

Utilization of Summarization Algorithms for a Better Understanding of Clustered Medical Documents



Ravi Seeta Sireesha, P. S. Avadhani

Abstract: Medical documents contain rich information about the diseases, medication, symptoms and precautions. Extraction of useful information from large volumes of medical documents that are generated by electronic health record systems is a complex task as they are unstructured or semi-structured. Various partitional and agglomerative clustering techniques are applied for grouping the medical documents into meaningful clusters [4]. Multi-document summarization techniques which are recent development in the field of Natural Language Processing are applied to condense the huge data present in the clustered medical documents to generate a single summary which conveys the key meaning. The summarization techniques can be broadly classified into two types [2]. They are: Extractive Summarization techniques and Abstractive Summarization techniques. Extractive Summarization techniques try to retrieve the most important sentences from the given document. Abstractive Summarization techniques try to generate summary with new sentences which are not present in the document. Extractive summarization techniques using Statistical Approaches are applied on the clustered medical documents. Medical summaries help the patients for a better and prior understanding of the disease and they can get a brief idea before consulting a physician. The generated summaries are evaluated using ROUGE (Recall Oriented Understudy of Gisting Evaluation) evaluation technique.

Keywords: Partitional and Agglomerative Clustering techniques, Multi-document summarization techniques.

I. INTRODUCTION

NLP-Natural Language Processing is a sub-field of Artificial Intelligence which focuses on increasing the ability of the computers to understand and process human languages, so that the computers get closer to a human-level understanding of language [6]. Medical documents contain important information about patients, such as diseases, symptoms, diagnoses, precautions and medication. Health care can be improved if these resources are utilized properly. As the medical documents are unstructured or semi-structured in nature, extraction of useful information from them is a difficult task. Sophisticated language processing techniques can be applied to extract useful information from the medical documents.

The medical documents can be clustered to group all the documents of same kind [4]. The clustered documents can then be summarized to condense the huge data into a meaningful summary.

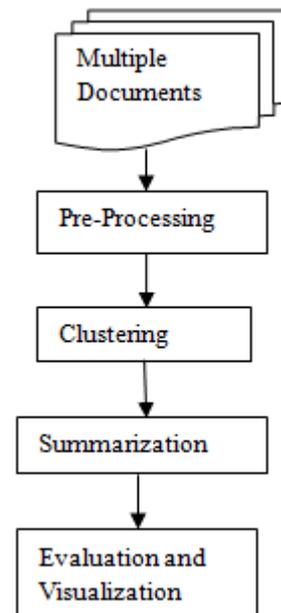
The characteristics of a summary are [1]:

- Single document or multiple documents can be given to summarizer to generate summary.
- The most important information should be preserved in a Summary.
- Compared to the original document, the Summary should be short.

Text summarization is of two types: 1) Extractive summarization system 2) Abstractive summarization system.

In Extractive Summarization, most important text segments of the actual document are identified and included in the summary. In abstractive summarization, interpretation and condensation of the original text is done such that the final summary contains the meaning of the original text [2].

The steps followed in Summarization of Multi-Document are:



Revised Manuscript Received on December 30, 2019.

* Correspondence Author

Mrs. Ravi Seeta Sireesha*, pursuing Ph. D Department of Computer Science and Systems Engineering, Andhra University College of Engineering (A), Andhra University, Visakhapatnam, India.

Prof. P S Avadhani, Professor, Department of Computer Science and Systems Engineering, Andhra University College of Engineering (A), Andhra University, Visakhapatnam, India.

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

➤ **Pre-Processing:** All the documents are pre-processed by performing the following: Tokenization, Stemming and Lemmatization, Removing stop words and punctuation.

➤ Clustering:

K-Means

The K-Means algorithm is a partitional clustering algorithm which groups a set of objects into k clusters. k centroids are defined one for each cluster in such a way that it is closely related to all the objects in that cluster.

Agglomerative Clustering

Agglomerative clustering also called Hierarchical Agglomerative Clustering is a “bottom up” type of clustering. In agglomerative clustering, each data point is defined as a cluster. Pairs of clusters are merged using various distance measures and linkages.

➤ Summarization:

Cluster wise Summary is generated for the documents.

➤ Evaluation and visualization:

Finally, various metrics can be applied to evaluate the summaries of the clustered medical documents.

II. METHODOLOGY

2.1 Statistical Approach-1

1. Create the word frequency table.
2. Score the sentences
/ Score a sentence by its words, adding the frequency of every non-stop word in a sentence dividing every sentence score by the number of words in the sentence */*
3. Find the threshold
/ The average score of the sentences is selected as a threshold */*
4. Generate the summary
/ Select a sentence for a summarization, If the sentence score is more than the average score */*

2.2 Statistical Approach-2

1. Extract the features.
2. Calculate the sentence score(S) as:
$$S = \sum_{i=1}^n S_F(S)$$
where S_F is the score of the feature and n is the number of features.
/ For each sentence in the clustered document calculate the sum of the scores of all the features*/*
3. Select the sentences with the highest score as the summary.
/ Select the sentences having S value more */*

The various features that are considered are as follows:

Feature Extraction

In Feature Extraction, an attribute vector of features is assigned to each sentence of the document [3]. Four features are considered as attributes that attempt to signify the data in the document.

1. Title feature

The word in the title is of high importance if it appears in any sentence. It is calculated as the ratio of the number of words in the sentence that arise in the title over the number of words in title.

$$S_F_1(S) = \frac{\text{Number of Title Words in } S}{\text{Number of Words in Title}}$$

2. Sentence Length

Sentences which are too short may not be much important. Such sentences are not predictable to belong to the summary. It is calculated as the fraction of length of the sentence and the number of words in the longest sentence of the document.

$$S_F_2(S) = \frac{\text{Number of Words in } S}{\text{Number of Words in longest sentence}}$$

3. Thematic Word

The thematic words are the most important words for the given context. It is calculated as the proportion of the number of thematic words in a sentence over the maximum of number of thematic words in all sentences in the document.

$$S_F_3(S) = \frac{\text{Number of Thematic Words in } S}{\text{Max}(\text{Number of Thematic Words in all sentences in a document})}$$

4. Numerical Data

The numerical data has high importance to be included in a medical summary. It is calculated as the proportion of the number of numerical data in a sentence over the sentence length.

$$S_F_4(S) = \frac{\text{Number of Numerical Words in } S}{\text{Number of Words in } S}$$

2.3 Evaluation

Rouge metric is used to evaluate the summaries generated. ROUGE is a set of metrics called Recall Oriented Understudy of Gisting Evaluation (ROUGE). It gives a score based on the similarity in the sequences of words between a human-written model summary and the machine summary [5]. Thus, it helps to automatically evaluate the summary. The ROUGE scores are generally mentioned in terms of F-score(f), Precision(p), Recall(r) help in evaluation of the machine generated summary.

$$\text{Precision} = \frac{\text{Number of overlapping words}}{\text{Total Number of Words in machine Summary}}$$

$$\text{Recall} = \frac{\text{Number of overlapping words}}{\text{Total Number of Words in Reference Summary}}$$

$$F_Score = 2 * \left(\frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})} \right)$$

III. EXPERIMENTAL RESULTS

In this paper output from four Document Clustering techniques are given as input to the two summarization techniques. The Clustering techniques are:

- k-means
- Complete Linkage
- Wards Linkage
- Average Linkage

The summarization techniques are:

- Statistical Approach-1
- Statistical Approach-2

The same medical documents are given as input to the above four techniques with different number of clusters (k=3,4,5). The clustered medical documents are then summarized using the above mentioned two techniques. The validation of the summaries of the clusters is analyzed using ROUGE value and they are shown in tables 1 to 12.

Table 1: f,p,r scores for Statistical approach1 and Statistical approach2 for kmeans with clusters=3

kmeans						
f-sa1	f-sa2	p-sa1	p-sa2	r-sa1	r-sa2	
cl1	0.36 0	0.37 4	0.50 0	0.500	0.2 81	0.29 8
cl2	0.35 7	0.58 1	0.77 8	0.980	0.2 31	0.41 3
cl3	0.29 8	0.51 5	0.92 3	0.656	0.1 77	0.42 4

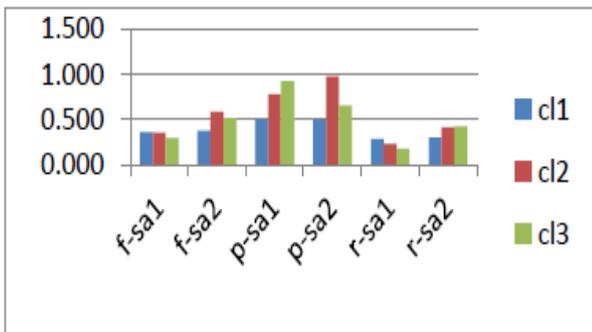


Figure 1: f,p,r scores for Statistical approach1 and Statistical approach2 for kmeans with clusters=3

Table 2: f,p,r scores for Statistical approach1 and Statistical approach2 for agglomerative (affinity='euclidean', linkage='ward') with clusters=3

agglomerative (affinity='euclidean', linkage='ward')						
f-sa1	f-sa2	p-sa1	p-sa2	r-sa1	r-sa2	
cl1	0.22 6	0.30 7	0.92 3	0.51 3	0.12 9	0.21 9
cl2	0.36 4	0.37 8	0.50 0	0.50 0	0.28 6	0.30 4
cl3	0.73 8	0.71 4	0.66 7	0.61 0	0.82 8	0.86 2

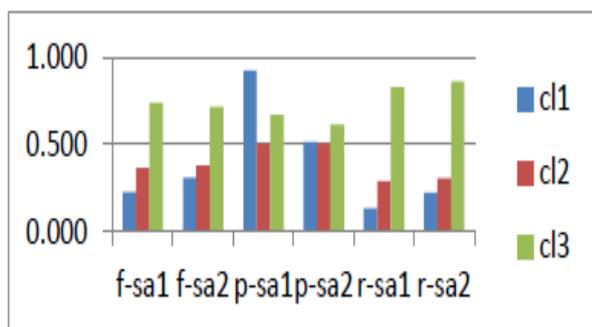


Figure 2: f,p,r scores for Statistical approach1 and Statistical approach2 for agglomerative (affinity='euclidean', linkage='ward') with clusters=3

Table 3: f,p,r scores for Statistical approach1 and Statistical approach2 for agglomerative (affinity='euclidean', linkage='complete') with clusters=3

agglomerative (affinity='euclidean', linkage='complete')						
f-sa1	f-sa2	p-sa1	p-sa2	r-sa1	r-sa2	
cl1	0.27 9	0.46 3	0.77 8	0.98 0	0.17 0	0.30 3
cl2	0.31 9	0.52 2	0.92 3	0.63 4	0.19 3	0.44 4
cl3	0.73 8	0.71 4	0.66 7	0.61 0	0.82 8	0.86 2

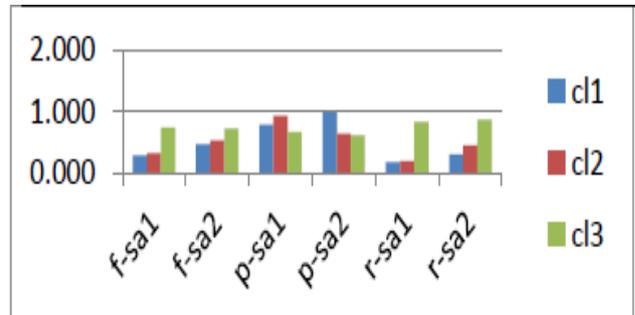


Figure 3: f, p, r scores for Statistical approach1 and Statistical approach2 for agglomerative (affinity='euclidean', linkage='complete') with clusters=3

Table 4: f,p,r scores for Statistical approach1 and Statistical approach2 for agglomerative (affinity='cityblock', linkage='complete') with clusters=3

agglomerative (affinity='cityblock', linkage='complete')						
f-sa1	f-sa2	p-sa1	p-sa2	r-sa1	r-sa2	
cl1	0.21 9	0.30 3	0.92 3	0.52 1	0.12 4	0.21 4
cl2	0.63 0	0.63 0	0.50 0	0.50 0	0.85 0	0.85 0
cl3	0.42 1	0.21 6	0.50 0	0.26 7	0.36 4	0.18 2

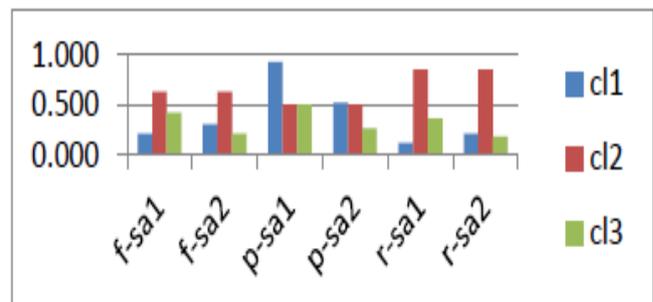


Figure 4: f, p, r scores for Statistical approach1 and Statistical approach2 for agglomerative (affinity='cityblock', linkage='complete') with clusters=3

Table 5: f,p,r scores for Statistical approach1 and Statistical approach2 for kmeans with clusters=4

kmeans						
f-sa1	f-sa2	p-sa1	p-sa2	r-sa1	r-sa2	
cl1	0.67 9	0.64 4	0.82 6	0.73 1	0.57 6	0.57 6
cl2	0.24 3	0.43 6	0.92 3	0.65 6	0.14 0	0.32 7
cl3	0.37 2	0.86 8	0.61 5	1.00 0	0.26 7	0.76 7
cl4	0.73 8	0.71 4	0.66 7	0.61 0	0.82 8	0.86 2

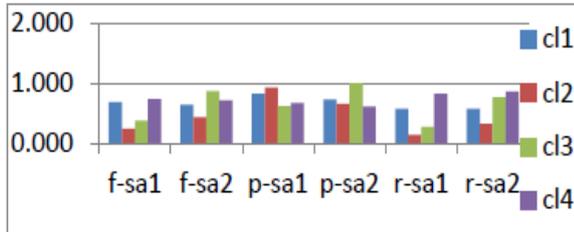


Figure 5: f,p,r scores for Statistical approach1 and Statistical approach2 for kmeans with clusters=4

Table 6: f,p,r scores for Statistical approach1 and Statistical approach2 for agglomerative (affinity='euclidean', linkage='ward') with clusters=4

agglomerative (affinity='euclidean', linkage='ward')						
f-sa1	f-sa2	p-sa1	p-sa2	r-sa1	r-sa2	
cl1	0.06 5	0.24 9	0.10 3	0.20 5	0.04 7	0.31 8
cl2	0.36 0	0.37 4	0.50 0	0.50 0	0.28 1	0.29 8
cl3	0.73 8	0.71 4	0.66 7	0.61 0	0.82 8	0.86 2
cl4	0.48 7	0.51 3	0.73 7	0.75 0	0.36 4	0.39 0

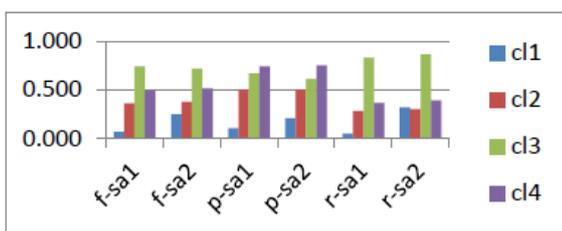


Figure 6: f,p,r scores for Statistical approach1 and Statistical approach2 for agglomerative (affinity='euclidean', linkage='ward') with clusters=4

Table 7: f,p,r scores for Statistical approach1 and Statistical approach2 for agglomerative (affinity='euclidean', linkage='complete') with clusters=4

agglomerative (affinity='euclidean', linkage='complete')						
f-sa1	f-sa2	p-sa1	p-sa2	r-sa1	r-sa2	
cl1	0.35 7	0.58 1	0.77 8	0.98 0	0.23 1	0.41 3
cl2	0.36 0	0.37 4	0.50 0	0.50 0	0.28 1	0.29 8

cl3	0.73 8	0.71 4	0.66 7	0.61 0	0.82 8	0.86 2
cl4	0.31 9	0.52 2	0.92 3	0.63 4	0.19 3	0.44 4

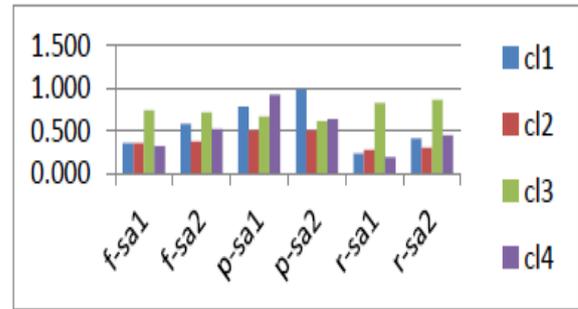


Figure 7: f,p,r scores for Statistical approach1 and Statistical approach2 for agglomerative (affinity='euclidean', linkage='complete') with clusters=4

Table 8: f,p,r scores for Statistical approach1 and Statistical approach2 for agglomerative (affinity='cityblock', linkage='complete') with clusters=4

agglomerative (affinity='cityblock', linkage='complete')						
	f-sa1	f-sa2	p-sa1	p-sa2	r-sa1	r-sa2
cl1	0.226	0.307	0.923	0.513	0.129	0.219
cl2	0.738	0.714	0.667	0.61	0.828	0.862
cl3	0.421	0.216	0.5	0.267	0.364	0.182
cl4	0.618	0.618	0.5	0.5	0.81	0.81

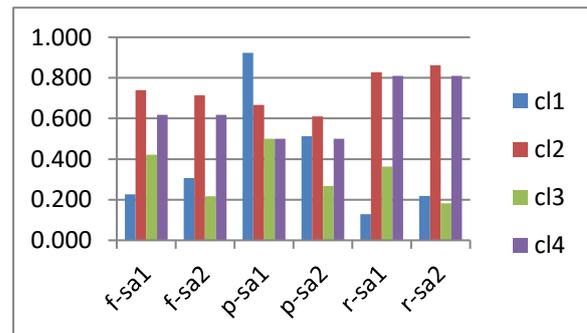


Figure 8: f,p,r scores for Statistical approach1 and Statistical approach2 for agglomerative (affinity='cityblock', linkage='complete') with clusters=4

Table 9: f,p,r scores for Statistical approach1 and Statistical approach2 for kmeans with clusters=5

kmeans						
	f-sa1	f-sa2	p-sa1	p-sa2	r-sa1	r-sa2
cl1	0.293	0.496	0.923	0.648	0.174	0.401
cl2	0.487	0.513	0.737	0.75	0.364	0.39
cl3	0.679	0.644	0.826	0.731	0.576	0.576
cl4	0.738	0.714	0.667	0.61	0.828	0.862
cl5	0.372	0.868	0.615	1	0.267	0.767



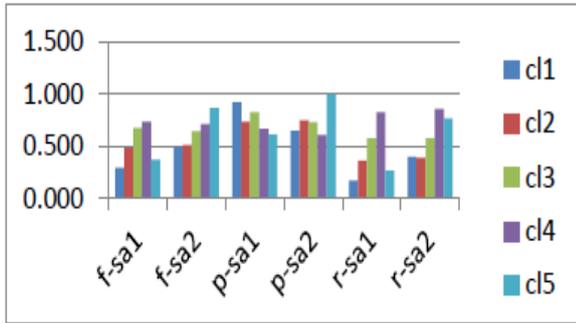


Figure 9: f, p, r scores for Statistical approach1 and Statistical approach2 for kmeans with clusters=5

Table 10: f,p,r scores for Statistical approach1 and Statistical approach2 for agglomerative (affinity='euclidean', linkage='ward') with clusters=5

agglomerative (affinity='euclidean', linkage='ward')						
	f-sa1	f-sa2	p-sa1	p-sa2	r-sa1	r-sa2
cl1	0.36	0.37	0.50	0.50	0.28	0.29
	0	4	0	0	1	8
cl2	0.31	0.52	0.92	0.63	0.19	0.44
	9	2	3	4	3	4
cl3	0.73	0.71	0.66	0.61	0.82	0.86
	8	4	7	0	8	2
cl4	0.48	0.51	0.73	0.75	0.36	0.39
	7	3	7	0	4	0
cl5	0.37	0.86	0.61	1.00	0.26	0.76
	2	8	5	0	7	7

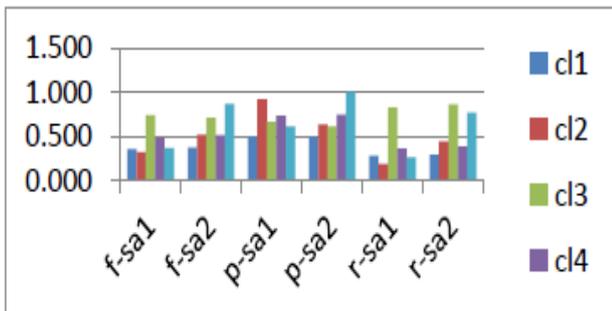


Figure 10: f,p,r scores for Statistical approach1 and Statistical approach2 for agglomerative (affinity='euclidean', linkage='ward') with clusters=5

Table 11: f,p,r scores for Statistical approach1 and Statistical approach2 for agglomerative (affinity='euclidean', linkage='complete') with clusters=5

agglomerative (affinity='euclidean', linkage='complete')						
	f-sa1	f-sa2	p-sa1	p-sa2	r-sa1	r-sa2
cl1	0.36	0.374	0.5	0.5	0.281	0.298
cl2	0.319	0.522	0.923	0.634	0.193	0.444
cl3	0.738	0.714	0.667	0.61	0.828	0.862
cl4	0.487	0.513	0.737	0.75	0.364	0.39
cl5	0.372	0.868	0.615	1	0.267	0.767

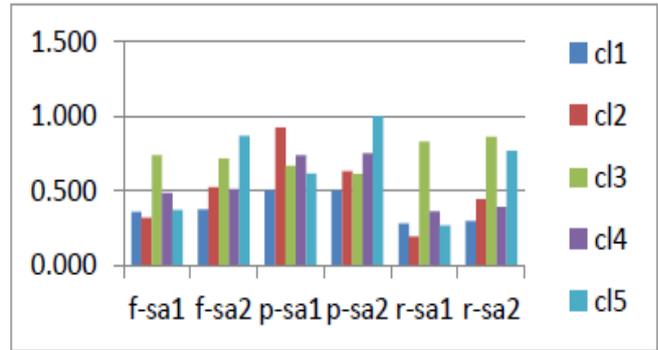


Figure 11: f,p,r scores for Statistical approach1 and Statistical approach2 for agglomerative (affinity='euclidean', linkage='complete') with clusters=5

Table12: f,p,r scores for Statistical approach1 and Statistical approach2 for agglomerative (affinity='cityblock', linkage='complete') with clusters=5

agglomerative (affinity='cityblock', linkage='complete')						
	f-sa1	f-sa2	p-sa1	p-sa2	r-sa1	r-sa2
cl1	0.257	0.339	0.923	0.513	0.149	0.253
cl2	0.738	0.714	0.667	0.61	0.828	0.862
cl3	0.421	0.216	0.5	0.267	0.364	0.182
cl4	0.63	0.63	0.5	0.5	0.85	0.85
cl5	0.372	0.868	0.615	1	0.267	0.767

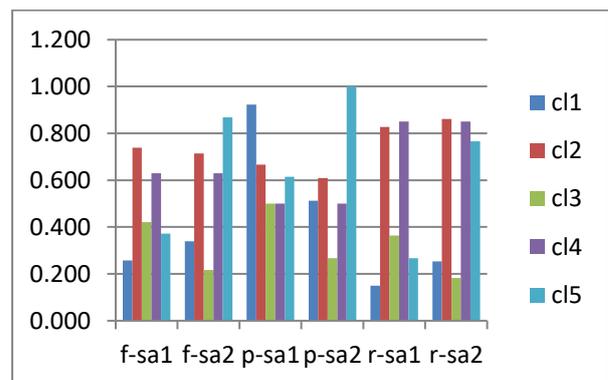


Figure 12: f,p,r scores for Statistical approach1 and Statistical approach2 for agglomerative (affinity='cityblock', linkage='complete') with clusters=5

IV. DISCUSSIONS

The results obtained from the Statistical Approach-1 and Statistical Approach-2 for different clustering algorithms with different number of clusters is analyzed in the tables 13 to 15.

Table 13: f, p, r scores for Statistical approach-1 and Statistical Approach-2 for different clustering algorithms with number of clusters=3

Statistical Approach-1												
	kmeans			agglomerative(affinity='euclidean', linkage='ward')			agglomerative(affinity='euclidean', linkage='complete')			agglomerative(affinity='city block', linkage='complete')		
	f	p	r	f	p	r	f	p	r	f	p	r
cl1	0.36	0.5	0.281	0.226	0.923	0.129	0.279	0.778	0.17	0.219	0.923	0.124
cl2	0.357	0.778	0.231	0.364	0.5	0.286	0.319	0.923	0.193	0.63	0.5	0.85
cl3	0.298	0.923	0.177	0.738	0.667	0.828	0.738	0.667	0.828	0.421	0.5	0.364

Statistical Approach-2												
	kmeans			agglomerative(affinity='euclidean', linkage='ward')			agglomerative(affinity='euclidean', linkage='complete')			agglomerative(affinity='city block', linkage='complete')		
	f	p	r	f	p	r	f	p	r	f	p	r
cl1	0.374	0.5	0.298	0.307	0.513	0.219	0.463	0.98	0.303	0.303	0.521	0.214
cl2	0.581	0.98	0.413	0.378	0.5	0.304	0.522	0.634	0.444	0.63	0.5	0.85
cl3	0.515	0.656	0.424	0.714	0.61	0.862	0.714	0.61	0.862	0.216	0.267	0.182

Table 14: f, p, r scores for Statistical approach-1 and Statistical Approach-2 for different clustering algorithms with number of clusters=4

Statistical Approach-1												
	kmeans			agglomerative(affinity='euclidean', linkage='ward')			agglomerative(affinity='euclidean', linkage='complete')			agglomerative(affinity='city block', linkage='complete')		
	f	p	r	f	p	r	f	p	r	f	p	r
cl1	0.679	0.826	0.576	0.065	0.103	0.047	0.357	0.778	0.231	0.226	0.923	0.129
cl2	0.243	0.923	0.14	0.36	0.5	0.281	0.36	0.5	0.281	0.738	0.667	0.828
cl3	0.372	0.615	0.267	0.738	0.667	0.828	0.738	0.667	0.828	0.421	0.5	0.364
cl4	0.738	0.667	0.828	0.487	0.737	0.364	0.319	0.923	0.193	0.618	0.5	0.81

Statistical Approach-2												
	kmeans			agglomerative(affinity='euclidean', linkage='ward')			agglomerative(affinity='euclidean', linkage='complete')			agglomerative(affinity='city block', linkage='complete')		
	f	p	r	f	p	r	f	p	r	f	p	r
cl1	0.644	0.731	0.576	0.249	0.205	0.318	0.581	0.98	0.413	0.307	0.513	0.219
cl2	0.436	0.656	0.327	0.374	0.5	0.298	0.374	0.5	0.298	0.714	0.61	0.862
cl3	0.868	1	0.767	0.714	0.61	0.862	0.714	0.61	0.862	0.216	0.267	0.182
cl4	0.714	0.61	0.862	0.513	0.75	0.39	0.522	0.634	0.444	0.618	0.5	0.81

Table 15: f, p, r scores for Statistical approach-1 and Statistical Approach-2 for different clustering algorithms with number of clusters=5

Statistical Approach-1												
	kmeans			agglomerative(affinity='euclidean', linkage='ward')			agglomerative(affinity='euclidean', linkage='complete')			agglomerative(affinity='city block', linkage='complete')		
	f	p	r	f	p	r	f	p	r	f	p	r
cl1	0.293	0.923	0.174	0.36	0.5	0.281	0.36	0.5	0.281	0.257	0.923	0.149
cl2	0.487	0.737	0.364	0.319	0.923	0.193	0.319	0.923	0.193	0.738	0.667	0.828
cl3	0.679	0.826	0.576	0.738	0.667	0.828	0.738	0.667	0.828	0.421	0.5	0.364
cl4	0.738	0.667	0.828	0.487	0.737	0.364	0.487	0.737	0.364	0.63	0.5	0.85
cl5	0.372	0.615	0.267	0.372	0.615	0.267	0.372	0.615	0.267	0.372	0.615	0.267

Statistical Approach-2												
	kmeans			agglomerative(affinity='euclidean', linkage='ward')			agglomerative(affinity='euclidean', linkage='complete')			agglomerative(affinity='city block', linkage='complete')		
	f	p	r	f	p	r	f	p	r	f	p	r
cl1	0.496	0.648	0.401	0.374	0.5	0.298	0.374	0.5	0.298	0.339	0.513	0.253
cl2	0.513	0.75	0.39	0.522	0.634	0.444	0.522	0.634	0.444	0.714	0.61	0.862
cl3	0.644	0.731	0.576	0.714	0.61	0.862	0.714	0.61	0.862	0.216	0.267	0.182
cl4	0.714	0.61	0.862	0.513	0.75	0.39	0.513	0.75	0.39	0.63	0.5	0.85
cl5	0.868	1	0.767	0.868	1	0.767	0.868	1	0.767	0.868	1	0.767

In all the tables and figures cluster, statistical approach-1 and statistical approach-2 are represented as cl, sa1 and sa2 respectively.

From the above results it was understood that the Statistical Approach-2 performs much better than Statistical Approach-1.

V. CONCLUSION AND FUTURE SCOPE

The summary generated from the clustered medical documents help the patients in a better and prior understanding of the diseases, dosage, symptoms and precautions. In the Statistical Approach-2, the scores of the features are definite or crisp values. Fuzzification of the values of these features can be performed to generate a much better summary.

REFERENCES

1. An Extractive Summarization Technique for Text Documents International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-8 Issue-6, August 2019
2. Document Summarization Techniques International Journal of Computer Science Engineering (IJCSSE) ISSN: 2319-7323 Vol. 5 No.02 Mar 2016.
3. Enforcing Text Summarization using Fuzzy Logic International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 5 (6), 2014, 8276-8279.
4. Application of Clustering Algorithms to Group Medical Documents. International Journal of Computer Applications (0975 – 8887) Volume 178 – No. 42, August 2019
5. Review On Text Summarization Evaluation Methods Indian Journal of Computer Science and Engineering (IJCSSE)
6. Stubbs et al., “Practical applications for NLP in clinical research: The 2014 i2b2/UTHealth shared tasks,” in Proc. i2b2 2014 Shared Task Workshop Challenges Natural Lang. Process. Clinical Data, 2014.

AUTHORS PROFILE



Mrs. Ravi Seeta Sireesha, has received B. Tech degree from Jawaharlal Nehru Technological University, Andhra Pradesh and M.Tech from Andhra University, Visakhapatnam. The author is currently pursuing her Ph.D in the Department of Computer Science and Systems Engineering, Andhra University College of Engineering (A), Andhra University, Visakhapatnam.

Her main research work focuses on Natural Language Processing, Artificial Intelligence and Machine Learning. She has published some research papers in various International Journals and attended numerous workshops and conferences during her career. The author has seven years of experience in teaching and dealt with various subjects in Computer Science like Data Mining, Data Structures, Operating Systems, Computer Organization etc.



Prof. P S Avadhani pursued M.Tech and Ph.D from IIT Kanpur. He is currently working as a Professor in the Department of Computer Science and Systems Engineering, Andhra University College of Engineering (A), Andhra University, Visakhapatnam. His areas of interest are Cyber Security, Fuzzy Logic, Cyber Forensics, Computer Algorithms,

Public Cryptography, and Data and Network Security. The author has written 6 books in Computer Science Engineering on various subjects.

The author has published 52 papers in International Journals, 8 papers in National Journals, 38 papers in International Conferences and 9 papers in National Conferences. He has received several awards for his excellence in Teaching, Research and Administration like “State Best Teacher award”, “Best Researcher award”, “Distinguished Principal award” from the Government of Andhra Pradesh. The author had trained various PhD, M.Tech & B.Tech Engineering students during his 35 years of long teaching career.

