Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 215 (2022) 361–369

Procedia Computer Science

www.elsevier.com/locate/procedia

4th International Conference on Innovative Data Communication Technology and Application

Text Summarization Using Document and Sentence Clustering Method

Dr Sumathi Pawar[a*], Dr Manjula Gururaj H[b], Dr Niranajan N Chiplunar[c]

[a,b]*Associate Professor,Nitte(Deemed to be University),NMAMIT, Karkala, Karnataka, India*
[c]*Professor, Nitte (Deemed to be University),NMAMIT, Karkala, Karnataka, India*

Abstract

Text documents have important information and it will be very large in size. Getting the relevant information from the text document is very much challenging criteria in the field of information retrieval. This can be done using the text summarization method. A text document is compressed using a summarizing system to produce a new form that conveys the core idea of the content it contains. The issue of information overload demands access to reliable and properly crafted summaries. Users can quickly find the information they need using data minimization. Saving the time and effort from browsing through the entire collection of documents is main advantage of text summarization. The proposed system is focused on an extractive technique of text summarization using a text clustering and word-graph approach. The proposed System uses the term Frequency, Inverse Document Frequency (TFIDF) , Jaccard similarity and Euclidian distance which are important techniques for clustering the text. This hybrid approach deals with the novel method for comprising of document and sentence clustering.

*Keywords:*Extractive Summarization; Abstractive Summarization; Document Clustering; Multi-document summarization; TFIDF, Jaccard Similarity

1. Introduction

Text summarization [23][24] has grown importance since the early 1960s, although the necessity was different, since storage space was so scarce. Large hard copies required to be replaced with summaries for storage. There are many more inexpensive storage devices accessible now, but there is vast volume of data readily available, which is necessary to transform it into knowledge and information. When a user submits a query, the system responds with a collection of lengthy Web pages that contain so much information that it would be nearly difficult for a reader to read them all.

10.1016/j.procs.2022.12.038

The present rapidly evolving world has given a lot of attention to studies in automatic text summarization due to complexity of information sources and the exponential development in the quantity. The information and knowledge collected can be used for a variety of purposes, such as market analysis, industrial control, engineering design, business management and scientific research. A text that has been created from one or more papers and contains all of the crucial details from the original text is approximately known as a summary.

Topic identification, interpretation and summary creation are steps in the text summarization process. The summaries might be generic or pertinent to a particular question (query based summaries).

Extractive and Abstractive summarization are the two different kinds [17]. Human-generated summaries are a reflection of the author's grasp of the subject, conceptual synthesis, assessment and other processes. Since the consequence is unique and not specifically covered by the input, it is required to have access to knowledge that is separate from the input.

Alternative approaches must be taken into account because computers do not yet possess the same linguistic skills as people. The selection-based technique has so far been the most popular method for automated text summarization. According to this method, summaries are created by removing essential text segments from the text, which are then located by analyzing factors like phrase frequency and sentence placement.

Summarization of multi document is an automated process developed to extract data from numerous texts written on the same topic. The resulting summary makes it simple for individual users, including business people who need to rapidly acquaint themselves with information in a huge collection of papers.

The effort of summarizing many documents is far more difficult and time-consuming than summarizing just one item. This challenge is brought on by the range of themes found in a vast collection of papers. Completeness, readability and conciseness should all be incorporated as the guiding principles of a good summarizing technique [2].

In the current implementation, an extraction approach that first clusters documents and then clusters the phrases inside sentences of those clusters have been developed. Then word graph technique is used to summarize the text. The efficiency and significant reduction in redundancy of the findings are outstanding.

2. Literature Review

Numerous researches already held in the field of text summarization. The sentence summarization is done using different types of methods. According to the process outlined in [3], a large number of texts are clustered using a document clustering method and a summary of the clusters is produced using a profile-oriented sentence extraction strategy of features. Relevant documents are grouped into a single cluster using a threshold-based approach to document clustering. Sentence-position, word-weight, sentence-centrality, sentence-length, proper nouns and numerical information in the sentence are all taken into account while creating a feature profile. Each sentence receives a sentence score based on its feature profile. The TSF-ISF (Term Synonym Frequency-Inverse Sentence Frequency) approach is used to determine each word's weight. Using varying compression rates, sentences are extracted from each cluster and then scored according to their importance. As result of text summarization, the extracted sentences are put in the same chronological sequence as the source documents and a summary is created from each cluster. The final result is a succinct summary of the input documents that is clustered according to information density.

The different types of clustering methods are used while performing the text summarization, one of them is proposed by Abualigah et al [4]. The authors described three crucial criteria to be considered into account during the process of text summarization. They are (1) Grouping of sentences (2) choosing representative phrases from the groups which is followed by cluster ranking. Authors used phrase similarity clustering and histogram-based incremental clustering. Both clustering algorithms are unsupervised learning algorithms, which are used have the property to build sentences, groups incrementally and dynamically. A cluster's significance is determined by the number of significant words in it. The top 'N'(number) clusters will be chosen after the clusters are assessed in decreasing order of scoring. One exemplary sentence is selected from each cluster and included to the summary. Sentences are still being chosen up to a predetermined summary size.

A query-based document summarizer using word frequency and sentence similarity is provided by Nenkova et al [5]. The suggested model uses the Vector Space Model and word frequency to identify words that are similar to the query. Redundancy is reduced using sentences and word frequency in this query-based summarizer, which brings

together inquiries that are linked to one another. The suggested approach starts with handling the query, then the summarizer compiles the necessary resources and generates the summary.

The authors used the following procedures to create the summary after pre-processing:

- Finding similarity between sentences in documents and the user query.
- After doing the similarity calculation, sorting sentences according to how similar they are.
- Calculating sentence score by utilizing the sentence location and word frequency data.
- Choosing the phrases from each category with the highest scores and including them in the summary.
- Keeping the summary to no more than 100 words.

Haque et al [6] proposes text document collection, training and testing procedures. Text files in Indonesian were used for the documents. The three main components of the training procedure were text characteristics, genetic algorithm modeling and a summary article. As part of this procedure, the document was manually summarized by three different people. The process for extracting text features made it possible to get the document's text. During the modelling phase of genetic algorithms, a search technique is utilized to determine the ideal weighting for each text feature extraction. The fitness function which is generated using stages, summaries and manually extracted text qualities were used to evaluate the chromosomes. The testing phase used 50 documents and the training phase used a different set of documents.

A novel approach for employing neural networks to summarize news stories was introduced by Widyassari et al [7]. Phrase selection, feature fusion and neural network training made up the three stages of the process. The neural network's input could be either binary or real vectors. In order to recognize the choice of sentences based on their word choice, the initial stage involved is teaching a neural network the essential characteristics of phrases that should be stored in the article summary. It was trained on a corpus of articles. The neural network was then modified to include and generalize the pertinent components that might be included in summaries. Finding patterns and linkages between the traits that are present in most sentences is necessary after learning the qualities using neural networks. It uses the feature fusion phase method, which entails two stages: removing odd features and condensing the impact of common features. Through feature fusion, the network discovers the importance (and unimportance) of many traits that are utilized to determine each sentence which is suitable for summarization. The upgraded neural network is then used to process the filtered news article summaries.

The Open NLP tool for natural language processing and word matching was introduced by Allahyari et al [8]. Data mining document clustering techniques are used to extract useful and query-dependent information from a vast collection of offline documents. This article serves as an example of how people can download paragraphs by uploading document files. A paragraph's individual sentences are treated as nodes. The syntactic relationship between each node is determined using the traditional edge weight method. A paragraph is represented by each node in this structure, which is known as a "document graph" and the connections between the nodes are shown by each edge. The clustering algorithm receives the document graph and uses the effective clustering method K-Means to categorize the nodes based on their edge weights and create clusters. The cluster's nodes all continue to be associated with one another. The bio medical text summarization is done by Wang at al [9].

Following a thorough analysis of Jani et al [10], text summarizing techniques are categorized as extractive and abstractive techniques. The authors analyzed each approach's benefits and drawbacks while working on the abstract method of summarizing. At the conclusion of the essay, the writers looked at potential problems and enhancements for various text summary techniques. Everything is presented in an organized manner. Multi document summarization is introduced by Shah et al. [11] and the authors examine various methods and their drawbacks. For writers who work in the field of text mining, this information is useful.

According to S. V. Mokhale et al. [12], summarizing content is necessary to extract pertinent information from several documents. The authors also addressed a review of several techniques to summarize multiple documents due to the significant increase in material that needs simply a summary to be retrieved in a short amount of time. M. Y. Saeed et al [13] combined metadata of documents and also analyzed unstructured documents by multi stage clustering. Adjacency graphs are generated to link generated clusters. Multistage-clustering and interlinking with sub-corpuses are performed by authors. Authors applied their approach to new data set and processed six different metadata

combinations over text queries and resulted into 67% associated text. This approach is evaluated by the SHAP (SHapley Additive exPlanations) model.

R. Alqaisi et al [14] proposed summarization system of multi document on to Arabic text. They proposed multi-objective clustering-based system which resulted into main topics by cluster-based methods. Relevance, Coverage and diversity are three main objectives of text summarization. Authors evaluated the system using [15] DUC (Data Set Understanding Conference) 2002 and [16] TAC 2011 data sets. Authors achieved F-Measure of 47% on DUC 2022 dataset.

Sumathi et.al[20][21][22] focused on creating, publishing, searching and composing Web services. If the proposed research method of text summarization is published as Web service, then it will be very convenient to the consumer to consume it whenever required independent of any platform.

H. Siddiqui et al [15] worked on document frequency (TF-IDF), latent Semantic analysis and text-rank calculation methods to get text summarization. They concluded that Bias will be less if text is summarized with machine than human. Thus through survey of these papers, we have found different methodologies for text summarization.

3. Methodology

The proposed system aims at summarizing the text for faster retrieval of information. Clustering of documents is depending on *TFIDF* and Jaccard similarity of lemmatized terms which is explained in this section in detail.

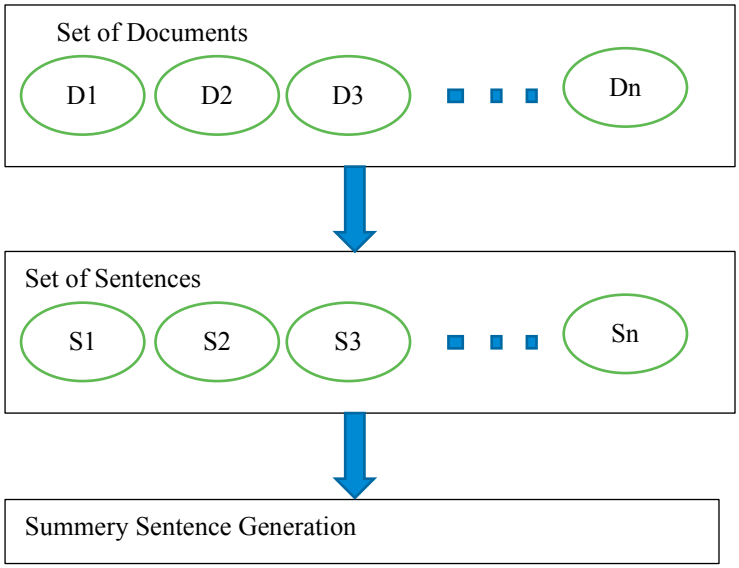

Fig.1. Outline of the system

The outline of the system is shown in figure 1. Set of documents are used for generating the summery of sentences using clustering method. A hybrid approach of clustering method is used in the proposed system. At first terms are extracted from the documents using lemmatization process. WordNet lexical dictionary [3] used to perform lemmatization and to extract terms. Clustering of document is performed based on lemmatized terms.

3.1 Clustering the documents

Text clustering is a technique where clustering of similar documents iscombinedinto one cluster. Consider the document set D1, D2,…Dn, where n represents number of documents in the dataset, Each term t_i in document d_j is converted into vector. Weight of the vector is represented in as W_{ij}, where W is weight of i^{th} term in j^{th} document.

The weight is calculated as $W_{i,j} = TFIDF_{(i,j)} = tf_{(i,j)} * \left(\frac{log_n}{df_{(j)}}\right)$ (1)

Where $tf_{(i,j)}$ is in document i frequency of the term j

$df_{(j)}$ is total number of documents that contain term j.

The documents which have 50% of *TFIDF* [19] weight age are grouped into one cluster. Thus, k number of document clusters are formed.

3.2 Jaccard similarity for sentence clustering

According to the similarity of sentences, the document need to be clustered using Jaccard technique[1]. Jaccard similarity is measured as follows.

$$J(A,B) = \frac{A \cap B}{A \cup B} \quad (2)$$

Here A, B are vectors contain terms of two documents A, B. A∩B is number of same terms that available in both documents, $A \cup B$ is number of unique terms available in both A and B.

To cluster the documents into n number of clusters following procedure is used.

- First we need to take two documents d_i, d_j, here d_i is source document and d_j is target document.
- Measure the similarity using Jaccard similarity using Jaccard(di,dj).
- Take a threshold value of Jaccard Co-efficient and cluster the documents into one cluster which gets Jaccard similarity above threshold.
- Repeat above procedure for remaining documents by incrementing ‘i’ and ‘j’ where i<=n and i > j< n, n is number of documents.
- Count the number of clusters formed.

3.3 Finding similarity between clusters

- To measure the Similarity between clusters, cluster centroid need to be calculated.
- Find the dissimilarity or distance between different cluster centroid using Euclidian distance.
- Take a threshold value and group the clusters which get the value above threshold
- Calculate the scores of each cluster according to the distance in reverse order
- Pick the Clusters which get more score and generate summery.

3.3.1 Euclidean distance to find similarity of clusters

Euclidean distance [18] is used to find distance between centroid of 2 clusters. Let us take centroid points of first cluster as p1, p2 and center of second cluster’s points as p3, p4.

The Euclidean distance is calculated by

$$d = \sqrt{[(p3 - p1)^2 + (p4 - p2)^2]} \quad (3)$$

Where,

"d" is the Euclidean distance
(p1, p3) are first coordinates of the center of first cluster and second cluster.
(p4, p2) are second coordinates of the center of second cluster and first cluster.

3.4 Generating summery of sentences

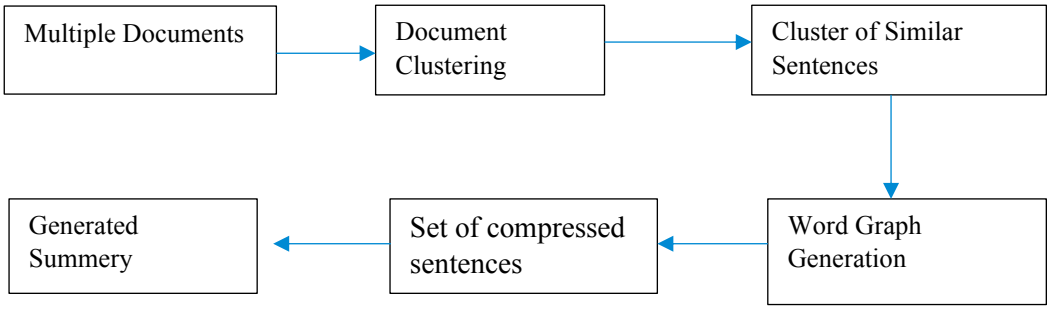

Fig. 2. Procedure of generating summery of sentences

A word-graph technique is used to generate summery from more than one cluster. Here abstractive summarization is ensured by removing very similar paths. Similar paths are calculated using cosine similarity. Multi sentence compression is used in order to generate one-sentence representation from a cluster of redundant sentences.

3.4.1 Word graph technique

In word graph technique the sequence of words is shown in the form of graphs. Fig 3 shows the word graph of given sentences. Consider following 2 sentences.

"Text summarization using clustering method"
"Clustering technique for text summarization method"

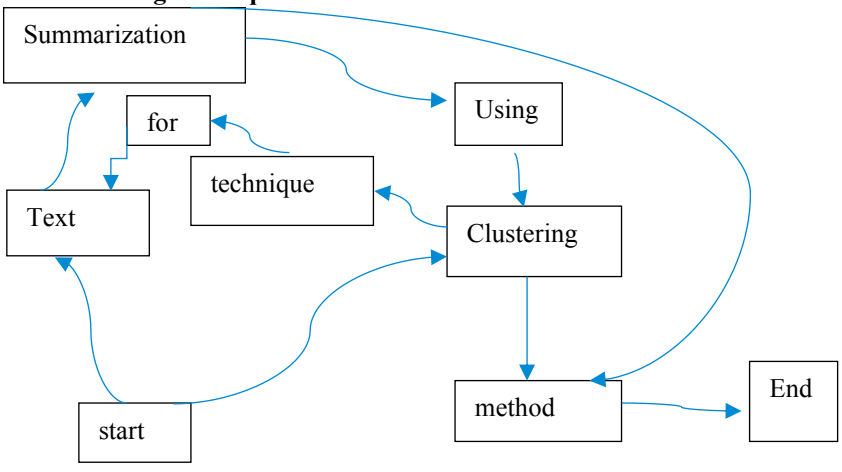

Fig.3. Word-graph

At first word graph is drawn for these sentences. Start with first sentence and add nodes for words connected with other nodes with edges. When another sentence appears, start from "start" node and connect to existing nodes where same word appears, otherwise add a new node for different words. A good compression method is finding shortest path for the graph to provide fast summery. Outline summery from multiple sentences is achieved by deleting similar paths.

4. Experimental results & analysis

Above procedure is applied with Kaggle [17] dataset for clustering. Using performance and accuracy metrics like Precision, Recall, and F-measure, this system compared the findings produced by various extraction approaches.

We used Python to implement with ***nltk*** libraries. Sample procedure is explained below.

```
from scipy.spatial.distance import jaccard

doc1 = set(d1.lower().split())
doc2 = set(d2.lower().split())

 # To find intersection between doc1 and doc2
 intersec = doc1.intersection(doc2)

# To find union of words between doc1 and doc2
   union = doc1.union(doc2)
 # To calculate Jaccard similarity

return float(len(intersection)) / len(union)
doc1 = "The mango is green in color"
doc2 = "The pineapple is yellow in color"
Jaccard(doc1,doc2)
```

Above code snippet gives Jaccard similarity of vectors doc1 and doc2.

In the above code doc1 and doc2 contain list of words in 2 documents after removal of stop words.

According to Jaccard similarity documents are clustered.

After summarization precision is calculated using below equation.

$$\text{Precision} = \frac{Retrieved \cap relevant}{retrieved} \quad (4)$$

$$\text{Recall} = \frac{Retrieved \cap relevant}{relevant} \quad (5)$$

$$\text{F-Measure} = \frac{2 * Precision * Recall}{(Precision + Recall)} \quad (6)$$

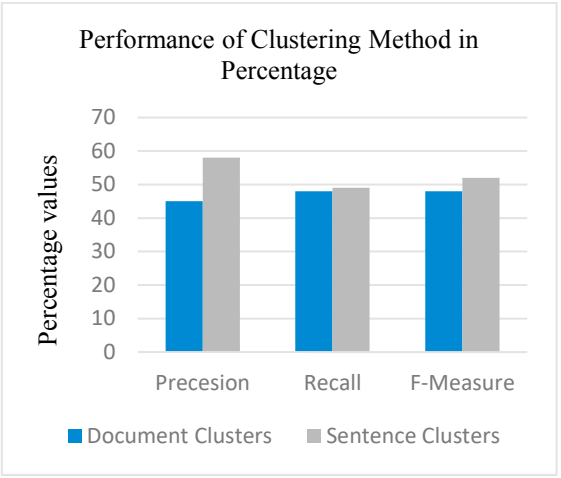

Fig.4. Performance of Clustering Method

In Fig 4, it is shown the outcomes of conducting summarising process on a document set. The result demonstrates that proposed research, cutting-edge "Sentence Clustering based Text Summarization" technique surpasses the document clustering technique.

5. Conclusion and future work

This research is focused on extractive summarization methods. According to the results, this innovative strategy performs better than the other methods and lessens clustering-related redundancy. In the future, we would like to improve the system by including sentence-simplification methods for generating summaries, which can be applied to shorten complex and long sentences. The future system may also use a paraphrase approach to give a summary in an abstract sense.

References

[1] Jaccard similarity and Jaccard distance in Python, https://pyshark.com/jaccard-similarity-and-jaccard-distance-in-python, Retrieved on 12/8/2021

[2]A. Kogilavani, Dr.P.Balasubramani, “Clustering And Feature Specific Sentence Extraction Based Summarization of Multiple Documents”, International journal of computer science & information Technology, vol.2, no.4, Aug. 2010.

[3] “Multi-document summarization”, Wikipedia, the free encyclopedia, 2012.

[4] Abualigah, L., Bashabsheh, M.Q., Alabool, H., Shehab, M. (2020), Text Summarization: A Brief Review. In: Abd Elaziz, M., Al-qaness, M., Ewees, A., Dahou, A. (eds) Recent Advances in NLP: The Case of Arabic Language. Studies in Computational Intelligence, vol 874. Springer,

[5]Nenkova, Ani, Kathleen McKeown "A survey of text summarization techniques." In Mining text data, pp. 43-76. Springer, Boston, MA, 2012.

[6] Haque, Majharul, Suraiya Pervin, and Zerina Begum. "Literature review of automatic multiple documents text summarization." International Journal of Innovation and Applied Studies 3, no. 1 (2013), 121-129.

[7] Widyassari, Adhika Pramita, Supriadi Rustad, Guruh Fajar Shidik, Edi Noersasongko, Abdul Syukur, and AffandyAffandy, "Review of automatic text summarization techniques & methods", Journal of King Saud University-Computer and Information Sciences (2020).

[8]Allahyari, Mehdi, SeyedaminPouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys Kochut, "Text summarization techniques: a brief survey." arXiv preprint arXiv:1707.02268 (2017).

[9] Wang, Mengqian, Manhua Wang, Fei Yu, Yue Yang, Jennifer Walker and Javed Mostafa, "A systematic review of automatic text summarization for biomedical literature and EHRs", Journal of the American Medical Informatics Association 28, no. 10 (2021): 2287-2297.

[10] Jani, D., Patel, N., Yadav, H., Suthar, S., Patel, S. (2022). A Concise Review on Automatic Text Summarization. In: Nayak, J., Behera, H., Naik, B., Vimal, S., Pelusi, D. (eds) Computational Intelligence in Data Mining. Smart Innovation, Systems and Technologies, vol 281. Springer, Singapore.

[11] Shah, C., Jivani, A. (2016). Literature Study on Multi-document Text Summarization Techniques. In: Unal, A., Nayak, M., Mishra, D.K., Singh, D., Joshi, A. (eds) Smart Trends in Information Technology and Computer Communications. SmartCom 2016. Communications in Computer and Information Science, vol 628. Springer, Singapore.

[12]S. V. Mokhale and G. M. Dhopawkar, "A Study on Different Multi-Document Summarization Techniques," 2019 Third International Conference on Inventive Systems and Control (ICISC), 2019, pp. 710-713, doi: 10.1109/ICISC44355.2019.9036387.

[13] M. Y. Saeed, M. Awais, R. Talib and M. Younas, "Unstructured Text Documents Summarization With Multi-StageClustering," in IEEE Access, vol. 8, pp. 212838-212854, 2020, doi: 10.1109/ACCESS.2020.3040506.

[14]R. Alqaisi, W. Ghanem and A. Qaroush, "Extractive Multi-Document Arabic Text Summarization Using Evolutionary Multi-Objective Optimization With K-Medoid Clustering," in IEEE Access, vol. 8, pp. 228206-228224, 2020, doi: 10.1109/ACCESS.2020.3046494.

[15] TAC 2011 Summarization Track, https://tac.nist.gov/2011/Summarization, Retrieved on 28/8/2021

[16]https://www-nlpir.nist.gov/projects/duc /guidelines/2002.html

[17] https://www.kaggle.com/datasets

[18] Euclidean Distance - Definition, Formula, Derivation & Examples (byjus.com) (https://byjus.com/maths/euclidean-distance/)

[19]Akash Panchal "NLP — Text Summarization using NLTK: TF-IDF Algorithm Towards Data Science", https://towardsdatascience.com/text-summarization-using-tf-idf-e64a0644ace3

[20] Sumathi, Dr. Niranjan N Chiplunkar, "Populating Parameters of Web Services by Automatic Composition Using Search Precision and WSDL Weight Matrix", IJCSE, 1742-7193, InderScience Publishers – Scopus indexed, vol 13, No 7, 2018, DOI: 10.1504/IJCSE.2016.10007953. Citations 3.

[21] Sumathi Pawar , Niranjan N. Chiplunkar , "Survey on Discovery of Web Services", Print ISSN : 0974-6846 , Online ISSN : 0974-5645, Indian Journal of Science & Technology, Vol 11(16), DOI: 10.17485/ijst/2018/v11i16/120397, April 2018 pp 1-10 (indexed by Thomson Reuters "Web of Science" Citations 1.

[22]Sumathi, Dr. Niranjan N Chiplunkar "Open Source APIs for Processing the XML Result of Web Services", International Conference on Advances in Computing, Communications and Informatics (ICACCI) (Scopus indexed & Thomson Reuters indexed conference) MIT – Manipal, September 14-16 2017 Published on December 04,2018

[23] Jacob, I. Jeena, "Performance evaluation of caps-net based multitask learning architecture for text classification" ,Journal of Artificial Intelligence 2, no. 01 (2020): 1-10.

[24] Adam, Edriss Eisa Babikir, "Deep Learning based NLP Techniques in Text to Speech Synthesis for Communication Recognition", Journal of Soft Computing Paradigm (JSCP) 2, no. 04 (2020): 209-215.