Text Summarization tries to capture the main ideas of a document by extracting key information, paraphrasing sentences that represent those information and giving it a new look that is different from the original document

yet keeping the theme of that document unchanged, also referred by many scholars as “Like a human generated summary”. Abstractive approach requires extreme computational resources and time. Hence, to avoid that Extractive Summarization process prevails in most of the summarization techniques out there. Extractive Summarization approach depends on extraction of key sentences from the original document by using some prominent methods such as- graph based model, sentence-based model, word-based model etc. [4].

Many previous works have been done on the basis of graph-based model. A graph-based model tries to compute the relation among the sentences. Based on the relation of the sentences a graph-based model formulates a graph representing connections that are most important to formulate a summary.

A sentence-based model iterates through all the sentences in the document to figure out the most important sentences, essentially known as “Sentence Ranking” method, which is based on some key features. Every sentence of a document is processed and analyzed to find the relativity of that sentence to the summary. The features can be as simple as sentence length and as complicated as term frequency. Many important features have been pointed out in previous works and in this paper, four of the best features among those [1] were combined and a new feature has been introduced that can potentially improve the performance of sentence ranking algorithms. Another way that the proposed model differs from the previous works is the approach towards implementing Fuzzy C-Means (FCM) clustering algorithm. FCM algorithm, a soft-computing technique, is also regarded as an unsupervised classification method [6]. However, in the presented model that feature of FCM has been taken advantage of by giving it an edge using an initial partition matrix in a fashion that reassembles hard clustering mechanism but cannot be denoted as such.

FCM model is mostly used for pattern recognition tasks. It has hardly been utilized for tasks involved in text summarization, let alone in Extractive text summarization processes. In the presented approach to abstractive text summarization sentence feature extraction tasks have been done using Natural Language Toolkit which is implemented in Python. Feature extraction later on contributes to sentence ranking which essentially tries to find out the most important sentences of a document to form the extractive summary. Sentence ranking procedure is done by FCM algorithm, an unsupervised classification process, that scores each sentence based on the features it has and how distributed those features are. To put in simpler words, sentences that are most important have all the features equally distributed and that is exactly the task FCM tries to do. In the presented work the FCM model has been given modified to give it an advantage, which is in all intense and purposes the initial partition matrix that have been mentioned earlier.

The modified FCM model combined with the merger of the best features found in previous works, as well as the addition of the new feature contributes to the factor that differentiates this work from previous works.

1.2 RELATED WORK

Text summarization has been a research concern for almost 70 years and there have been a number of noteworthy works accomplished by researchers who explored multiple genres in order to represent the gist of text documents in the most limited way possible [2] [3] [4]. The very first text summarization technique that used the thematic feature, “Term- Frequency” was introduced by Luhn [7] back in 1958. Again in 1958, Baxndale [8] introduced the new feature, sentence location for assessing sentence importance. The work of Rath et al. [9] in 1961 demonstrated practical evidences indicating difficulties innate in the idea of perfect summary. In contrast to these surface-level approaches, syntactic analysis [10] in text summarization introduced the notion of entity-level approaches. Moreover, the use of Bayesian classifier [11] introduced a probabilistic outlook in sentence selection for summarization. Other notable methods from the late 90s are bushy path and aggregate similarity [12] used for extracting summary from text documents. Furthermore, inspired by the graph based page rank model, Mihalcea et al. proposed the TextRank [13] algorithm and Erkan proposed LexRank [14] algorithm in 2004 for document summarization.

However, all of the mentioned approaches depend on feature weights and selection of these feature weights can become challenging due to the element of ambiguity in natural languages. Fuzzy sets [15] provide a solution to this issue by denoting a parameter to measure the degree ambiguity in a context. However, despite of a number of works done in fuzzy logics based text summarization, known for pattern recognition, the Fuzzy C-Means (FCM) clustering is hardly explored in this area.

However, the most inspirational and relevant works that motivated the research of this book would be the works in [4] [15] [16] [17] [18] [19] [20]. In Fuzzy Logic Based Method for Improving Text Summarization [17] a fuzzy logic based model for improving text summary has been implemented. The extractive approach was based on a sentence based model where the sentences were extracted from the original document based on eight key features. After preprocessing the document via sentence segmentation, tokenization, removing stop words etc. the sentences were evaluated based on eight features. Sentence length, term frequency, number of numerical data etc. being the key ones among those eight. The analyzed sentences were then put into the fuzzy logic model to recognize patterns and improve the choice of important sentences for forming extractive summary.

Sentence scoring techniques later became widely implemented methods in sentence based models where the sentences are scored based on some key features. For evaluation and experimentation of those extractive summarization methods different datasets were used. In Assessing Sentence Scoring Techniques for Text Summarization [1] the assessment of different sentence scoring techniques was done using three different popular datasets. Valuable suggestions were made in the paper to make the algorithms better. 15 algorithms were tested in the paper which became widely credited throughout the years for sentence based models. Word Frequency, TF/IDF, Upper Case, Cue-phrase, Inclusion of Numerical data, Sentence position 1, Sentence position 2, Sentence Centrality 1, Sentence Centrality 2, Resemblance to the title, TextRank Score, Bushy path are the most mention worthy among those algorithms and most of which were adopted as features for extracting sentences. Assessments were made using the CNN dataset, blog summarization dataset and SUMMAC dataset.

Although the presented work is based on sentence based model, graph based models have been key inspirations behind the exploration in extractive

summarization approaches. In A four dimension Graph Model for Automatic Text Summarization [21] a graph model has been utilized to form extractive summary. Graph models are based on the relationship of the sentences. The relation of each sentence with the other is presented by a graph that later helps to extract the key sentences for the summary. This work relied on four key features for forming the graph model which are as follows. Semantic Similarity, Co-reference resolution, Discourse Relations and Similarity. These four features have been introduced in previous works although the use of Co-reference resolution for extractive summarization was done first in the paper. These features formed the edges of the graph that connected the sentences which were essentially the nodes of the graph model. The similarity methods were based on four key features Centrality, Cosine Similarity, Entropy measure and Word Co-occurrence. The semantic similarity feature relied on characteristics of the words of a sentence. Synonyms, hyponyms and hyponyms are the most popular way to find the semantic similarity between two words. Resnik measure was implemented for figuring out semantic similarity. Co-reference resolution of the sentences was based on name-entity recognition of the document and discourse relations were used to find the grammatical and sequential connection among the sentences. Cause-effect, Violated Expression, Condition, Similarity, Contrast, Temporal Sequence, Attribution, Example and Generalization are the key features that formulated discourse relations.

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. Research work in extractive field have resorted more to sentence ranking or scoring techniques, although some works have relied on features that are based on relations of the sentences but cannot be regarded as Graph based models as these relations were taken into account as features of each sentences. Among the many sentence scoring methods FCM model has proved to be

promising and has left room for upgrades that can improve the model by a significant margin to provide better results.

In this paper, a noble approach is proposed to utilize the FCM algorithm to aid sentence extraction for generating summaries. The FCM algorithm is an unsupervised soft computing clustering technique that uses fuzzy sets and fuzzy partition matrix to denote the membership of an element across multiple clusters [6]. The idea behind the proposed model is to assess the membership value of a sentence (represented as a vector of features) in the partition matrix in order to determine its importance in the document and subsequently, its inclusion in the resulting summary.

CHAPTER 2
FUZZY C–MEANS CLUSTERING MODEL

2.1 DATASET & PREPROCESSING

2.1.1 DATASET

The CNN dataset that contains CNN news articles as texts [1] has been used for experimental purposes. One vital advantage of this dataset is its complex consistency and cleanness, with elegantly composed writings of general intrigue. Other than that, every article has its "highlights", a great quality and compact outline composed by the first creator of 3 or 4 sentences, which might be taken as gold standard. These highlights are bullet points that can be considered as the core for sentence selection. 30 sample texts from the CNN dataset were chosen and for each text, the authors of this paper selected the sentences that best resemble the highlights in order to perform comparison with the generated summary of this model. The chosen sentences met with general consensus in almost all the cases with the exceptions where majority voting decided the inclusion of the sentences in the set.

2.1.2 PREPROCESSING

The preprocessing of the input data is done using the NLTK library of python [22]. The tasks involved are:

- **Paragraph Splitting:** It is the process of splitting the whole text into paragraphs. The process is executed by implementing regular expressions that identifies the intervals between paragraphs.
- **Sentence Splitting:** After splitting the text into paragraphs, the paragraphs are further split into sentences. The splitting of paragraphs into sentences is done via NLTK's built-in sentence tokenizer (*nltk.tokenize.sent_tokenize*).
- **Stop Word Removal:** The words which have hardly any representative value for the document are called stop words and they are removed to provide additional computational ease. The stop words are identified using *"stopwords"* from *nltk.corpus* and filtered out by cross matching afterwards.
- **POS tagging:** The POS tagging or Parts Of Speech tagging is the process of assigning each token with its morphological classification such as – noun, verb, adjective, etc. This is done using the POS-tagger from NLTK (*nltk.pos_tag*).
- **Lemmatization:** The process of lemmatization refers to the mapping of verb forms such as the infinite tense and nouns into the singular form so that the form of the word can be known. Lemmatization is also done using the WordNet Lemmatizer in NLTK (*nltk.stem.WordNetLemmatizer*).

2.2 FEATURE EXTRACTION

2.2.1 TERM FREQUENCY – INVERSE SENTENCE FREQUENCY

This feature is based on the concept of Term Frequency vs. Inverse Document Frequency (TF-IDF) which is often used to calculate the uniqueness of a term in a

document [1] [2] [3] [7]. It is comprehensible that the words that appear in every document tend to contain no significance to a particular document. Therefore, the TF-IDF equation is developed on the concept that the words that are only found in a particular document potentially holds the unique information of that particular document. In this model, each sentence is treated as a document, in order to obtain the aggregate TF-ISF value for a sentence and then each value is normalized with the maximum value calculated [16]. Equation (1) illustrates the formula for calculating TF-ISF score for terms and equation (2) shows the formula for normalizing the aggregate value of the sentences.

$$\text{TF} - \text{ISF (term)} = frequency\,(term) \times \log\left(\frac{No.\ of\ sentences}{frequency\,(term)}\right) \quad (1)$$

For a sentence S_i,

$$\text{TF} - \text{ISF} - \text{Score}, f_1(S_i) = \frac{\text{sum of TF}-\text{ISF (term) in } S_i}{\text{Max (sum of TF}-\text{ISF (term) in a sentence)}} \quad (2)$$

2.2.2 PROPER NOUN DENSITY

This feature is based on the premise that in a document, a sentence containing a high number of proper nouns is comparatively more important than other sentences in the document [1] [11]. To elaborate, in a sentence, proper nouns are often the names of the important characters, places, attributes, etc. around which the entire context of a document revolves. Hence, a sentence containing proper nouns are more likely to contain important relations, transitions or plots that include those key characters, places, attributes, etc. and thus are vital for

summaries made on that document. Equation (3) gives the formula for calculating the proper noun density of a sentence.

For a sentence S_i,

$$\text{PND} - \text{Score}, f_2(S_i) = \frac{\text{No. of proper Nouns in } S_i}{\text{Max (No. of proper nouns in a sentence)}} \quad (3)$$

2.2.3 SENTENCE LENGTH

This feature filters out the sentences that are too long or too short as they are not considered important for summary or sentence ranking [1]. The logic behind this feature is that usually the short sentences on a document are the sentences that only contain single words, author names and exclamatory phrases whereas long sentences are usually seen in grabbers and quotation which are vague related to the gist of the document. Therefore, while generating a summary these sentence holds no importance whatsoever and thus needs to be removed. Equation (4) illustrates the formula for calculating this feature.

For a sentence S_i,

$$ratio(S_i) = \frac{\text{Length}(S_i)}{\text{Average Length}}$$

$$\text{Sentence} - \text{Length} - \text{Score}, f_3(S_i) = \begin{cases} ratio & if\ ratio \leq 1 \\ 2 - ratio & if\ ratio \leq 2 \\ 0 & if\ ratio > 2 \end{cases} \quad (4)$$

2.2.4 NUMERICAL VALUE

Sentences containing numerical values are considered to be important as they can potentially contain useful data about the document [1] [16] [23]. It is often seen that numbers in the document offer precise information regarding the document and the sentences that contain numbers are viewed with greater concerns compared to sentences that do not. This makes the inclusion of sentences that contain numerical data more crucial than the others. Equation (5) shows the formula for calculating the score for this feature.

For a sentence S_i,

$$ratio(S_i) = \frac{\text{No.of Numerical data in } S_i}{\text{Length } (S_i)}$$

$$\text{Numerical} - \text{Value} - \text{Score}, f_4(S_i) = \frac{\text{No.of Numerical data in } S_i}{\text{Length } (S_i)} \quad (5)$$

2.2.5 TOPIC SENTENCE SCORE

Similar sentences that contain the topics which are emphasized at both the beginning and the end of the text typically contain the overall view of the text. This is a combination of the sentence position and sentence similarity feature mentioned in [1] [8] [16] [18]. In the sentence position feature, the sentences that appear at the beginning of a paragraph are given greater importance than others and in the sentence similarity feature; sentences that contain greater lexical similarity with other sentences are given greater importance than the others. However, even though these are very insightful features, in order to implement them, it adds greater time complexity. On the other hand, it is a common practice in academic literature to include sentences named “Topic Sentence” that provides

the complete overview of the context both at the beginning and at the end (Abstract and Conclusion) of the document. Such sentences understandably contain the greater portion of gist of the document than other sentences. Moreover, these sentences are actually paraphrased versions of each other, residing in the most distant paragraphs of the document. As a result, it makes sense to combine the features that focus on position and similarity of sentences to identify these topic sentences and provide a metric of evaluation of their importance in the document. Here, the similarity between the concerned sentences is calculated using WordNet [22] [24]. The formula in equation (6) is used to calculate this feature.

Let, k be an integer between 10 to 50,

Set_b = set of sentences in $k\%$ length of the document

Set_e = set of sentences after $(100 - k)\%$ length of the document

For a sentence S_i in Set_b and S_j in Set_e,

$$\text{Score}_{ij} = Similarity\ (S_i, S_j)$$

$$\text{Topic} - \text{Sentence} - \text{Score},\ f_5(S_i)\ = \frac{\text{sum (Score}_{ij})}{\text{Max (sum (score))}} \tag{6}$$

$$\text{Topic} - \text{Sentence} - \text{Score}, f_5(S_j)\ = \frac{\text{sum (Score}_{ij})}{\text{Max (sum (score))}}$$

2.3 FUZZY C – MEANS CLUSTERING

In computation, the term Fuzzy refers to a form of set theory and logic in which predicates may have degrees of applicability, rather than simply being true or false [6] [15]. This fuzzy set theory was introduced by Zadeh [15] in 1965. The contrast of fuzzy set theory against ordinary set theory is in fact this measurement of membership in fuzzy sets which is derived from concept adjectives used in natural languages that do not correspond to precise mathematical values [6].

In ordinary set theory,

For a set A and an element x,

$$x: x \in A \mid x \notin A$$

A more mathematically profound way to demonstrate this arrangement would be, for an indicator function I_A,

$$I_A(x) = 1 \qquad if\ x \in A$$
$$I_A(x) = 0 \qquad if\ x \notin A$$

In contrast to this, Zadeh defined fuzzy sets as ordinary sets with a membership parameter obtained from the membership function, $u_A(x)$. The exceptionality of this membership value is that it can take any value from the interval $[0, 1]$ [6].

$$u_A(x) \in [0, 1]$$

This arrangement illustrates that where in ordinary set theory, a statement takes binary value, meaning that a statement can be either true (1) or false (0), in fuzzy set theory the statement can take fractions, which means it can be true with a membership value of 0.35, 0.5, 0.6, etc. [15].

Fuzzy C-Means Clustering algorithm is a soft computing technique that was developed by Dunn [19] in 1973 and improved by Bezdek [20] in 1981. The FCM clustering algorithm uses the fuzzy sets introduced by Zadeh which vary from ordinary sets in terms of membership of a particular element as mentioned earlier. Hence, in fuzzy sets, the membership function denotes the value of the degree of membership of an element to multiple sets. This idea of membership is stretched in Fuzzy C-Means clustering algorithm where a membership matrix, named partition matrix denotes the degree of membership of an element across multiple clusters.

2.3.1 PARTITION MATRIX

Defined in reference [6], in the regular C – Partition where, $C \geq 2$ and C is an integer, of a set, $S = \{x_1, x_2 \dots \dots x_N\}$ represented as, $P(A_1, A_2 \dots \dots A_C)$ the definition of the partition stands,

1. $A_i \neq \emptyset \qquad for\ all\ i = 1, 2 \dots \dots C$
2. $A_i \cap A_j = \emptyset \qquad for\ all\ i \neq j$
3. $\bigcup_1^C A_i = S$

In contrast to this, the fuzzy *C* partition of a set S is represented by (U, S) where,

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

Here,

N = Number of elements in S, $\quad i = 1, 2 \ldots\ldots N$

C = Number of clusters or fuzzy sets, $\quad j = 1, 2 \ldots\ldots C$

μ_{ij} = membership value of i^{th} element to j^{th} cluster

The partition matrix U must satisfy the following constraints,

1. $0 \leq \mu_{ij} \leq 1$
2. $\sum_{j=1}^{C} \mu_{ij} = 1, \quad for\ all\ i = 1, 2 \ldots\ldots N$
3. $0 < \sum_{i=1}^{N} \mu_{ij} \leq N, \ for\ all\ j = 1, 2 \ldots\ldots C$

Here, the first constraint denotes the range of the membership value for each element. The second constraint illustrates that for all the elements in the set S, the sum of their membership to the clusters must be equal to 1. This is a correspondence to second constraint of ordinary C-partition. Finally, the third constraint illustrates that for each cluster, the sum of the membership values for all element of S must be strictly between 0 to N (number of element) which is a complementary to the third constraint of ordinary C-partition.

2.3.2 OBJECTIVE FUNCTION

The primary focus of the FCM algorithm is until any termination criterion is met, iteratively minimize the value objective function *J* denoted as,

$$J = \sum_{i=1}^{N} \sum_{j=1}^{C} {\mu_{ij}}^{m} \left\| x_i - c_j \right\|^2 \quad (7)$$

Where, x_i is the data element and c_j is the cluster center. The fuzzifier, $m \in [1, \infty)$ dictates the cluster fuzziness as higher value of m results in smaller μ.

2.3.3 CLUSTER CENTER

The formula for calculating cluster center c_j is,

$$c_j = \frac{\sum_{i=1}^{N} ({\mu_{ij}}^{m} . x_i)}{\sum_{i=1}^{N} ({\mu_{ij}}^{m})} \quad (8)$$

2.3.4 MEMBERSHIP VALUE

The formula for updating the membership values, μ_{ij} of the partition matrix is,

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_k \right\|} \right)^{\frac{2}{m-1}}} \quad (9)$$

2.4 INITIALIZATION

2.4.1 INPUT DATA

The input data that is to be clustered is the document broken down into sentences where each sentence is represented as a 5 – dimensional vector. Therefore, an input document consisting of N sentences is represented as a matrix of size $N \times 5$. Here, the input matrix is $S_{N\times 5}$.

2.4.2 CLUSTERS

For this experiment, the number of clusters, C is set to 3 for 3 basic classifications of the input document sentences based on their importance.

$$Cluster\ 1 \rightarrow High$$
$$Cluster\ 2 \rightarrow Mid$$
$$Cluster\ 3 \rightarrow Low$$

2.4.3 FUZZIFIER

Six experiments were done by changing the value of the fuzzifier, m. Values used in the experiments were: $1.5, 2, 2.5, 3, 4, 5$.

2.4.4 TERMINATION CRITERIONS

The initial objective function, $J^{(0)} = \infty$; and 2 terminations criterion are provided,

Error limit, $e = 0.00001$

Maximum Iteration, $t_{max} = 1000$

2.4.5 INITIAL PARTITION MATRIX

The initial partition matrix,

$$U^{(0)} = \left(\left(\mu_{ij}\right)\right)_{N \times C}$$

is calculated using the following formula,

In the input matrix $S_{N \times 5}$ for a sentence $S_i(f_1, f_2, f_3, f_4, f_5)$ and for all $i = 1, 2 \ldots\ldots N$,

$$M = \frac{\sum_{j=1}^{5}(S_i, f_j)}{\sum_{j=1}^{5} \text{Max}\,(f_j)}$$

$$\mu_i = \begin{cases} (1 \;\; 0 \;\; 0) & if\ M \geq A \\ (0 \;\; 1 \;\; 0) & if\ M \geq B \\ (0 \;\; 0 \;\; 1) & if\ M < B \end{cases} \qquad (10)$$

Where, $(A, B) \in [0, 1]\ and\ A > B$

Here, for $A = 0.7$ and $B = 0.3$, the formula illustrates that based on the magnitude of M compared to A and B, S_i is hard Clustered,

$$if\ M \geq 0.7\ \ S_i \in Cluster\ 1$$
$$if\ M \geq 0.3\ \ S_i \in Cluster\ 2$$
$$if\ M < 0.3\ \ S_i \in Cluster\ 3$$

2.5 ITERATION

I. Calculate cluster centers using equation (8)

II. Calculate objective function, J using equation (7)

III. Update partition matrix using equation (9)

IV. IF $\left(J^{(t)} - J^{(t-1)}\right) \leq e | t = t_{max}$: STOP

ELSE: loop back to step 1

2.6 SENTENCE EXTRACTION

The FCM Algorithm breaks down the full membership of a sentence belonging to a single cluster in $U^{(0)}$ to partial membership to all the three clusters in $U^{(t)}$ (final partition matrix). For a sentence S_i, the hard clustered '1' in $U^{(0)}$ is distributed across all the clusters in $U^{(t)}$ where μ_j indicates its degree of membership to j^{th} cluster. Therefore, the general idea of selecting sentences is to assess their

membership value to the high cluster ($j = 1$). The following algorithm illustrates the selection process of the sentences $S_{i=1,2....N}$, based on their membership in $U^{(t)} = \left(\left(\mu_{ij}\right)\right)_{N\times C}$.

$Document = \{S_{i=1,2....N}\}$

$Summary = \{\}$

$Z \in [0, 1]$

$for\ i = 1, 2.... N$:

$if\ \mu_{i1} \geq \mu_{i2} \geq \mu_{i3}\ \&\ \mu_{i1} + \mu_{i2} \geq Z$:

$Summary = Summary \cup \{S_i\}$

$if\ length(Summary) < desiredLength$:

$Summary_2 = \{\}$

$for\ i = 1, 2.... length(Document - Summary)$:

$if\ \mu_{i2} \geq \mu_{i3}\ \&\ \mu_{i1} + \mu_{i2} \geq Z$:

$Summary_2 = Summary_2 \cup \{S_i\}$

$Summary = Summary \cup Summary_2$

[The value of Z is used to adjust the length of the summary]

CHAPTER 3
EVALUATION & RESULT

The task of evaluating something as abstract as summaries is quite difficult and approximate as there cannot be only one perfect summary for a document. Hence, for evaluating a summary, ROUGE (Recall Oriented Understudy for Gisting Evaluation) has become the standard method that requires human generated summaries which are regarded as gold standards and are compared with the machine produced summaries. ROUGE [25] uses n-gram statistics approach to measure the precision, recall and f-measure of a summarizer quantitatively. Following are the equations of these measures,

$$recall, r = \frac{\{gold\ summary\} \cap \{generated\ summary\}}{\{gold\ summary\}} \quad (11)$$

$$precision, p = \frac{\{gold\ summary\} \cap \{generated\ summary\}}{\{generated\ summary\}} \quad (12)$$

$$f - measure, f = \frac{2 \times r \times p}{r + p} \quad (13)$$

Our proposed model was tested with ROUGE-1 and has been compared with different models from different perspective. Table I presents the feature based comparison of the Fuzzy C-Means summarizer (for fuzzifier, m = 3) to portray the progressing evaluation metrics as more features are added to the sentence ranking task.

Here,

- TF-ISF = TF-ISF-Score
- PND = PND-Score
- SL = Sentence-Length-Score
- NV = Numerical-Value-Score
- TS = Topic-Sentence-Score

And,

- Rec = Recall
- Prec = Precision,
- FM = F-Measure

TABLE I. FEATURE BASED COMPARISON OF THE FCMS

Feature / Score	TF-ISF+PND			TF-ISF+PND +SL			TF-ISF+PND +SL+NV			TF-ISF+PND+SL+NV+TS		
	Rec	**Prec**	**FM**	**Rec**	**Prec**	**FM**	**Rec**	**Prec**	**FM**	**Rec**	**Prec**	**FM**
Maximum	0.75	0.5	0.6	0.75	0.5	0.6	0.75	0.6	0.67	0.75	0.6	0.67
Average	0.48	0.32	0.38	0.54	0.45	0.49	0.54	0.45	0.49	0.56	0.46	0.51
Minimum	0.17	0.07	0.1	0.18	0.05	0.08	0.33	0.13	0.18	0.25	0.17	0.2

As mentioned in 2.4.3, different values of m were used in the experimentation of this FCM model. A comparison based on different values of the Fuzzifier, m is shown via the line graph in figure 1.

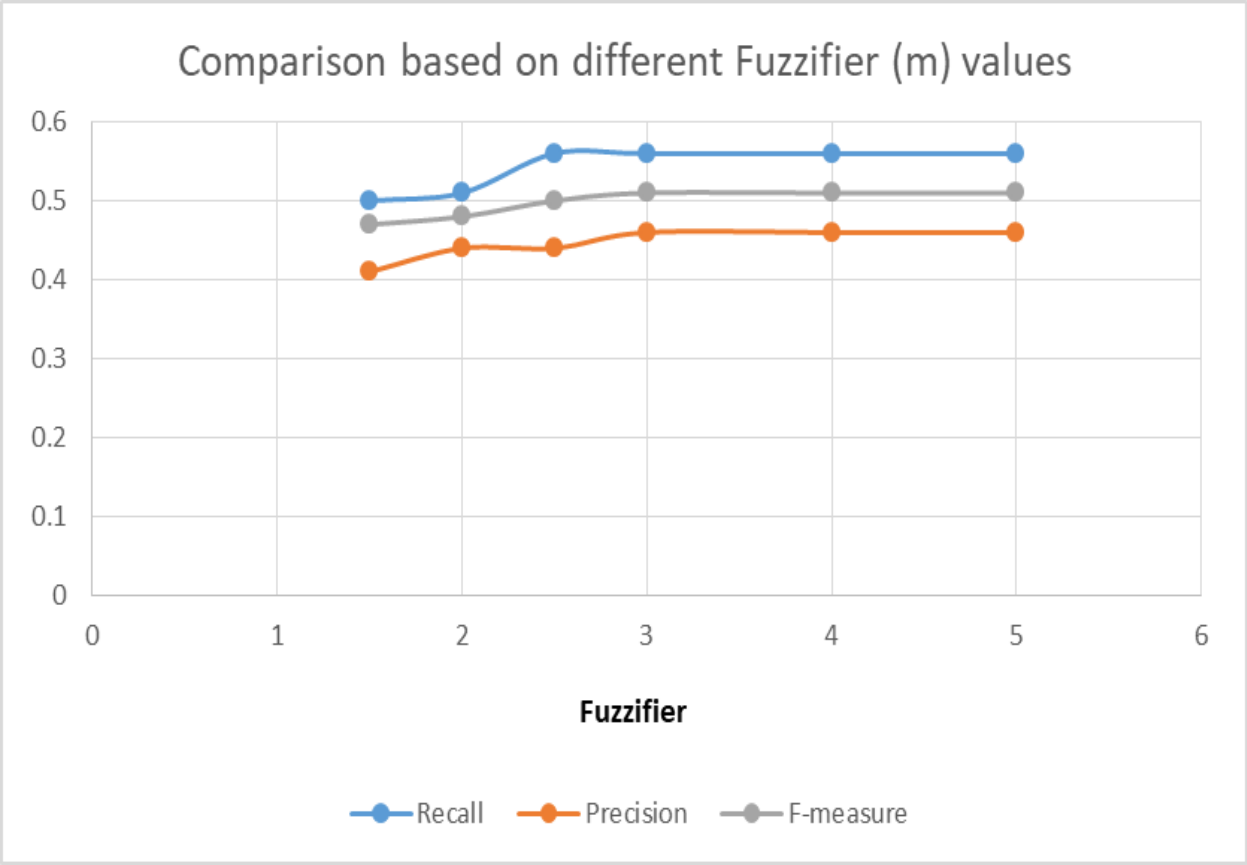

Fig. 1. Comparison based on different Fuzzifier value

Most of the standard datasets available for evaluating automatic summarizers contain golden summaries [1] (bullet points, important snippets and such) and the evaluation metrics of the same summarizer vary significantly, if not drastically, from dataset to dataset. Fig 2 depicts a histogram comparing the results of ROUGE-1 analysis carried out on extractive summaries generated from the CNN dataset by the FCM summarizer and other well-known models [1] [21].

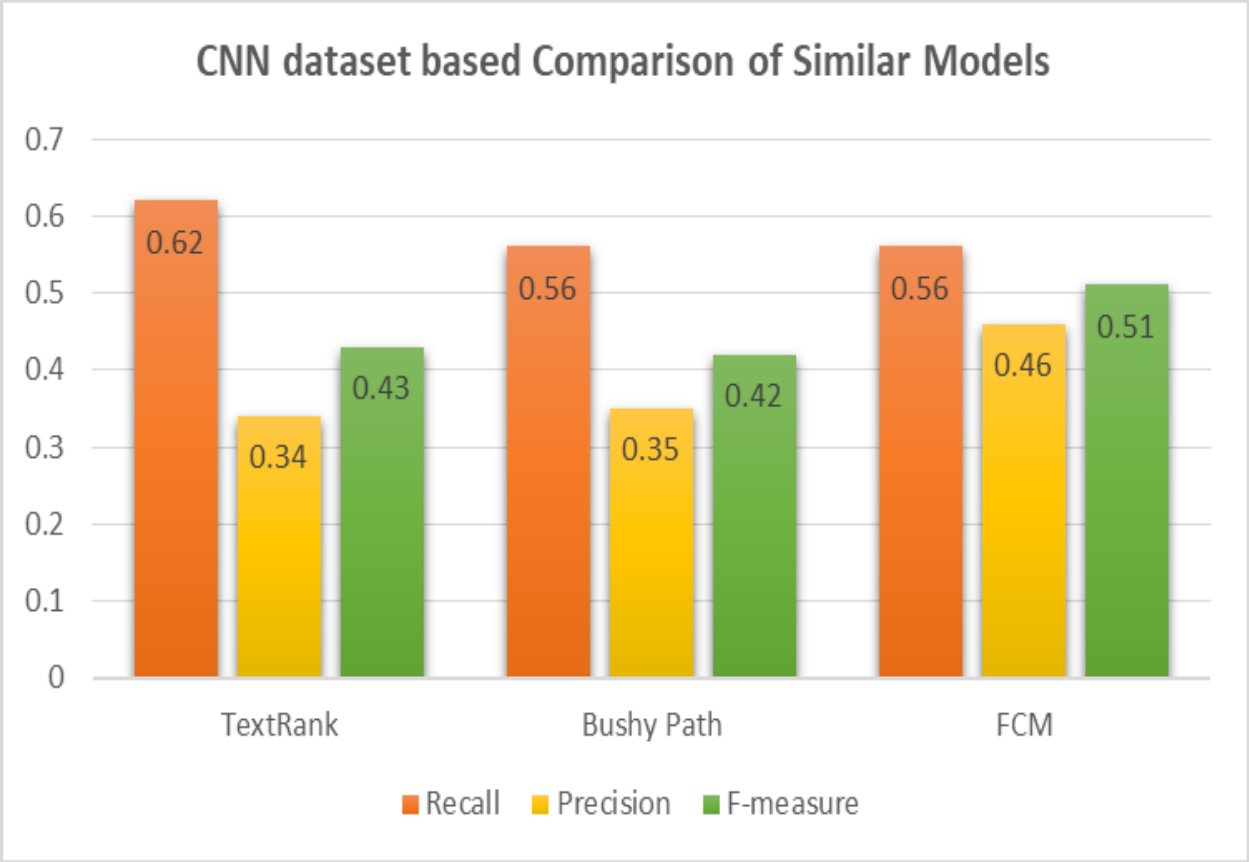

Fig. 2. CNN dataset based comparison of similar models

Graph based models are formulated using graphs that contain nodes as sentences and edges as relations among those sentences. Graph based models have been implemented in many previous works, but the work in A four dimension graph based model [21] have combined the best features that are useful in graph based models. The 4D graph based model has used CNN dataset for experimentation. The following figure (Fig. 3) shows a comparison of recall, precision and f-measure between the 4D graph model and the FCM model.

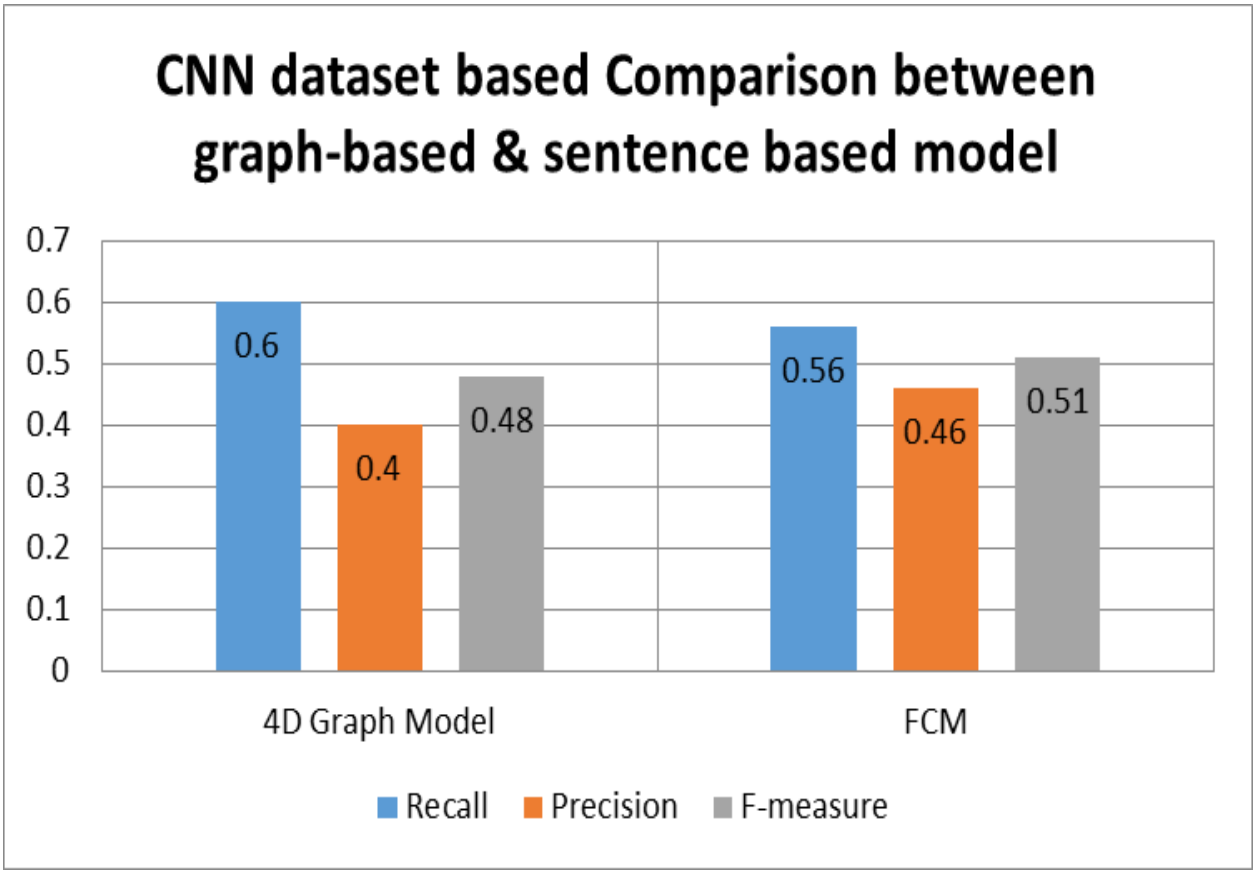

Fig. 3. CNN dataset based comparison between sentence and graph based model

The proposed model in this paper uses a soft computing clustering technique. There are other implementations of summarizer models such as Baseline, MS-Word, GSM etc. [16] [17] that use different approach from that of a clustering model. The following figure (Fig 3) demonstrates how the fuzzy C-Means clustering model gives a better f-measure than these other popular summarizing models (results taken from their performance on DUC2002 dataset).

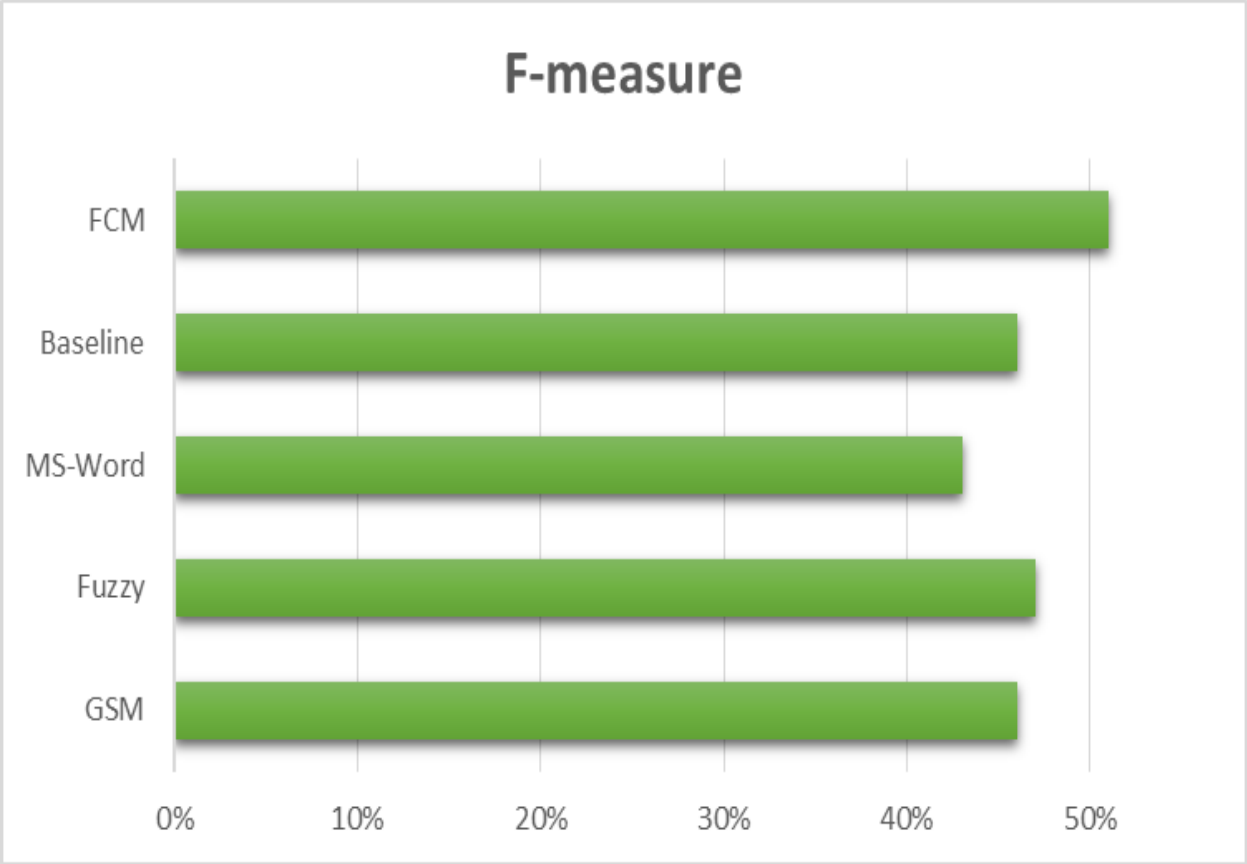

Fig. 4. F-Measure comparison with different models

The use of ROUGE method have been widely accepted and credited for text summarization tasks. The experiments done on the CNN dataset have been compared with different models using different procedures and different datasets. The effectiveness of any probabilistic or statistical approach towards Natural Language Processing is often hard to measure as Natural Language, although follow a lot of rules and regulations, can be extremely random at times. Therefore, it becomes evident that to avoid the randomness of Natural Language, any process or model be compared with another from different angles i.e. different datasets, different procedures and such. In the presented results this notion has been kept in consideration to be as precise as possible regarding the comparison of the proposed model with the existing ones.

The following 3 tables illustrate 3 sample experimental results of the FCM summarizer on CNN news articles and their comparison with the gold summaries.

TABLE II. SAMPLE RESULT – 1

Article	Gold Summary	FCM summary
At least 162 passengers and 11 crew members have reported being ill on board Princess Cruises' Caribbean Princess, according to the Centers for Disease Control and Prevention. The institute said health officers would board the ship in Houston to investigate the gastrointestinal illness, which is causing vomiting and diarrhea. The news follows reports of sickness this week on another cruise ship, this one from the Royal Caribbean line. Nearly 700 crew and passengers fell ill aboard the Royal Caribbean's Explorer of the Seas, the highest number of sick people reported on any cruise ship in two decades, CDC data show. That ship returned home Wednesday, two days earlier than expected. To compare the cruises, 5.22% of passengers on the Caribbean Princess reported being ill, versus 20.5% on the Explorer of the Seas. The outbreak on board the Caribbean Princess has been confirmed as norovirus, according to Julie Benson, a spokeswoman for Princess Cruises. Are cruise ships floating petri dishes? Noroviruses spread easily and are a common cause of gastroenteritis, which produces vomiting and diarrhea. Norovirus is also suspected on board the Explorer of the Seas, though the cause of the illness there has not been confirmed. Caribbean Princess is expected in Houston early Friday. The seven-day cruise is being cut short by one day. Sick passengers are being asked to stay in their cabins, while staff disinfect public areas such as restrooms and elevators. The decision to cut the trip short was made based on forecasts for heavy fog, not the outbreak, Benson said. CNN first learned of the stricken Princess ship from a Twitter post by the Houston Chronicle. Royal Caribbean cruise ship returns home - with a sickness record CNN's Miriam Falco contributed to this report.	At least 162 passengers and 11 crew members have reported being ill on board Princess Cruises' Caribbean Princess, according to the Centers for Disease Control and Prevention. To compare the cruises, 5.22% of passengers on the Caribbean Princess reported being ill, versus 20.5% on the Explorer of the Seas. The outbreak on board the Caribbean Princess has been confirmed as norovirus, according to Julie Benson, a spokeswoman for Princess Cruises. Noroviruses spread easily and are a common cause of gastroenteritis, which produces vomiting and diarrhea.	At least 162 passengers and 11 crew members have reported being ill on board Princess Cruises' Caribbean Princess, according to the Centers for Disease Control and Prevention. Nearly 700 crew and passengers fell ill aboard the Royal Caribbean's Explorer of the Seas, the highest number of sick people reported on any cruise ship in two decades, CDC data show. To compare the cruises, 5.22% of passengers on the Caribbean Princess reported being ill, versus 20.5% on the Explorer of the Seas. The outbreak on board the Caribbean Princess has been confirmed as norovirus, according to Julie Benson, a spokeswoman for Princess Cruises. CNN first learned of the stricken Princess ship from a Twitter post by the Houston Chronicle. Royal Caribbean cruise ship returns home - with a sickness record CNN's Miriam Falco contributed to this report.

TABLE III. SAMPLE RESULT – 2

Article	Gold Summary	FCM summary
Taliban militants, who implemented Islamic law in Pakistan's violence-plagued Swat Valley last week, have now taken control of a neighboring district. Protests in Karachi against the creation of sharia courts in Swat Valley. Here are some answers about the Swat Valley, its history and what's taking place there. What is Swat Valley? Swat Valley is located in Pakistan's North West Frontier Province, near the border with Afghanistan and about 185 miles (300 kilometers) from the Pakistani capital of Islamabad. The alpine region once was one of Pakistan's premier tourist destinations, boasting the nation's only ski resort until it was shut down after Taliban militants overran the area. It also was a draw for trout-fishing enthusiasts and those wishing to visit the ancient Buddhist ruins in the area. What's happening in Swat Valley? In recent years Taliban militants unleashed a wave of violence that claimed hundreds of lives in the province. The militants wanted sharia law -- or Islamic law -- imposed in the region. They took over the valley in 2008. The central government of Pakistan, which long exerted little control in the area, launched an intense military offensive in late July to flush out the militants. In retaliation, the Taliban carried out a series of deadly attacks and began gaining ground, setting up checkpoints in the area. Has the government intervened? The militants and the Pakistani government reached a peace deal earlier this year, which was recently signed into law by Pakistani President Asif Ali Zardari. Under the deal, sharia law was imposed in the region. While the peace deal drew criticism for the Pakistani government, some analysts and political observers say the government had little choice but to capitulate, as militants have terrorized the region with beheadings, kidnappings and the destruction of schools. What's happening now? This week, the Taliban moved to seize control of the neighboring Buner district, bringing it closer to Islamabad than it has been since Taliban insurgency began. What is sharia law? Sharia law is Islamic law. While there are different interpretations of it, the Taliban's strict interpretation forbids women from being seen in public without their husbands and fathers, requires veils for women and beards for men, and bans music and television. Consequences are severe; during the Taliban struggle to impose sharia law, anyone found disobeying was pinned to the ground and lashed. Others were beheaded and hung from poles, with notices attached to their bodies that anyone daring to remove the corpse before 48 hours had passed would also be beheaded and hanged.	Taliban militants, who implemented Islamic law in Pakistan's violence-plagued Swat Valley last week, have now taken control of a neighboring district. The alpine region once was one of Pakistan's premier tourist destinations, boasting the nation's only ski resort until it was shut down after Taliban militants overran the area. It also was a draw for trout-fishing enthusiasts and those wishing to visit the ancient Buddhist ruins in the area. This week, the Taliban moved to seize control of the neighboring Buner district, bringing it closer to Islamabad than it has been since Taliban insurgency began.	Taliban militants, who implemented Islamic law in Pakistan's violence-plagued Swat Valley last week, have now taken control of a neighboring district. Swat Valley is located in Pakistan's North West Frontier Province, near the border with Afghanistan and about 185 miles (300 kilometers) from the Pakistani capital of Islamabad. The alpine region once was one of Pakistan's premier tourist destinations, boasting the nation's only ski resort until it was shut down after Taliban militants overran the area. The central government of Pakistan, which long exerted little control in the area, launched an intense military offensive in late July to flush out the militants. The militants and the Pakistani government reached a peace deal earlier this year, which was recently signed into law by Pakistani President Asif Ali Zardari. This week, the Taliban moved to seize control of the neighboring Buner district, bringing it closer to Islamabad than it has been since Taliban insurgency began.

TABLE IV. SAMPLE RESULT – 3

Article	Gold Summary	FCM summary
Nissan Motor Company announced the return of the Datsun brand after 30 years, with plans to introduce a low-cost car in several emerging markets in 2014 . In a nod to the growing importance of developing markets, Nissan CEO Carlos Ghosn made the announcement Tuesday in Indonesia, one of three markets the new car line will debut. "The Datsun brand has a global mission, but when you go on a global mission you start with priorities and today we said our priority are three: They are Indonesia, they are India and they are Russia," Ghosn said. "That doesn't mean that we are going to be limiting it to there. This is the first step of development, which are these three markets, but absolutely not excluding other high growth markets or other emerging markets from the range of Datsun," he added. The Datsun brand first emerged in Japan in 1923 but was phased out in the 1980s as the company focused on mid-market buyers and its upscale Infiniti brand, launched in 1989 . The new line will capitalize on the Datsun tradition of low-cost, sporty cars marketed in burgeoning economies where many young drivers are first-time buyers. The company also plans to include a "green" line of Datsun cars. Ghosn drew an analogy to Japan's economy when it launched the Datsun brand and today's Indonesian economy, which has grown steadily at about 6% per year. The country is the fourth most populous nation in the world with 250 million people. "Datsun was introduced to Japan to bring reliable, afford and contemporary technology to consumers who had too few choices," Ghosn said at the Indonesia press conference. "Today, this is the very situation in Indonesia. There is a growing segment of up and coming customers who have few choices. They can buy motorcycles, they can buy used vehicles or they can buy a new vehicle but one based on older technology. We will offer them an exciting new choice." The company also announced plans to invest $400 million in Indonesia production, adding 3,300 jobs and producing 250,000 vehicles by 2014 . Nissan sales outlets will increase to 150 in Indonesia by 2015, the company said. "Datsun is part of our company heritage and will now contribute to its future," Ghosn said.	Nissan Motor Company announced the return of the Datsun brand after 30 years, with plans to introduce a low-cost car in several emerging markets in 2014 . "The Datsun brand has a global mission, but when you go on a global mission you start with priorities and today we said our priority are three: They are Indonesia, they are India and they are Russia," Ghosn said. In a nod to the growing importance of developing markets, Nissan CEO Carlos Ghosn made the announcement Tuesday in Indonesia, one of three markets the new car line will debut. The company also announced plans to invest $400 million in Indonesia production, adding 3,300 jobs and producing 250,000 vehicles by 2014 . Nissan sales outlets will increase to 150 in Indonesia by 2015, the company said.	Nissan Motor Company announced the return of the Datsun brand after 30 years, with plans to introduce a low-cost car in several emerging markets in 2014 . In a nod to the growing importance of developing markets, Nissan CEO Carlos Ghosn made the announcement Tuesday in Indonesia, one of three markets the new car line will debut. "The Datsun brand has a global mission, but when you go on a global mission you start with priorities and today we said our priority are three: They are Indonesia, they are India and they are Russia," Ghosn said. The Datsun brand first emerged in Japan in 1923 but was phased out in the 1980s as the company focused on mid-market buyers and its upscale Infiniti brand, launched in 1989 . Ghosn drew an analogy to Japan's economy when it launched the Datsun brand and today's Indonesian economy, which has grown steadily at about 6% per year. "Datsun was introduced to Japan to bring reliable, afford and contemporary technology to consumers who had too few choices," Ghosn said at the Indonesia press conference.

Chapter 4
CONCLUSION

Text summarization has become a very important research area as the data stored in the cloud are increasingly progressively. Two main approaches of automatic text summarization are extractive and abstractive text summarization process. Abstractive process requires a lot of computational effort and resources and can be avoided using extractive summarization process as it mostly relies on sentence extraction methods among other things. Extractive summarization can be implemented in many ways. Graph based model, sentence based model, word based model are some of the branches of extractive summarization. Sentence based approach has been implemented in many previous works and showed promising results. In the presented work of this paper, a new approach to FCM model based extractive summarization process has been discussed. A hard clustered initialization to provide a significant edge over the FCM model has been implemented in the sentence ranking procedure which depends on four of the most important features selected from many sentence extracting methods from previous works and one additional feature introduced in this paper. The addition of the new feature and the novel way of approaching FCM clustering method improves extractive summarization by a significant margin. The FCM model proves to be generating the key ideas from the original document and can be implemented in abstractive summarization techniques in prospective future works.

REFERENCES

[1] R. Ferreira, L. de Souza Cabral, R. Lins, G. Pereira e Silva, F. Freitas, G. Cavalcanti, R. Lima, S. Simske and L. Favaro, "Assessing sentence scoring techniques for extractive text summarization", Expert Systems with Applications, vol. 40, no. 14, pp. 5755-5764, 2013.

[2] D. Radev, Natural Language Processing - Coursera (FULL) | University of Michigan. 2016.

[3] D. Jurafsky and C. Manning, Dan Jurafsky & Chris Manning: Natural Language Processing. 2014.

[4] N. Moratanch and S. Chitrakala, “A Survey on Extractive Text Summarization”, presented at the Conf. of IEEE International Conference on Computer, Communication, and Signal Processing, 2017.

[5] N. Moratanch and S. Chitrakala, "A Survey on Abstractive Text Summarization", in International Conference on Circuit, Power and Computing Technologies [ICCPCT], KK Dist. India, 2016.

[6] C. Murthy, Mod-06 Lec-41 FCM and Soft-Computing Techniques. 2014.

[7] H. Luhn, "The Automatic Creation of Literature Abstracts", IBM Journal of Research and Development, vol. 2, no. 2, pp. 159-165, 1958.

[8] P. Baxendale, "Machine-Made Index for Technical Literature—An Experiment", IBM Journal of Research and Development, vol. 2, no. 4, pp. 354-361, 1958.

[9] G. Rath, A. Resnick and T. Savage, "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] I. Mani and M. Maybury, Advances in automatic text summarization. Cambridge, Mass.: MIT Press, 2001.

[11] J. Kupiec. , J. Pedersen, and F. Chen, “A Trainable Document Summarizer,” In Proceedings of the Eighteenth Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR), Seattle, WA, pp. 68-73, 1995.

[12] G. Salton, A. Singhal, M. Mitra and C. Buckley, "Automatic text structuring and summarization", Information Processing & Management, vol. 33, no. 2, pp. 193-207, 1997.

[13] R. Mihalcea and P. Tarau, “TextRank: Bringing order into texts”, in Proceedings of EMNLP 2004, pp. 404-411, 2004.

[14] G. Erkan and D. R. Radev, “LexRank: Graph-based lexical centrality as salience in text summarization”, J. Artif. Int. Res., vol. 22, no. 1, pp. 457-479, 2004.

[15] L. Zadeh, "Fuzzy sets", Information and Control, vol. 8, no. 3, pp. 338-353, 1965.

[16] L. Suanmali, N. Salim and M. Binwahlan, "Feature-Based Sentence Extraction Using Fuzzy Inference rules", presented at the Conf. of International Conference on Signal Processing Systems, Santorini, Greece, 2009.

[17] L. Suanmali, N. Salim and M. Binwahlan, "Fuzzy Logic Based Method for Improving Text Summarization" in International Journal of Computer Science and Information Security, Vol. 2, No. 1, 2009.

[18] J. Yadev and Y. K. Meena, “Use of fuzzy logic and WordNet for improving performance of extractive automatic text summarization”, presented at the International Conference on Advances in Computing, Communications and Informatics, Jaipur, India, 2016.

[19] J. Dunn, "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters", Journal of Cybernetics, vol. 3, no. 3, pp. 32-57, 1973.

[20] J. Bezdek, Pattern recognition with fuzzy objective function algorithms. New York: Plenum Press, 1987.

[21] R. Ferreira, F. Freitas, L. Cabral, R. Lins, R. Lima, G. Franca, S. Simskez and L. Favaro, “A Four Dimention Graph Model for Automatic Text Summarization”, 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013.

[22] S. Bird, E. Klein and E. Loper, Natural language processing with Python. Beijing [etc.]: O'Reilly, 2009.

[23] C.Y. Lin, “Training a selection function for extraction,” In Proceedings of the eighth international conference on Information and knowledge management, Kansas City, Missouri, United States, pp. 55–62,1999.

[24] G. Miller, R. Beckwith, C. Fellbaum, D. Gross and K. Miller, "Introduction to WordNet: An On-line Lexical Database*", International Journal of Lexicography, vol. 3, no. 4, pp. 235-244, 1990.

[25] C. Lin, "Rouge: A package for automatic evaluation of summaries", in Text summarization branches out: Proceedings of the ACL-04 workshop, S. Szpakowicz and M. Moens, Ed. Barcelona, Spain: Association for Computational Linguistics, 2004, pp.74-81.