IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 2, 2020, accepted November 19, 2020, date of publication November 25, 2020,
date of current version December 9, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3040506

Unstructured Text Documents Summarization With Multi-Stage Clustering

MUHAMMAD YAHYA SAEED[1,2], MUHAMMAD AWAIS[2], RAMZAN TALIB[1], AND MUHAMMAD YOUNAS[1], (Member, IEEE)
[1]Department of Computer Science, Government College University Faisalabad, Faisalabad 38000, Pakistan
[2]Department of Software Engineering, Government College University Faisalabad, Faisalabad 38000, Pakistan

Corresponding author: Muhammad Awais (muhammadawais@gcuf.edu.pk)

ABSTRACT In natural language processing, text summarization is an important application used to extract desired information by reducing large text. Existing studies use keyword-based algorithms for grouping text, which do not give the documents' actual theme. Our proposed dynamic corpus creation mechanism combines metadata with summarized extracted text. The proposed approach analyzes the mesh of multiple unstructured documents and generates a linked set of multiple weighted nodes by applying multistage Clustering. We have generated adjacency graphs to link the clusters of various collections of documents. This approach comprises of ten steps: pre-processing, making multiple corpuses, first stage clustering, creating sub-corpuses, interlinking sub-corpuses, creating page rank keyword dictionary of each sub-corpus, second stage clustering, path creation among clusters of sub-corpuses, text processing by forward and backward propagation for results generation. The outcome of this technique consists of interlinked sub-corpuses through clusters. We have applied our approach to a News dataset, and this interlinked corpus processing follows step by step clustering to search the most relevant parts of the corpus with less cost, time, and improve content detection. We have applied six different metadata processing combinations over multiple text queries to compare results during our experimentation. The comparison results of text satisfaction show that Page-Rank keywords give 38% related text, single-stage Clustering gives 46%, two-stage Clustering gives 54%, and the proposed technique gives 67% associated text. Furthermore, this approach covers/searches the relevant data with a range of most to less relevant content. It provides the systematic query-relevant corpus processing mechanism, which automatically selects the most relevant sub-corpus through dynamic path selection. We used the SHAP model to evaluate the proposed technique, and our evaluation results proved that the proposed mechanism improved text processing. Moreover, combining text summarization features, shown satisfactory results compared to the summaries generated by general models of abstractive & extractive summarization.

INDEX TERMS Cosine similarity, page rank keywords, k-means, word2vec, summarized parallel corpus.

I. INTRODUCTION

The large number of unstructured text documents exist for use in daily life. It is not easy to process them without an automatic approach. Automatic corpus processing approaches rely on dynamic information grouping in text retrieval. The efficiency of text-documents grouping decreases the size of the large corpus. However, grouping text documents by splitting multiple text-documents into related subsets is a problematic text processing task [1], [3], [7]. Furthermore, the efficiency of information retrieval approaches decreases in the absence of such grouping of corpus [2], [3]. Less efficient grouping ultimately increases the cost and effort involved in the information retrieval process. Text mining experts determine the processing cost & time by the size and information diversity of the underlying corpus [4], [5]. Text processing efficiency improves by consuming less time with a small related corpus. Therefore, the optimal use of time and cost for corpus processing is crucial in the information retrieval system. To resolve this issue, metadata analysis and relevant text extraction have a significant role in optimizing the processing effort, cost, and time [6], [8], [17].

The associate editor coordinating the review of this manuscript and approving it for publication was Muhammad Afzal.

The text summarization processes rely on text classification by machine learning to perform text mining. This extraction process attempts to extract important sentences or those sentences carrying salient words. In the text extraction process, we refer these sentences as key-sentences [2], [4], [8], [59], [61]. The reason to select abstractive & extractive summarization is to overcome sentence structures, active or passive voice, etc. These forms of unstructured-text summarizations give the language-independence advantages in text-query processing [1], [3], [9], [58], [60].

The automatic summarization of unstructured-text has various challenges and problems. The basic reason corresponds to its robustness & completeness to cover/represent the actual text, which fewer words/lines. It is challenging to guarantee that short-extracted text can give all the underlying text information. In literature, researchers study this problem of text representation as two strategies: abstraction & extraction. Abstractive or metadata processing technique gives a keywords-based theme about the text as abstractive summarization [13], [24]. Collective metadata of similar documents leads to a relevant set of documents rather than processing documents with individual metadata. The process of keywords-based Clustering generates groups of the metadata in the form of a dictionary [8], [9]. This process faces challenges like semantic-restrictions language-dependency. However, still, it has vast application over those documents which contain relatively short unstructured text. It is very much suitable for tasks such as headlines, various documents title, website or research paper keywords [5], [21], sentence compression [13], [19], [24], and sentence fusion [11], [17]. The keyword selection has many concerns and constraints, and as per literature, one cannot guarantee it as stand-alone the comprehensive text representation.

After the metadata processing, extractive summarization reduces the corpus by removing less necessary text [3], [10]. The potential sentences are either selected through the sentence score algorithm or word embedding principles, i.e., Word2Vec model [11], [18], [29]. Recently, the frequency-based weighting techniques are the preliminary researches in sentence extraction studies [9], [14], [31]. Similarly, the semantic assessment by latent analysis [19], [27], [39], Markov-models, and graph-oriented supervised and unsupervised techniques are under investigation [4], [16], [27]. In short, the extractive-summarization is a key area of research as per the current literature analysis [2], [18], [29], [37]. For simplifying the extractive summarization, the sentence selection process considers the summary-worthiness. The sentence ranking process extracts the most salient 'n' sentences. These processes essentially relate to the classification problems of machine learning. In literature, researchers mostly apply either syntactic and semantic approaches for deciding sentence scores. We implement the predetermined features during the syntactic approach, which are usually hand-crafted for those sentences that we consider for scoring the sentences [8], [13], [22], [39]. Semantic approaches involve the word's meaning and various types of text phrases to create a semantic relation [9], [17], [28], [41]. However, during summarization, both the text's meanings and their semantic relations in written sentences rely on the text's structural properties.

As per our literature review, there is an intellectual space to comprehensively combine abstractive & extractive summarization to handle the text representation, and these techniques need attention. There exist few studies which target syntactic & semantic issues in a combined manner. This combination can solve unstructured text document analysis for both the document-dependent & document-independent issues of unstructured text summarization. Although the presented researches have shown the capacity of identifying the relevant text, even then, these studies have less attention to minimize corpus diversity by creating a runtime relevant reduced corpus by utilizing multistage Clustering & summarizations. These studies rely only on sentences and overlook various syntactic text, which exists in unstructured text. Therefore, these techniques are not mature or final solutions for handling diverse information in an unstructured text corpus.

The proposed technique is a new text processing way to give a path-oriented visualization to the actual large corpus. The current study's first objective is to systematically create improved corpus visualization, combining clustering, abstractive, and extractive summarization. This form of corpus visualization has path-oriented query processing to handle any extended text corpus's diverse nature. A further objective of this study is to perform improved metadata analysis by relying on pre-extracted text. This study binds the keywords metadata with reduced extracted text, and it achieves the goal of extracting crucial parts of the actual-text. Another objective of this technique is to develop a mechanism that has combined features of single&multiple document summarizations. We have achieved all these objectives by interlinking single document summarization metadata through cluster-oriented inter-linking & summarization.

This research article consists of eight parts. Section-I describes the main aspects of the paper and contains a complete description of the current research paper. Section-II gives a briefing about the existing related work. It shows the comparison of the associated studies. Section-III provides an overview of the techniques applied in the current study. The next two sections cover the methodology and results. Section-IV is about the implemented steps of the proposed technique. The experimental results as graphs and tables, including their brief Description, are part of Section-V. Section-VI is the summary of our presented work. We have mentioned the detailed summary of the entire work in Section-VII. Section-VIII is the conclusion of this article.

II. RELATED WORK

Several text processing systems exist to process unstructured textual documents from multiple unstructured sources of free text. The existing information extraction methods use the dictionary-based approaches, pattern matching approaches, and rule-based systems [14], [15]. The detail of related work is as follows: Liu, K., & El-Gohary, N. (2017) applied abstractive summarization over maintenance action reports of bridge conditions and processed them by applying automatic metadata analysis. Abstractive extraction served as a tool for bridge deterioration prediction related to decision making about bridge maintenance. Their work presented a semi-supervised information extraction procedure for decision-making information mining. Researchers generated keywords dictionaries to unfold current deficits and maintenance actions from inspection reports of the bridges. Their approach defined dependencies structures by abstractive summarization for both labeled and unlabeled data.

Ying, Y. *et al.* (2017) implemented abstractive text summarization for key phrases extraction. They applied a graph-based technique for key phrases ranking during text extraction. They selected an abstracted text extraction approach to create connections between the keywords and ignore the sentencing impact. They proposed certain types of metadata relationships between standard terms to their related sentences. First, they grouped multiple documents as a set of clusters, and these clusters presented the main topics of the given papers. They developed three different metrics and proved their work better than just extracting the key phrases.

Gupta, A. *et al.* (2018) analyzed the radiology reports by abstractive summarization and Clustering to overcome the previous difficulties of manual text processing. They improved dictionary-based and rule-based approaches with better keywords extraction and text grouping. At earlier processing techniques, the named entity recognition process missed the cluster relationship analysis. They extracted named entities relations by unsupervised approach, without prior knowledge. They covered text processing dependencies by parse trees and related them in the form of distributed semantics.

Moradi, M. (2018) used extractive text summarization and reduced the text by eliminating fewer essential parts. This technique applied text summarization over bio-medical data, and they named it as Clustering & Itemset-mining Biomedical Summarizer (CIBS). This approach processed the text input to extract biomedical concepts. The concepts represented the main topics by applying the itemset mining algorithm over reduced text, and CIBS placed sentences into the clusters' relevant set.

Azadani, M. N., *et al.* (2018) worked over abstractive & extractive text summarization. They developed a solution for scientific and clinical literature summaries to maintain valuable, informative content. They applied graph-oriented summarization in domain-specific knowledge mining for frequent itemset identification. This summarizer implemented the Unified Medical Language System and provided a concept-based text mining model. It processed source documents by mapping concepts to actual records. It provided interrelated item-sets with correlation similarity function to generate graphs.

Uçkan *et al.* (2020) proposed a text summarization technique names KUSH. This technique identifies the maximum possible sets of un-overlapping abstractive summarization. They marked these sets as nodes to determine the context of various paragraphs in unstructured text documents. They focused their work to generate coherent text visualization. Their proposed KUSH technique used abstractive summarization, integrated with set theory and graph visualization methods.

Sanchez-Gomez *et al.* (2020) developed a multi-objective abstractive & extractive summarization-based context analyzer. Context defines critical objectives and operational usability. For the functional usability analysis, the keywords-based approach identified the main parts of context in the paragraphs. The summarizer utilizes pre-defined criteria for text selection. This work focus on the rising demand for automatic text summarization methods.

Deng *et al.* (2020) pointed out the lack of unknown words and incomplete sentences. They presented an abstractive-text summarization approach as an alternative to the sequence-to-sequence text summarization models, which integrated text-summarization with sequence-to-sequence models in Chinese text assessment. They included adversarial learning in their proposed text summarizer. Their comparison results proved that the proposed method improved text assessment with the addition of abstractive text summarization.

Mohd *et al.* (2020) tried to capture & preserve the text's semantics as the fundamental feature for summarizing a document. To generate high-quality text, summarization researchers applied the distributional-semantic-model with abstractive & extractive summarization. These summaries proved suitable by ROUGE-summarizer over the DUC-2007-dataset. The main contribution of this article is to include summarized semantics as part of text assessment features.

Bidoki *et al.* (2020) worked over a multi-document text summarizer and developed an extractive summarization-based semantic framework. This semantic framework combines machine learning & graphs with text summarization and extracts sentences representing semantics from a set of text-documents with the word2vec model. It calculates meaningful sentences by applying graph relations to the documents. This method helped to identify relevant topics to text documents.

Mutlu *et al.* (2020) presented a new dataset for abstractive and extractive summarization tasks in this study. This dataset consists of miscellaneous academic-publications with the abstracts and the human extracted text from these papers. This method combines the keywords of these three parts and joins them with the critical extracted sentences. This form of academic publication presentation helps to assess the validity of the text extraction process. This method reinvestigates the robustness of the extractive summarization process over their dataset. This study focused the semantic features by using GloVe & word2vec embeddings.

Liu, K., & El-Gohary, N. (2017) only made keywords metadata & phrases comparisons for logical grouping and did not consider sentences to predicted possible defects in text grouping. Ying, Y. *et al.* (2017) processed metadata to build a graph-based representation of extracted text. This graph-based approach lacks a clustering approach for query-specific corpus processing. Gupta, A. *et al.* (2018) proposed a formal structured system reduced manual processing. This work performed knowledge finding main with the pre-defined output structure and without applying corpus assessment approaches. Moradi, M. (2018) developed the topic selection technique with the clustering algorithm only over key phrases metadata. Azadani, M. N., *et al.* (2018) applied a minimum spanning tree and included tree traversal to mark subthemes. The knowledge finding process only depended on the metadata and information processing trees without sentence analysis. Uçkan *et al.* (2020) only applied graph theory and evaluated the designated nodes by abstractive summarizations. Sanchez-Gomez *et al.* (2020), instead of developing summarizers & text classification, followed multiple pre-defined criteria to achieve their multi-objective analysis of the unstructured text. Deng *et al.* (2020) propose a two-stage abstractive summarization method with adversarial learning, but they did not apply text clustering at any stage. Mohd *et al.* (2020) reduced redundancies from the input source only as part of text cleansing but instead corpus reduction. Bidoki *et al.* (2020) do not select the best sentences from appropriate clusters for the final summary concerning essential issues like information salience, minimum redundancy, and adequate coverage. Mutlu *et al.* (2020) lack to provide syntactic closeness and the impact of text grouping on the resulting summaries.

Table 1 contains the summary and comparisons of the features of the existing studies. Most of these approaches involve significant text processing and ultimately consume more time. In these approaches, less attention focused on creating a strong relationship between metadata to actual text with some pre-extracted text. These techniques handle the metadata quality to overcome the corpus's diversity to obtain the relevant text query results. Therefore, these studies relate to our research regarding the enhanced corpus processing visualization. Table 1 presents the essential work and limitations of the described tasks, and it shows the salient features of our proposed technique [1]–[11].

III. TEXT PROCESSING TECHNIQUES OVERVIEW

Text processing relies on multiple approaches to identify the exactly required text. These approaches are different as compared to the data processing applied in spreadsheets and databases [6], [17]. As the initial step, text processing techniques analyze the textual data using morphological analysis and wordless analysis [8], [13], [21]. Next, these techniques classify text into useful to less useful grouping [16], [17], [21]. This section contains a description of the text processing techniques.

A. TEXT PRE-PROCESSING

Preliminary text analysis is the first step in text processing and comprises information search, formation, and extraction by text mining methods [11], [16], [28]. The data cleansing removes elements like exclusive characters, punctuation, and tags. All such items are useless for the underlying text processing and only result as unnecessary noise [5], [8], [21]. However, there exists no fixed rule to categorically declare the specific noise in any type of textual data. Removing anything from the source text depends on the problem statement [2], [14], [19]. Generally, after further tokenization, steps are removing stop words, stemming, and lemmatization [4], [13], [22].

B. DICTIONARY-BASED APPROACHES

There are two different approaches for Morphological Analysis (MA), i.e., dictionary-based MA and MA without dictionaries. The vocabulary of the language provides a base for MA. The wording of natural language has many compound words or constructions using punctuation marks like hyphens and apostrophes [13], [25], [41]. The vocabulary approach comprising MA faces two challenges, i.e., the incompleteness of the dictionaries and the constant appearance of new words [21], [22], [49]. In MA methods, wordless analysis supplements all those words which are not present in the dictionary. Wordless morphology uses the regularities according to inflections and grammatical meanings [21], [37], [44]. The principle of this morphology relies on analyzing the end of a word for predicting its morphological features. [26], [37]. The English language MA is based on the use of finite state machines and finite state transducer [23], [24], [31].

C. TEXT RELEVANCE BY TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (TF-IDF)

The method of TF-IDF determines text relevance in search engines. TF is the ratio of the number of occurrences of a certain word, to the number of words in the document. For example, if a document contains 500 words, and the word 'cosine' occurs five times, then TF $= 5/500 = 0.01$. IDF is the inverse document frequency and the inverse of the frequency with which a word occurs in a large collection of documents. This collection consists of all the pages which indexer processes. For example, if the collection contains 10000 documents, and the word 'cosine' appears in 100 documents, then the value IDF $= \log2\ (10000 / 100) = \log10\ (100) = 2$. Thus, popular words occurring in every tenth document more often will have IDF<1. The words found in every hundred documents and occurs less often have IDF>2. Almost all topics characterization likelihood, by certain words, has IDF close to 2 [24], [25], [44]. TF-IDF gives maximum value if rare words have many occurrences in the document [26], [29].

TABLE 1. Comparative analysis of existing studies.

Basic work and study references	Meta data processing.	Corpus processing prioritized.	Key phrases linked.	Summarized text involved.	Path selection made.	Knowledge path finding.	Similar text grouping.	Layered approach.	Sub-corpuses creation.	Neural network involved.	Similarity continuously checked.	Limitations
Clustering and abstractive-summarization based text mining [1]	✓	X	✓	X	X	X	✓	X	X	X	X	Supplement instead actually Improving decision making and prediction process.
Graph-based keyphrase extraction & ranking by text summarization algorithms [2]	✓	✓	✓	X	✓	X	X	✓	X	X	✓	Group the documents for only similarity recommendations.
Unsupervised parse tree processing approach for named entity recognition by abstractive summarization based distributed semantics [3]	✓	✓	✓	X	X	✓	X	X	X	X	X	Problem of manual and dictionary-based approach is not fully automated.
Item-set mining by bio abstractive text-summarizer & keyphrase matching [4]	✓	X	X	✓	✓	X	X	✓	✓	X	✓	Domain-specific Clustering and mining based over bio summarizer.
Concept-based abstractive & extractive text mining implemented by Unified Medical Language System [5]	✓	X	X	✓	✓	✓	X	✓	X	X	✓	Minimized clinical literature with limited critical information
Graph-visualization combined with abstractive & extractive summarization in multi-document summarization with dictionary-based approach [6]	✓	✓	X	✓	X	✓	X	✓	✓	X	✓	This approach lack in rule-based automatic text analysis
Combinational analysis of text documents for multi-document summarization [7]	✓	X	✓	✓	X	✓	✓	✓	X	X	✓	Limited analysis of unstructured text document because of some predefined criteria. It is also not capable for every sort of text corpus
Worked over incomplete words and sentences by abstractive & extractive summarization [8]	✓	X	✓	X	X	✓	X	✓	X	X	X	This technique does not assess corpus diversity.
Preserved text documents semantics by distributional-semantic-model with abstractive & extractive summarization [9]	✓	✓	X	✓	X	✓	X	✓	X	X	✓	Mainly focused on text cleansing with no corpus processing reduction
Extractive summarization based semantic framework [10]	✓	X	✓	✓	X	✓	X	X	X	✓	✓	Sentence selection has limited applicability as a new extractive approach.
Academic documents processing with abstractive & extractive summarization [11]	✓	X	✓	X	✓	X	✓	✓	X	X	X	Impact of text grouping and syntactic closeness have no weighted connection
Propose approach	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	Not applicable on very short text like one or two lines,etc.,

This indicator calculation works as below:

$$TF(t,d) = \frac{t_i}{\sum_k t_k} \tag{1}$$

$$IDF(t,d) = log\frac{N}{|\{d \in D : t \in d\}|} \tag{2}$$

where t_i = *Word t in Document d and*
$\sum_k t_k$ = *Total Words in Document d*
N = *total number of documents in the corpus:* $N = |\ D\ |$

D. PAGE RANK ALGORITHM

This algorithm processes a collection of documents linked by hyperlinks such as web pages. It assigns each of them a numerical value to measure its importance relating to other documents [11], [17], [33]. This algorithm applies to any set of objects joined by reciprocal links. The Page Rank (PR) value indicates the significance of a page. This value depends on the number and PR values of all pages that contain links to the current page [27], [43]. If each of pages B, C, and D have links only to page A, then PR (A) will be equal to the sum of PR (B), PR (C), and PR (D) since all links in this simple method will point to A:

PR (A) = PR (B) + PR (C) + PR (D)

E. WORD EMBEDDINGS AND TEXT SUMMARIZATION

Word2Vec model, trained on the body of text, maps words into a small dimension vector space. It keeps the distance between words smaller and closer to the meanings of the words, which occur in close relations [15], [29], [46]. It requires the training of an artificial neural network to predict context from a vector of words. This mapping gives those words which appear in similar contexts to close vectors [28], [30], [43]. This technique gives the cluster of words occurring close in real-time. These words keep getting closer or keep moving away as per the passage of time. This technique is a useful method for topic or subject assessment [17], [33], [48]. In our technique, we have used Python language package Gensim-Word2Vec for text summarization. This extractive text summarization technique detects text lines with words of high mutual occurrence in real space [53]–[55]. Table 3 shows the comparison of sentence score and Gensim-Word2Vec abstractive summaries.

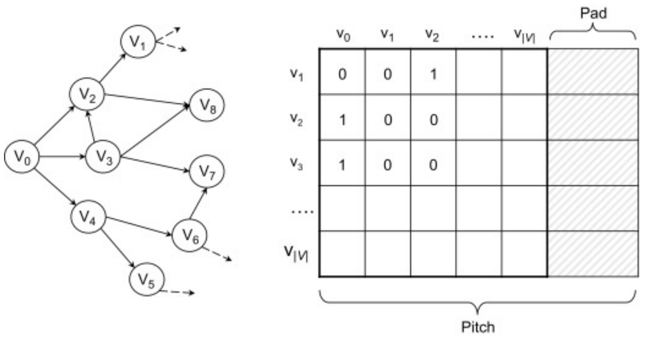
F. K-MEANS CLUSTERING

KMC has an iterative stabilization process for cluster centroids [29], [39], [43]. The main characteristic of a cluster is its centroid. This algorithm stabilizes or at best completely stops the change in the centroid of the cluster [31], [34], [35]. At first, it selects the initial centroids for multiple documents. All document distribution takes place among the clusters. The document falls into only one cluster; the metric of the proximity of the centroid of the document is important. The centroid calculation of the cluster relies on the new set of documents in each cluster. The cycle of calculations repeats if the centroid of the cluster has moved. Otherwise, if the centroid has stabilized, the clustering process completes [32], [33], [46]. The algorithm's theoretical speed is O (n); n is the number of documents in the set. This algorithm has a linear rate and uses the values of the matrix TF-IDF. This method does not need training, and if necessary, it can accumulate information to increase further the accuracy of work using Bayesian estimates of clustering parameters [36], [50].

G. COSINE SIMILARITY

Cosine Similarity (CS) is a good measure to give the similarity between different articles. It recommends those most likely items in which the users are interested. Item similarity recommendation depends on the value of CS [28], [30], [36]. CS has the following formula.

$$Cosine(x, y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \tag{3}$$



	v0	v1	v2	....	v|V|	Pad
v1	0	0	1			
v2	1	0	0			
v3	1	0	0			
....						
v|V|						

FIGURE 1. Adjacency graph and adjacency matrix.

CS is the right choice when attributes are high-dimensional, especially in information retrieval and text analysis [37], [38], [42].

H. ADJACENCY MATRIX

An adjacency matrix represents a graph by a matrix (V × V) with M = [f (i, j)]. The element f (i, j) has the edge (i, j) attributes [20], [39]. The edges without attributes have a graph representation as a matrix to save memory, as shown in Figure 1.

We can also define it as below.

Let G is the graph having, n vertices ordered from v_1 to v_n.

Matrix A (n × n), in which

$a_{ij} = 1$, in a path exists from v_1 to v_j

$a_{ij} = 0$, if path does not exist

The above is adjacency matrix [39], [46].

IV. METHODOLOGY OF DYNAMIC CORPUS CREATION BY MULTI-STAGE METADATA EXTRACTION

This study presents a Dynamic Corpus Creation (DCC) approach for information retrieval with multi-stage metadata selection. It helps to reduce the cost and time for information extraction. It has multiple steps of corpus processing which jointly utilize the principles of both single & multiple document summarization. Section-IV-A to Section-IV-J contains the step by step description of DCC.

A. PREPROCESSING OF TEXT

The unstructured text has specific abnormalities and unnecessary elements. Removal of these redundant elements saves the overhead of cost and time during text processing. Our dataset for experimentation consisted of over thirty-five hundred long paragraphs, having miscellaneous topics [52]. This step resolved the text processing issues of handling of UTF & none-ASCII characters. This issue is step 1 in Figure 2.

B. MAKING MULTIPLE SUMMARIZED PARALLEL CORPUSES

The corpus having long text documents is prone to irrelevant document processing [40], [41]. The proposed approach combined abstractive & extractive techniques. The second approach of crucial line extraction has two other methods.

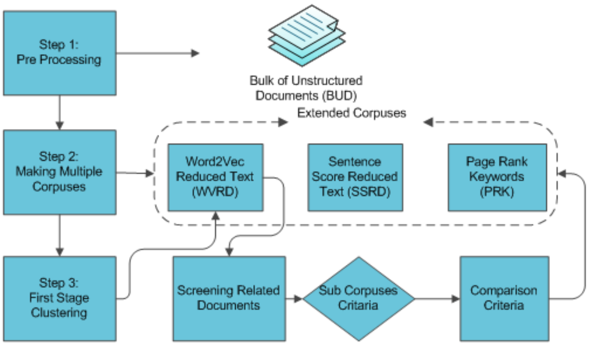


FIGURE 2. Step by step hierarchy of FSC.

In the first approach, the sentence score algorithm extracts the high score text lines [39], [48]. In the second approach, the word embeddings algorithm extracts those lines with a high frequency of mutually occurring words. We also refer to the word embeddings as Word2Vec model [31], [37], [44]. We converted the unstructured text corpus into the three parallel reduced corpuses. i.e., Gensim Based (GB) Word2Vec Summarized Parallel Corpus (SPC), FB-SPC, and Page Rank (PR). We presented all three corpuses as a tuple in Table 3. Each tuple consists of the actual text, GB reduced text, frequency-based reduced text, and page rank keywords, shown in step 2 in Figure 2.

C. FIRST STAGE CLUSTERING

We applied K-Means clustering to divide a single large corpus into multiple sub-corpuses. Contrary to the traditional approaches, our approach used reduced corpus as a substitute for the large corpus [46], [47], [51]. In our experiment, the extended corpuses compared with the actual text by calculating CS values. We applied KMC and produced first stage clusters from the GB-SPC. In Figure 2, we have presented it as step 3.

D. SUB-CORPUS CREATION

In our approach, the Group of documents relating to one cluster is a sub-corpus. Sub-corpus consisted of closely related parts of a larger corpus. We made a comparison of KMC clusters with the entire AEC by calculating CS values. We identified the documents relating to a cluster through these CS comparisons. Documents with high CS towards a specific cluster associated with one sub-corpus. Distinct clusters represented cluster-level corpus summarization. We have mentioned this summarization as step 4 in Figure 3.

E. INTERLINKING MULTIPLE SUB-CORPUSES

We used cluster similarities to join clusters in a path-oriented manner. The adjacency matrix and adjacency graph provide interlinked processing of SC. By using the adjacency matrix, we related each cluster to all other clusters. This matrix has $c_1, c_2, \ldots, c_{24}$ rows and columns, as presented in Table 4. In this path-oriented approach, the processing begins from the most relevant sub-corpus and gradually tends towards the inter-linked sub-corpuses. We have mentioned it as step 5 in Figure 3.

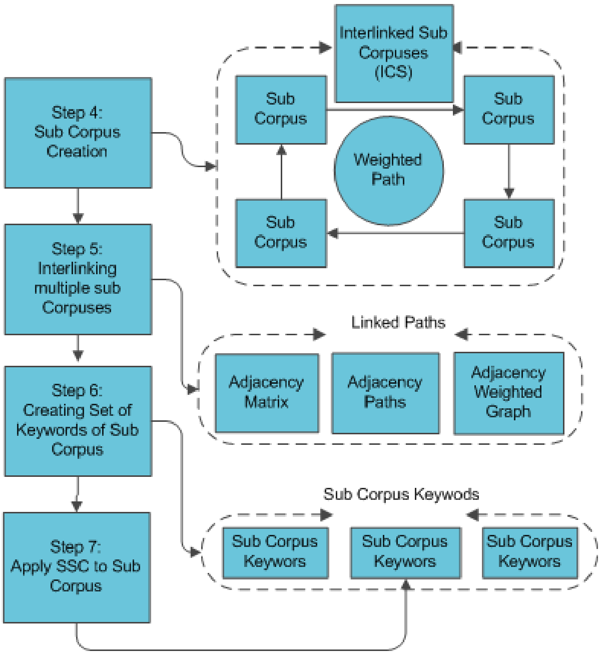


FIGURE 3. Step by step processing of SSC.

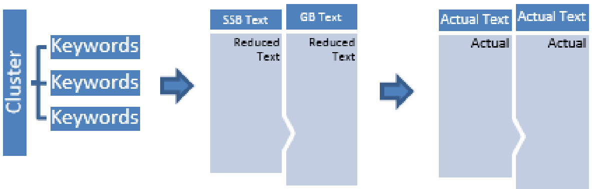


FIGURE 4. Corpus view after first stage KMC.

F. DICTIONARY OF KEYWORDS OF SUB-CORPUS

After the corpus creation, the PRK of sub-corpus collectively combined as a dictionary of keywords. Each document in our experiment contains a specific set of unique keywords, and these keywords form a superset keyword dictionary of a particular sub-corpus. We have processed these dictionaries of sub-corpuses in second stage K-Means clustering. The collection of FSC clusters and sub-corpus dictionaries jointly form enriched metadata to identify the required text. We have mentioned it as step 6 in Figure 3. This metadata includes the first layer of query processing in the proposed approach, as presented in Figure 4.

G. APPLY SECOND STAGE CLUSTERING (SSC) SC

In our approach, the second stage KMC applied over grouped PRK of sub-corpuses. It further identifies relevant portions of SC and related it to a specific cluster. KMC produced unique sub-corpus clusters. CS calculated for each cluster

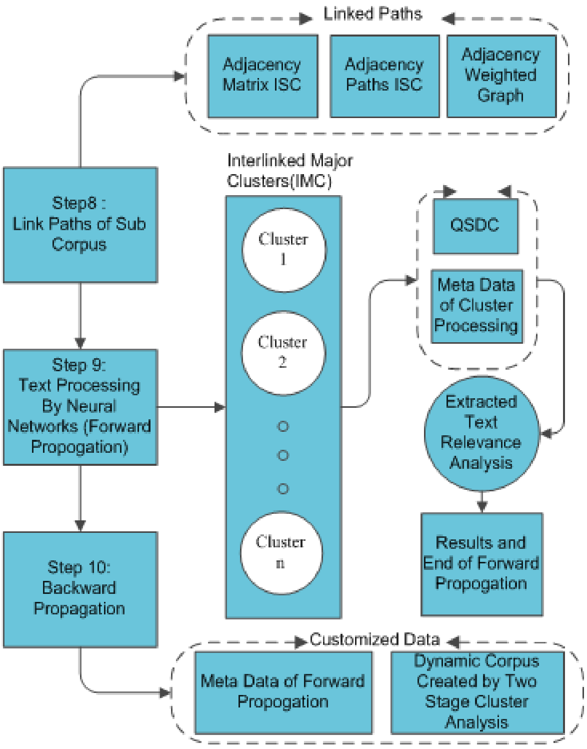

Linked Paths
Adjacency Matrix ISC
Adjacency Paths ISC
Adjacency Weighted Graph
Interlinked Major Clusters(IMC)
Step8 : Link Paths of Sub Corpus
Cluster 1
Cluster 2
Cluster n
QSDC
Meta Data of Cluster Processing
Step 9: Text Processing By Neural Networks (Forward Propogation)
Extracted Text Relevance Analysis
Results and End of Forward Propogation
Step 10: Backward Propagation
Customized Data
Meta Data of Forward Propogation
Dynamic Corpus Created by Two Stage Cluster Analysis

FIGURE 5. Text processing over linked nodes of the corpus.

to determine the relevant part of the sub-corpus. We have mentioned it as step 7 in Figure 3.

H. CREATING LINKED PATHS OF SUB-CORPUS

After the second stage K-Means clustering, these clusters inter-related to each other in the same manner as followed to interlink the first-stage clusters. An adjacency matrix for one of sub-corpuses has $c_1, c_2, \ldots, c_6$ rows and columns as presented in Table 5. All the remaining sub-corpuses have similar matrices.

I. TEXT PROCESSING BY FORWARDING PROPAGATION

We created systematic processing layers for hierarchical text processing. The first two layers consist of the clusters of FSC and SSC, joined in a path-oriented manner. The third layer contains keywords of the documents. The fourth layer has the GB summarized text, as presented in Figure 6. We generated the Query Specific Dynamic Corpus (QSDC) by processing the queries through these layers. At first, the query matches the clusters' set in the first layer to identify the relevant clusters. Secondly, the query matches with the clusters of the sub-corpus. Thirdly text query matches the keywords of selected potential paragraphs. Next, the query matches with relevant summarized GB extracted text. After the complete processing, the technique extracted relevant actual text at the last step. We have mentioned it as step 9 in Figure 5. The proposed path-oriented process reduces corpus comparisons and provides three primary processing outcomes, i.e., QSDC, strongly related corpus segments, and query processing metadata.

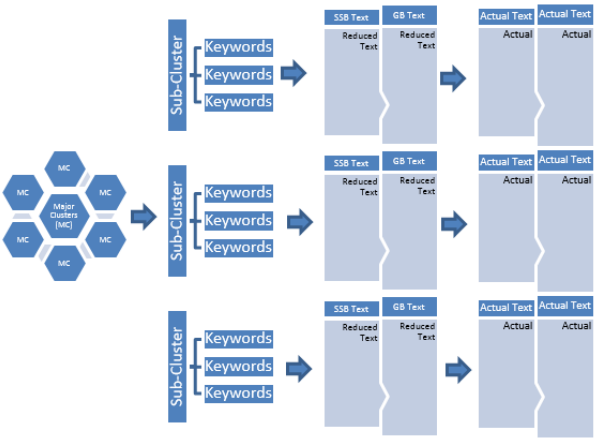
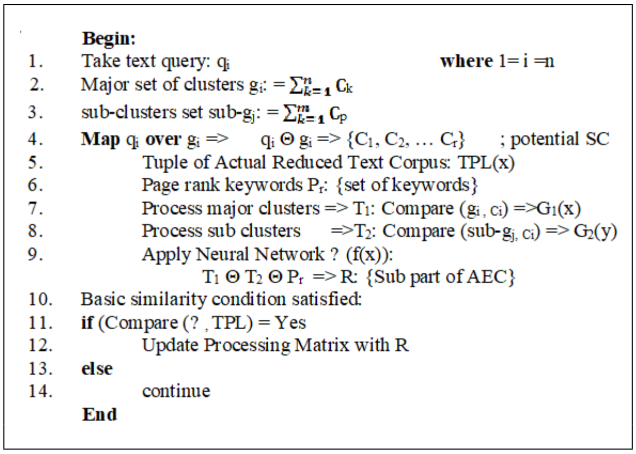


FIGURE 6. Corpus view after second-stage KMC.

TABLE 2. Algorithm 1. Text and query processing.

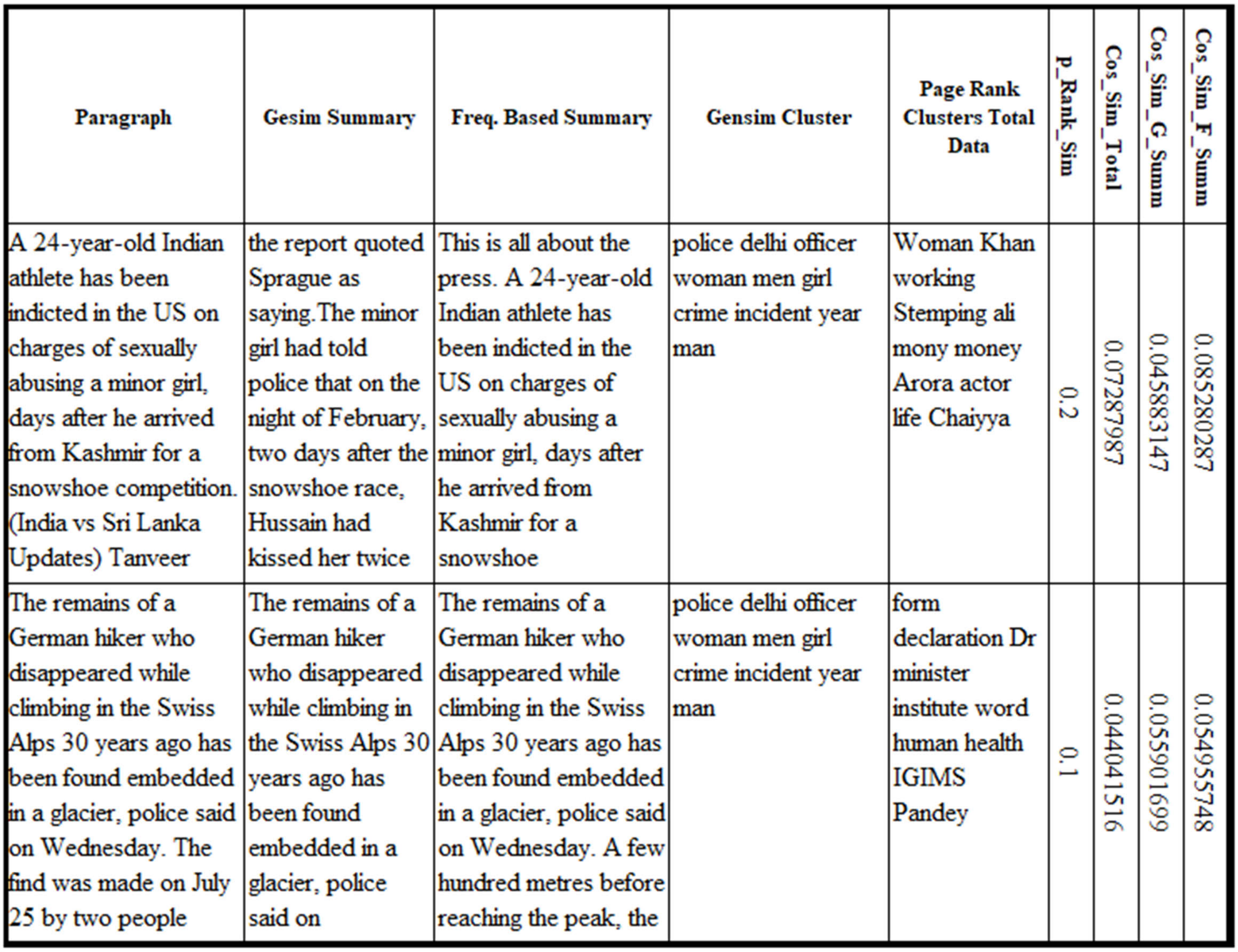


```
Begin:
Take text query: $q_i$                                  where 1= i =n
Major set of clusters $g_i$: $= \sum_{k=1}^{n} C_k$
sub-clusters set sub-$g_j$: $= \sum_{k=1}^{m} C_p$
Map $q_i$ over $g_i$ =>     $q_i \Theta g_i$ => {$C_1$, $C_2$, … $C_r$}     ; potential SC
        Tuple of Actual Reduced Text Corpus: TPL(x)
        Page rank keywords $P_r$: {set of keywords}
        Process major clusters => $T_1$: Compare ($g_i$ , $c_i$) =>$G_1$(x)
        Process sub clusters    =>$T_2$: Compare (sub-$g_j$, $c_i$) => $G_2$(y)
        Apply Neural Network ? (f(x)):
                $T_1 \Theta T_2 \Theta P_r$ => R: {Sub part of AEC}
Basic similarity condition satisfied:
if (Compare (? , TPL) = Yes
        Update Processing Matrix with R
else
        continue
End
```

J. TEXT PROCESSING BACKWARD PROPAGATION

In this step, we performed text processing over the marked set of potential paragraphs. The extracted text relied on the keyword's selection, text extraction, and KMC. The current technique generates the text query results either by permanent text extraction for further processing or uses text temporarily/ad-hoc textual-results. The query processing starts from strongly to weakly related nodes. Subsequent query processing relies on the metadata by applying the principles of neural networks. More relevant text portions have default priorities in text processing. Text queries kept in processing history with identified related metadata. The queries related to metadata facilitated to classify similar questions based on clusters & text similarities. We have mentioned it as step 10 in Figure 5.

TABLE 3. Actual and extended summarized parallel corpuses with cosine similarity values.



Paragraph	Gesim Summary	Freq. Based Summary	Gensim Cluster	Page Rank Clusters Total Data	P_Rank_Sim	Cos_Sim_Total	Cos_Sim_G_Summ	Cos_Sim_F_Summ
A 24-year-old Indian athlete has been indicted in the US on charges of sexually abusing a minor girl, days after he arrived from Kashmir for a snowshoe competition. (India vs Sri Lanka Updates) Tanveer	the report quoted Sprague as saying.The minor girl had told police that on the night of February, two days after the snowshoe race, Hussain had kissed her twice	This is all about the press. A 24-year-old Indian athlete has been indicted in the US on charges of sexually abusing a minor girl, days after he arrived from Kashmir for a snowshoe	police delhi officer woman men girl crime incident year man	Woman Khan working Stemping ali mony money Arora actor life Chaiyya	0.2	0.07287987	0.045883147	0.085280287
The remains of a German hiker who disappeared while climbing in the Swiss Alps 30 years ago has been found embedded in a glacier, police said on Wednesday. The find was made on July 25 by two people	The remains of a German hiker who disappeared while climbing in the Swiss Alps 30 years ago has been found embedded in a glacier, police said on	The remains of a German hiker who disappeared while climbing in the Swiss Alps 30 years ago has been found embedded in a glacier, police said on Wednesday. A few hundred metres before reaching the peak, the	police delhi officer woman men girl crime incident year man	form declaration Dr minister institute word human health IGIMS Pandey	0.1	0.044041516	0.055901699	0.054955748

K. DESCRIPTION OF ALGORITHM

This algorithm improves query processing. It executes text processing in a staged manner. At first stage the most related major (C_k) and sub-corpus (C_p) cluster generate reduced corpus. For example, if the 20% corpus relates to the given query, then ultimately 80% corpus processing becomes reduced. At the next stage, the query comparison with abstractive summarization identifies more related text. For example, if the related text is 13%, then 87% corpus processing becomes reduced. At the last stage, query & extracted summarized text comparison provides most related actual text documents. This algorithm utilizes interlinked metadata of a complete corpus and relies on the actual text as well.

V. RESULTS AND DISCUSSION

Our proposed DCC approach processed a News dataset [52]. This dataset has more than thirty-five hundred long paragraphs with miscellaneous topics. In our experiment, we used long paragraphs of an average of eight hundred words from this dataset. We have applied Python language package NLTK to perform tokenization. Unnecessary token removal gave a useful set of tokens. We applied procedures like speech tagging, stemming & lemmatization, and identified the various types of text-tokens.

A. SUMMARIZED PARALLEL CORPUSES

There are two approaches to assess the theme of underlying text documents without processing the whole-text, i.e., 1) Creating the dictionary of keywords, 2) Extracting the crucial lines of the unstructured text. We generated the reduced corpuses by applying these text reduction techniques as described in Section-IV-B, and this step reduced the large corpus by extracting important words & lines from it as presented in Table 3. This step makes text query rely on reduced actual-text & page rank keywords, and Figure 7 shows the text reduction statistics obtained by the three text extraction techniques. The corpus processing becomes almost seventy percent reduced by extractive summarization, as shown

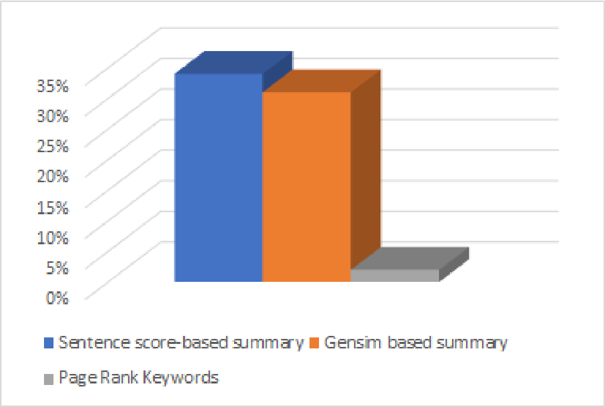
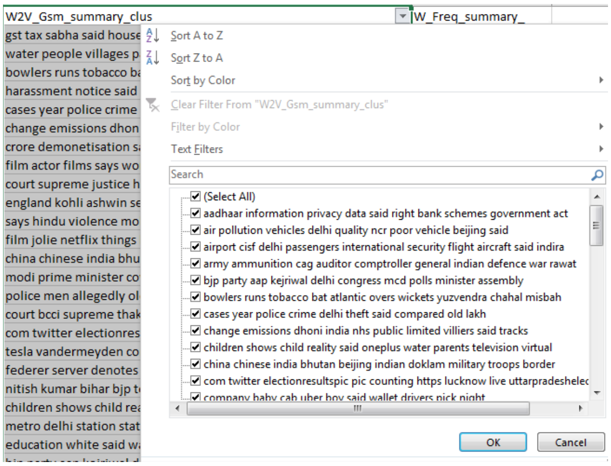


FIGURE 7. Comparison of text reduction by summarization techniques.



W2V_Gsm_summary_clus | W_Freq_summary_

gst tax sabha said house
water people villages p
bowlers runs tobacco ba
harassment notice said
cases year police crime
change emissions dhon
crore demonetisation sa
film actor films says wo
court supreme justice h
england kohli ashwin se
says hindu violence mo
film jolie netflix things
china chinese india bhu
modi prime minister co
police men allegedly ol
court bcci supreme thak
com twitter electionres
tesla vandermeyden co
federer server denotes
nitish kumar bihar bjp t
children shows child rea
metro delhi station stat
education white said wa

☑ (Select All)
☑ aadhaar information privacy data said right bank schemes government act
☑ air pollution vehicles delhi quality ncr poor vehicle beijing said
☑ airport cisf delhi passengers international security flight aircraft said indira
☑ army ammunition cag auditor comptroller general indian defence war rawat
☑ bjp party aap kejriwal delhi congress mcd polls minister assembly
☑ bowlers runs tobacco bat atlantic overs wickets yuzvendra chahal misbah
☑ cases year police crime delhi theft said compared old lakh
☑ change emissions dhoni india nhs public limited villiers said tracks
☑ children shows child reality said oneplus water parents television virtual
☑ china chinese india bhutan beijing indian doklam military troops border
☑ com twitter electionresultspic pic counting https lucknow live uttarpradeshelec
☑ company baby cab uber boy said wallet drivers pick night

FIGURE 8. Main corpus with First Stage Clustering.

in Figure 7. We have the text reduction statistics in the form of a bar-chart in Figure 7. As per our experimentation, GB-SPC & FB-SPC consists of almost one-third of the actual text. The extracted-text becomes approximately thirty percent of the total text. PRK-SPC gave the top keywords from each paragraph. PR keywords of the given paragraph obtained from the pre-processed text cleansed text.

B. SELECTION OF REDUCED CORPUS TO APPLY FSC

We identified similar documents by applying the K-Means clustering process. KMC gives optimal results in many situations [17], [34], [46]. We compared the GB reduced text with {Actual-Text, SS-reduced text, PRK} by calculating the CS values. Similarly, the SS-reduced text compared with the {Actual-Text, GB-reduced text, PRK}. Figure 9 shows the comparison results obtained from GB & SS comparisons. GB reduced corpus has better text similarity to the actual text. Based on these results, we selected GB-reduced corpus as a substitute for the actual-corpus. We compared KMC from GB-corpus with each tuple, as illustrated in Table 3. Figure 4 is the representation of corpus view after applying the first stage K-Means clustering. Table 3 represents the formation of the underlying dataset.

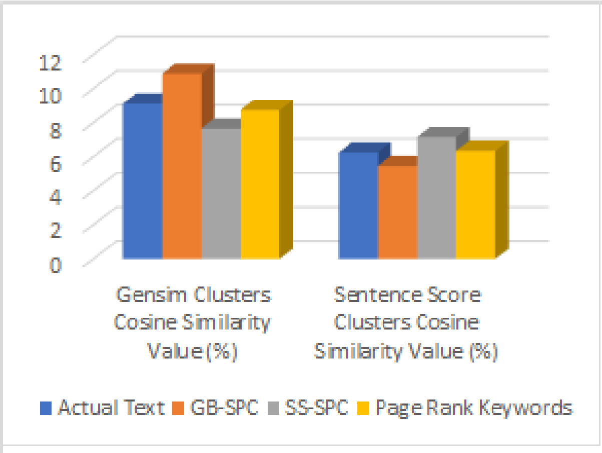
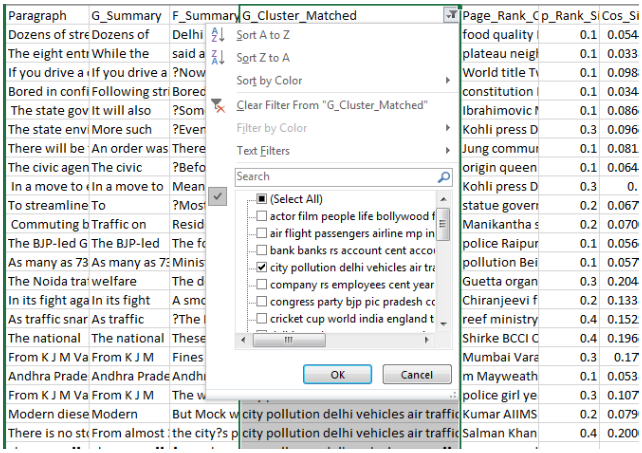


FIGURE 9. Comparison of extractive summarization approaches.



Paragraph	G_Summary	F_Summary	G_Cluster_Matched	Page_Rank_C	p_Rank_Si	Cos_Si
Dozens of stre	Dozens of	Delhi		food quality	0.1	0.054
The eight entr	While the	said a		plateau neigl	0.1	0.033
If you drive a	If you drive a	?Now		World title T	0.1	0.098
Bored in confi	Following stri	Bored		constitution	0.1	0.034
The state gov	It will also	?Som		Ibrahimovic	0.1	0.086
The state envi	More such	?Even		Kohli press D	0.3	0.096
There will be	An order was	There		Jung commu	0.1	0.081
The civic agen	The civic	?Befo		origin queen	0.1	0.064
In a move to	In a move to	Mean		Kohli press D	0.3	0.
To streamline	To	?Most		statue gover	0.2	0.067
Commuting b	Traffic on	Resid		Manikantha s	0.2	0.070
The BJP-led G	The BJP-led	The fc		police Raipur	0.1	0.056
As many as 73	As many as 73	Minis		pollution Bei	0.1	0.057
The Noida tra	welfare	The d		Guetta organ	0.3	0.204
In its fight aga	In its fight	A smc		Chiranjeevi f	0.2	0.133
As traffic snar	As traffic	?The		reef ministry	0.4	0.152
The national	The national	These		Shirke BCCI C	0.4	0.196
From K J M Va	From K J M	Fines		Mumbai Vara	0.3	0.17
Andhra Prade	Andhra Prade	Andhi		m Mayweath	0.1	0.053
From K J M Va	From K J M	The w		police girl ye	0.3	0.107
Modern diese	Modern	But Mock w	city pollution delhi vehicles air traffic	Kumar AIIMS	0.2	0.079
There is no st	From almost	the city?s p	city pollution delhi vehicles air traffic	Salman Khan	0.4	0.200

■ (Select All)
□ actor film people life bollywood f
□ air flight passengers airline mp in
□ bank banks rs account cent acco
☑ city pollution delhi vehicles air tra
□ company rs employees cent year
□ congress party bjp pic pradesh cc
□ cricket cup world india england t

FIGURE 10. Sub-corpus view with K-Means clustering.

C. CREATING SUB-CORPUSES BY USING KMC

We applied FSC to the GB reduced corpus to generate the set of KMC clusters. Each cluster matched with actual and extended corpuses. The sub-corpus creation followed the steps as mentioned in Section-IV-D. Figure 8 represents the distinct first stage clusters. We used twenty-four distinct clusters to identify the related groups of documents. We calculated the CS values between these clusters and placed them in the next AEC columns, as presented in Table 3. Based on text similarity, we grouped a document under a designated sub-corpus. This step created a cluster-based text grouping of a large corpus into manageable sub-corpuses. Figure 10 presents the sub-corpus view, and each cluster in this figure has a specific set of documents. These sub-corpuses form a subset of similar documents in one large corpus.

D. CONNECTING SUB-CORPUSES BY DENSE GRAPH

We interrelated the clusters in a path-oriented manner by following the steps mentioned in Section-IV-E. We developed a dense adjacency matrix, as presented in Table 4. Figure 11 represents this matrix by an adjacency graph to show the

TABLE 4. Adjacency matrix of FSC.

Clusters	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}	c_{11}	c_{12}	c_{13}	c_{14}	c_{15}	c_{16}	c_{17}	c_{18}	c_{19}	c_{20}	c_{21}	c_{22}	c_{23}	c_{24}
c_1	1	0	0	0	0	0	0	0	0	0	0.1	0	0	0	0	0.1	0.1	0	0.1	0	0	0.2	0	0
c_2	0	1	0.3	0.1	0.1	0	0	0	0.1	0	0	0	0.2	0	0	0	0.2	0	0.1	0.1	0	0	0	0.1
c_3	0	0.3	1	0.2	0.1	0	0.2	0	0.1	0	0	0	0.1	0	0	0	0.2	0	0.1	0	0	0	0.1	0.1
c_4	0	0.1	0.2	1	0	0	0	0	0.1	0	0.1	0	0	0.1	0	0	0.2	0	0	0	0	0	0	0
c_5	0	0.1	0.1	0	1	0	0	0	0	0	0	0	0	0	0	0	0.1	0	0.2	0	0	0	0	0.2
c_6	0	0	0	0	0	1	0.1	0	0	0.2	0.1	0.2	0	0.2	0.1	0	0	0	0.1	0	0.1	0.1	0	0
c_7	0	0	0.2	0	0	0.1	1	0.1	0	0	0.1	0	0	0	0.3	0	0	0	0.1	0	0	0	0.1	0
c_8	0	0	0	0	0	0	0.1	1	0	0	0	0	0	0	0.1	0	0	0	0	0	0	0	0	0
c_9	0	0.1	0.1	0.1	0	0	0	0	1	0	0	0	0.1	0	0.1	0	0.2	0.1	0	0.1	0	0	0.2	0
c_{10}	0	0	0	0	0	0.2	0	0	0	1	0.1	0.2	0	0.2	0	0	0	0	0	0	0	0	0	0
c_{11}	0.1	0	0	0.1	0	0.1	0.1	0	0	0.1	1	0.2	0	0.2	0.1	0	0.1	0	0	0	0	0	0	0
c_{12}	0	0	0	0	0	0.2	0	0	0	0.2	0.2	1	0	0.3	0	0	0	0	0	0	0.1	0	0	0
c_{13}	0	0.2	0.1	0	0	0	0	0	0.1	0	0	0	1	0	0	0	0.1	0.1	0	0.3	0	0	0	0
c_{14}	0	0	0	0.1	0	0.2	0	0	0	0.2	0.2	0.3	0	1	0	0	0.1	0	0	0	0.1	0	0	0
c_{15}	0	0	0	0	0	0.1	0.3	0.1	0.1	0	0.1	0	0	0	1	0	0	0.1	0.1	0.1	0	0	0.1	0
c_{16}	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0.1	0	0.1	0	0	0.1	0.1	0
c_{17}	0.1	0.2	0.2	0.2	0.1	0	0	0	0.2	0	0.1	0	0.1	0.1	0	0.1	1	0.1	0.2	0	0	0.1	0	0.1
c_{18}	0	0	0	0	0	0	0	0	0.1	0	0	0	0.1	0	0.1	0	0.1	1	0	0.4	0	0	0.1	0
c_{19}	0.1	0.1	0.1	0	0.2	0.1	0.1	0	0	0	0	0	0	0	0.1	0.1	0.2	0	1	0	0	0.1	0	0.2
c_{20}	0	0.1	0	0	0	0	0	0	0.1	0	0	0	0.3	0	0.1	0	0	0.4	0	1	0	0	0.1	0
c_{21}	0	0	0	0	0	0.1	0	0	0	0	0	0.1	0	0.1	0	0	0	0	0	0	1	0	0	0
c_{22}	0.2	0	0	0	0	0.1	0	0	0	0	0	0	0	0	0	0.1	0.1	0	0.1	0	0	1	0	0
c_{23}	0	0	0.1	0	0	0	0.1	0	0.2	0	0	0	0	0	0.1	0.1	0	0.1	0	0.1	0	0	1	0
c_{24}	0	0.1	0.1	0	0.2	0	0	0	0	0	0	0	0	0	0	0	0.1	0	0.2	0	0	0	0	1

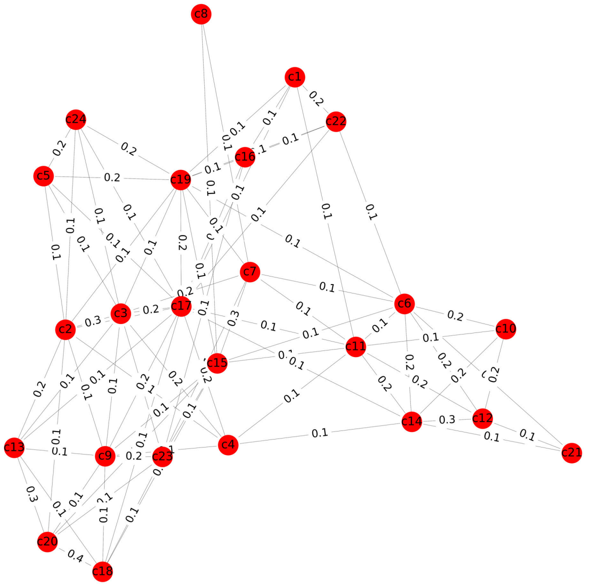


FIGURE 11. Weighted dense graph of connected FSC clusters.

connected nodes of sub-corpuses by weighted paths. This graph provides path selection during corpus processing. Each row in Table 4 depicts a cluster's relation to another cluster, and vice versa. AM of each SC presents a complete mapping of clusters. We generated adjacency lists by processing the AM & AG. These lists relate every sub-corpus to its related set of sub-corpuses.

E. SUB-CORPUS KEYWORDS DICTIONARY, SSC, AND INTERLINKING SUB-CORPUSES

We created the keyword dictionaries of each sub-corpus, as discussed in Section-IV-F. We applied the second stage KMC over these dictionaries to perform further sub-grouping of the documents. In this step, unique words of multiple documents jointly form the collective dictionary of each sub-corpus. We refer to it as the multi-document abstractive summarization. It provides the collective text relevance, specific to a group of documents [6], [14], [29]. For each cluster, we obtained CS values for a tuple of actual and extended corpuses. We linked the clusters of sub-corpuses by a sparse adjacency matrix. Table 5 presents the adjacency matrix of one of the sub-corpuses. In Table 5, each row shows the relation of a cluster to another and vice versa. AM of each sub-corpus presents a complete mapping of clusters. We have mentioned it as step 8 in Figure 5. By utilizing this matrix, path-oriented information processing follows weighted edges between connected nodes, as shown in Figure 12. This graph shows the relationship in the form of weights between the inter-related clusters of the sub-corpus. Figure 21 presents the corpus view after complete processing.

TABLE 5. Adjacency matrix with CS values for sub-corpus.

Clusters	c_1	c_2	c_3	c_4	c_5	c_6
c_1	1	0.3	0.1	0.1	0.1	0.1
c_2	0.3	1	0.1	0.2	0	0.1
c_3	0.1	0.1	1	0.1	0	0.1
c_4	0.1	0.2	0.1	1	0	0.1
c_5	0.1	0	0	0	1	0
c_6	0.1	0.1	0.1	0.1	0	1

F. EVALUATION AND COMPARISONS OF IMPLEMENTED TECHNIQUE

Our approach has many advantages over traditional metadata extraction and processing techniques. We have carried out different text processing comparisons to assess the proposed

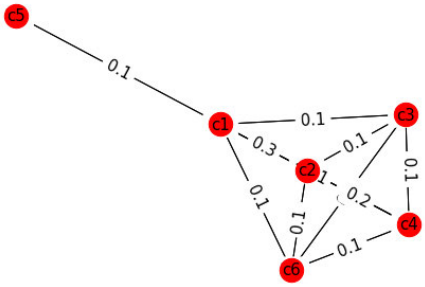
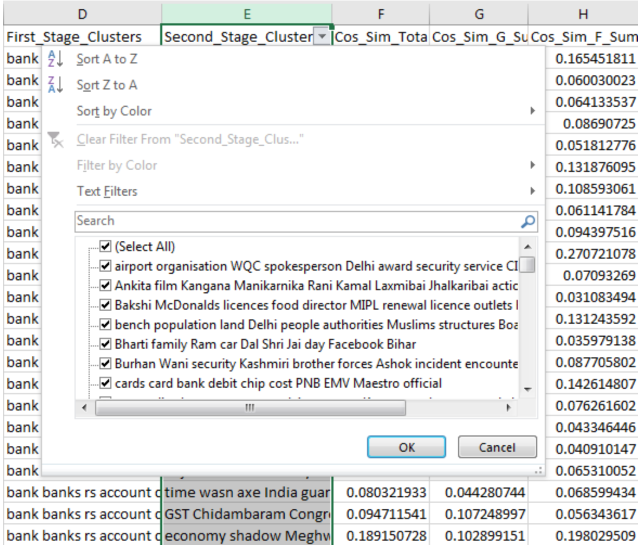


FIGURE 12. Weighted sparse graph of for a sub-corpus after SSC.



D	E	F	G	H
First_Stage_Clusters	Second_Stage_Cluster	Cos_Sim_Tota	Cos_Sim_G_Su	Cos_Sim_F_Sum
bank				0.165451811
bank				0.060030023
bank				0.064133537
bank				0.08690725
bank				0.051812776
bank				0.131876095
bank				0.108593061
bank				0.061141784
bank				0.094397516
bank				0.270721078
bank				0.07093269
bank				0.031083494
bank				0.131243592
bank				0.035979138
bank				0.087705802
bank				0.142614807
bank				0.076261602
bank				0.043346446
bank				0.040910147
bank				0.065310052
bank banks rs account c	time wasn axe India guar	0.080321933	0.044280744	0.068599434
bank banks rs account c	GST Chidambaram Congr	0.094711541	0.107248997	0.056343617
bank banks rs account c	economy shadow Meghw	0.189150728	0.102899151	0.198029509

☑ (Select All)
☑ airport organisation WQC spokesperson Delhi award security service CI
☑ Ankita film Kangana Manikarnika Rani Kamal Laxmibai Jhalkaribai actic
☑ Bakshi McDonalds licences food director MIPL renewal licence outlets I
☑ bench population land Delhi people authorities Muslims structures Boa
☑ Bharti family Ram car Dal Shri Jai day Facebook Bihar
☑ Burhan Wani security Kashmiri brother forces Ashok incident encounte
☑ cards card bank debit chip cost PNB EMV Maestro official

FIGURE 13. Corpus view after Second Stage Clustering.

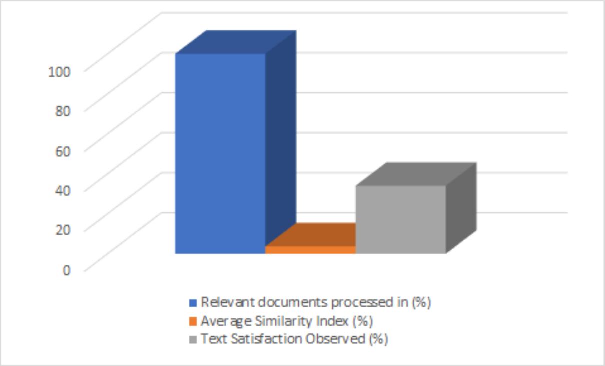


FIGURE 14. Text processing by simple keywords.

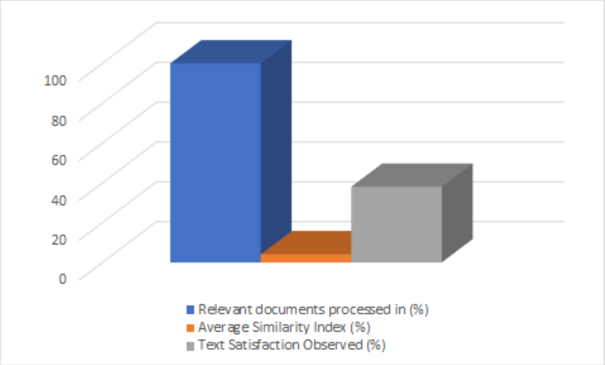


FIGURE 15. Text processing by Page Rank keywords.

method's efficiency by executing text queries. We processed these text queries by six other techniques—this analysis performed in the same way as discussed in Section-IV-I and IV-J. This section presents the results of the comparisons of different techniques. In our experimentation, the path-oriented text processing gave better results than other commonly used text processing techniques.

Abstractive Summarization: Every webpage has specific keywords to get the attention of audiences. In our experimentation, this technique provides less than 7% relevant text, and it involved 100% text document processing, as presented in Figure 14. This technique uses human analysis and sometimes gives misleading results because of human errors, and it has widespread implementation as the simplest form of metadata. TF-IDF based Keywords selection gives text-based keywords. These keywords have better relation to text as compared to the simple keyword's selection techniques. These two forms of abstractive summarization are similar. Both approaches follow the dictionary approach and both approaches depend on selecting important terms as keywords. In both techniques selection of misleading words causes wastage of time and resources. In our experiment TF-IDF, based PRK, utilized 100% text documents. The text similarity remained about 7%, as presented in Figure 15. However, it provided better results than the simple keyword selection process, as presented in Figure 20.

Clusters based Text Summarization: In our experiment, First Stage & Second Stage Clustering generated cluster level corpus summarization. These clusters facilitate to determine the corresponding portions of a corpus. We carefully matched the clusters with all forms of existing & extended-corpuses and selected those clusters that gave adequate similarity with TFIDF-PRK, word2vec-reduced-text & sentence-score-text. This technique generates the topic assessment in text analysis studies [4], [17], [21], [25]. These clusters serve as nodes in our adjacency graph, and various clusters form a weighted path for corpus traversal and processing. This technique generates efficient path-oriented metadata. The first stage clusters gave about 10% text similarity and processed approximately 50% documents, as presented in Figure 16. In the second stage, K-Means clustering over the sub-corpus dictionary of words formed many interconnected connected nodes of a text corpus. These nodes provide intense corpus traversal in which dense graphs serve as a central entity. Our experimentation shows that the cluster-based text query processing facilitated better text analysis with more than 10% text similarity and better-observed result satisfaction. This technique processed less than 40% of text documents, as presented in Figure 17. This technique has better text processing

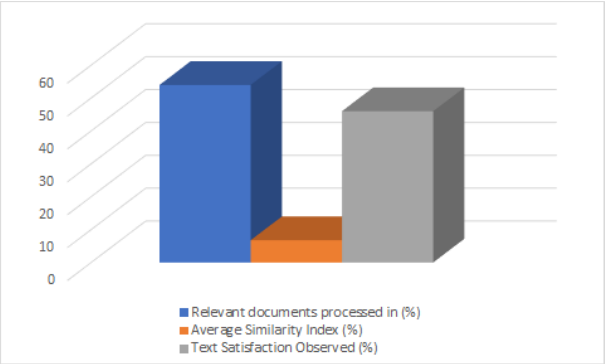


FIGURE 16. Text processing by first stage KMC.

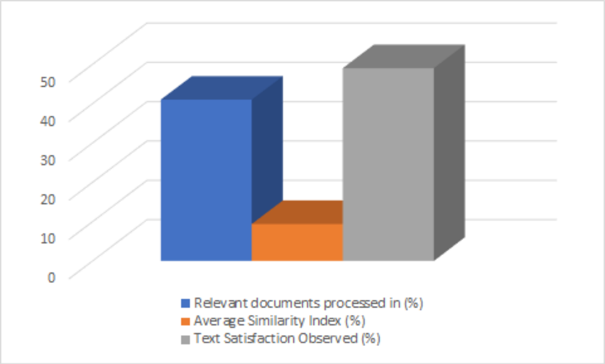


FIGURE 17. Text processing by Page Rank Keywords & first stage KMC.

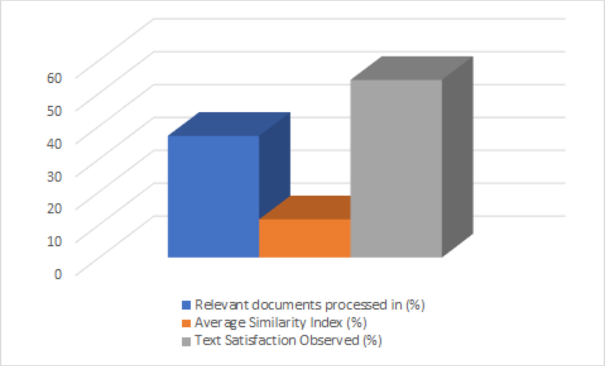


FIGURE 18. Text processing first & second stage KMC.

than simple abstractive text summarization, as presented in Figure 20.

Abstractive & Cluster-based Summarization: We joined query processing with cluster & keywords selection. In this way, one cluster becomes connected with the sub-corpus keyword's dictionary & individual document's keywords. This form of text processing is an improved form of abstractive text summarization based on a better set of related text documents. This text processing gives better text similarity as compared to just relying on abstractive text summarization. It gives about 10% text similarity with processing approximately 40% actual text documents, as presented in Figure 18.

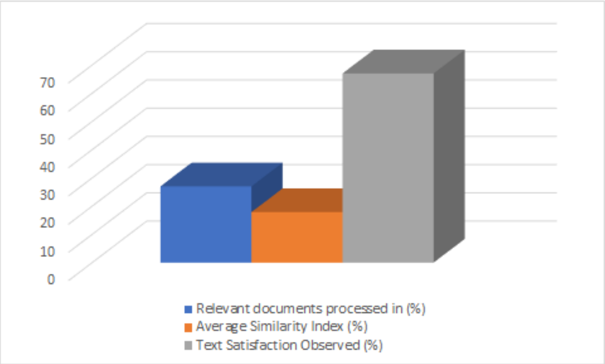


FIGURE 19. Text processing by first & second stage KMC with abstractive & extractive summarization.

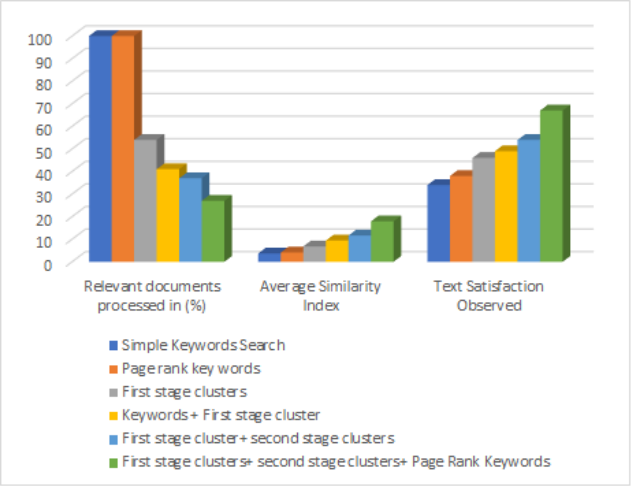
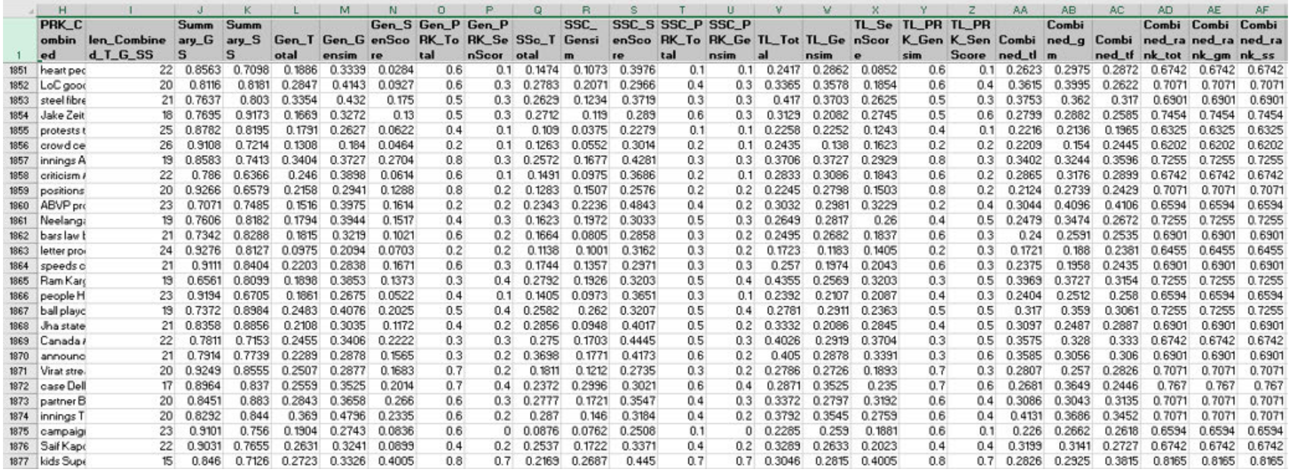


FIGURE 20. Text processing by first & second stage KMC with abstractive & extractive summarization.

However, multistage Clustering has better text processing results than Clustering and dictionary-based abstractive-summarization mechanism, as presented in Figure 20.

Path-oriented approach with FSC & SSC to select the sub-corpuses: The sub-corpus cluster's processing with metadata and summarized text causes the extraction of most related text documents. This technique provides path-based metadata processing and utilizes the advantages of all the text extraction and summarization techniques. This technique offers an efficient starting point for corpus processing. In this way, text processing resources become pre-prioritized in a systematic fashion. The significant unmanageable corpus becomes small and less diverse. In this way, the query-based dynamic corpus becomes available. Each query has its own set of the cluster, dictionary of keywords, and extracted text. This reduced form of metadata utilized actual extracted-text and facilitated the most query-relevant real documents presented in Figure 19. This technique uses about 33% text document and provides approximately 20% text-similarity in actual-documents. This technique process 66% fewer documents and give four times better results. As compared to simple



	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	V	X	Y	Z	AA	AB	AC	AD	AE	AF
1	PRK_Combined	len_Combined_T_G_SS	Summary_GS	Summary_SS	Gen_Total	Gen_Gensim	Gen_SenScore	Gen_PRK_Total	Gen_PRK_SenScor	SSc_Total	SSC_Gensim	SSC_SenScore	SSC_PRK_Total	SSC_PRK_Gensim	TL_Total	TL_Gensim	TL_SenScore	TL_PRK_Gensim	TL_PRK_SenScore	Combined_tl	Combined_gm	Combined_tf	Combined_rank_tot	Combined_rank_gm	Combined_rank_ss
1851	heart pec	22	0.8563	0.7098	0.1886	0.3339	0.0284	0.6	0.1	0.1474	0.1073	0.3976	0.1	0.1	0.2417	0.2862	0.0852	0.6	0.1	0.2623	0.2975	0.2872	0.6742	0.6742	0.6742
1852	LoC gooc	20	0.8116	0.8181	0.2847	0.4143	0.0927	0.6	0.3	0.2783	0.2071	0.2966	0.4	0.3	0.3365	0.3578	0.1854	0.6	0.4	0.3615	0.3995	0.2622	0.7071	0.7071	0.7071
1853	steel fibre	21	0.7637	0.803	0.3354	0.432	0.175	0.5	0.3	0.2629	0.1234	0.3719	0.3	0.3	0.417	0.3703	0.2625	0.5	0.3	0.3753	0.362	0.317	0.6901	0.6901	0.6901
1854	Jake Zeit	18	0.7695	0.9173	0.1669	0.3272	0.13	0.5	0.3	0.2712	0.119	0.289	0.6	0.3	0.3129	0.2082	0.2745	0.5	0.6	0.2799	0.2882	0.2585	0.7454	0.7454	0.7454
1855	protests t	25	0.8782	0.8195	0.1791	0.2627	0.0622	0.4	0.1	0.109	0.0375	0.2279	0.1	0.1	0.2258	0.2252	0.1243	0.4	0.1	0.2216	0.2136	0.1965	0.6325	0.6325	0.6325
1856	crowd ce	26	0.9108	0.7214	0.1308	0.184	0.0464	0.2	0.1	0.1263	0.0552	0.3014	0.2	0.1	0.2435	0.138	0.1623	0.2	0.2	0.2209	0.154	0.2445	0.6202	0.6202	0.6202
1857	innings A	19	0.8583	0.7413	0.3404	0.3727	0.2704	0.8	0.3	0.2572	0.1677	0.4281	0.3	0.3	0.3706	0.3727	0.2929	0.8	0.3	0.3402	0.3244	0.3596	0.7255	0.7255	0.7255
1858	criticism /	22	0.786	0.6366	0.246	0.3898	0.0614	0.6	0.1	0.1491	0.0975	0.3686	0.2	0.1	0.2833	0.3086	0.1843	0.6	0.2	0.2865	0.3176	0.2899	0.6742	0.6742	0.6742
1859	positions	20	0.9266	0.6579	0.2158	0.2941	0.1288	0.8	0.2	0.1283	0.1507	0.2576	0.2	0.2	0.2245	0.2798	0.1503	0.8	0.2	0.2124	0.2739	0.2429	0.7071	0.7071	0.7071
1860	ABVP prc	23	0.7071	0.7485	0.1516	0.3975	0.1614	0.2	0.2	0.2343	0.2236	0.4843	0.4	0.2	0.3032	0.2981	0.3229	0.2	0.4	0.3044	0.4096	0.4106	0.6594	0.6594	0.6594
1861	Neelang:	19	0.7606	0.8182	0.1794	0.3944	0.1517	0.4	0.3	0.1623	0.1972	0.3033	0.5	0.3	0.2649	0.2817	0.26	0.4	0.5	0.2479	0.3474	0.2672	0.7255	0.7255	0.7255
1862	bars law t	21	0.7342	0.8288	0.1815	0.3219	0.1021	0.6	0.2	0.1664	0.0805	0.2858	0.3	0.2	0.2495	0.2682	0.1837	0.6	0.3	0.24	0.2591	0.2535	0.6901	0.6901	0.6901
1863	letter pro	24	0.9276	0.8127	0.0975	0.2094	0.0703	0.2	0.2	0.1138	0.1001	0.3162	0.3	0.2	0.1723	0.1183	0.1405	0.2	0.3	0.1721	0.188	0.2381	0.6455	0.6455	0.6455
1864	speeds c	21	0.9111	0.8404	0.2203	0.2838	0.1671	0.6	0.3	0.1744	0.1357	0.2971	0.3	0.3	0.257	0.1974	0.2043	0.6	0.3	0.2375	0.1958	0.2435	0.6901	0.6901	0.6901
1865	Ram Kar	19	0.6561	0.8099	0.1898	0.3853	0.1373	0.3	0.4	0.2792	0.1926	0.3203	0.5	0.4	0.4355	0.2569	0.3203	0.3	0.5	0.3969	0.3727	0.3154	0.7255	0.7255	0.7255
1866	people H	23	0.9194	0.6705	0.1861	0.2675	0.0522	0.4	0.1	0.1405	0.0973	0.3651	0.3	0.1	0.2392	0.2107	0.2087	0.4	0.3	0.2404	0.2512	0.258	0.6594	0.6594	0.6594
1867	ball playc	19	0.7372	0.8984	0.2483	0.4076	0.2025	0.5	0.4	0.2582	0.262	0.3207	0.5	0.4	0.2781	0.2911	0.2363	0.5	0.5	0.317	0.359	0.3061	0.7255	0.7255	0.7255
1868	Jha state	21	0.8358	0.8856	0.2108	0.3035	0.1172	0.4	0.2	0.2856	0.0948	0.4017	0.5	0.2	0.3332	0.2086	0.2845	0.4	0.5	0.3097	0.2487	0.2887	0.6901	0.6901	0.6901
1869	Canada /	22	0.7811	0.7153	0.2455	0.3406	0.2222	0.3	0.3	0.275	0.1703	0.4445	0.5	0.3	0.4026	0.2919	0.3704	0.3	0.5	0.3575	0.328	0.333	0.6742	0.6742	0.6742
1870	announc	21	0.7914	0.7739	0.2289	0.2878	0.1565	0.3	0.2	0.3698	0.1771	0.4173	0.6	0.2	0.405	0.2878	0.3391	0.3	0.6	0.3585	0.3056	0.306	0.6901	0.6901	0.6901
1871	Virat stre.	20	0.9249	0.8555	0.2507	0.2877	0.1683	0.7	0.2	0.1811	0.1212	0.2735	0.3	0.2	0.2786	0.2726	0.1893	0.7	0.3	0.2807	0.257	0.2826	0.7071	0.7071	0.7071
1872	case Dell	17	0.8964	0.837	0.2559	0.3525	0.2014	0.7	0.4	0.2372	0.2996	0.3021	0.6	0.4	0.2871	0.3525	0.235	0.7	0.6	0.2681	0.3649	0.2446	0.767	0.767	0.767
1873	partner B	20	0.8451	0.883	0.2843	0.3658	0.266	0.6	0.3	0.2777	0.1721	0.3547	0.4	0.3	0.3372	0.2797	0.3192	0.6	0.4	0.3086	0.3043	0.3135	0.7071	0.7071	0.7071
1874	innings T	20	0.8292	0.844	0.369	0.4796	0.2335	0.6	0.2	0.287	0.146	0.3184	0.4	0.2	0.3792	0.3545	0.2759	0.6	0.4	0.4131	0.3686	0.3452	0.7071	0.7071	0.7071
1875	campaigi	23	0.9101	0.756	0.1904	0.2743	0.0836	0.6	0	0.0876	0.0762	0.2508	0.1	0	0.2285	0.259	0.1881	0.6	0.1	0.226	0.2662	0.2618	0.6594	0.6594	0.6594
1876	Saif Kap	22	0.9031	0.7655	0.2631	0.3241	0.0899	0.4	0.2	0.2537	0.1722	0.3371	0.4	0.2	0.3289	0.2633	0.2023	0.4	0.4	0.3199	0.3141	0.2727	0.6742	0.6742	0.6742
1877	kids Supe	15	0.846	0.7126	0.2723	0.3326	0.4005	0.8	0.7	0.2169	0.2687	0.445	0.7	0.7	0.3046	0.2815	0.4005	0.8	0.7	0.2826	0.2925	0.3815	0.8165	0.8165	0.8165

FIGURE 21. Corpus view after complete text processing.

cluster summarization, it utilizes half text document and gives two times better results. As compared to all the other techniques, this technique provides much better text satisfaction results, as presented in Figure 20. Further, it facilitates the corpus enhancement by placing additional text in the most related sub-corpus.

G. SHAP EXPLANATION

The purpose of the SHAP is the comprehension and the prediction of the given instance. We carried out this task by predicting, computing, and assessing the proposed model's basic feature's contributions. The SHAP method performs the explanation by computing the estimation values. These estimation values are Shapley values, which we have computed from the coalitional game theory. It assesses the values of given features as the data about players who are playing in a coalition. Shapley values fairly distribute the payout/prediction among the features. In this model, a player is either an individual feature or a Group of features. A player can also be a group of feature values [11], [14], [31], [43]. Figure 22 shows the contribution of various text processing. Among these simple keywords-based approach has less contribution to assess the relevant documents. The clustering approach provided improvement in the proposed model. Multistage Clustering, based on the summarized-text, has a key role in identifying the related text documents.

VI. SUMMARY

Section-I formally defines the contents of our paper. It provides an understanding of the document summarization, text clustering, and cluster-level summaries. It makes the technical work simple to follow in other sections. It shows how we are stringing together various standard information retrieval techniques and machine learning approaches by stating their reasoning. Our news dataset has diverse nature and carries various topics that justify the Clustering as a reasonable idea. The experimentation presented in Section-V confirms that this technique reduces the computational cost. It handles information diversity by following the principles of text summarization and path-oriented text processing. Section-II contains various corpus processing applications, and it covers the main issues and findings of these related studies. These studies show that efficient corpus processing helps decision making for multiple types of reports analysis [1]. Text processing has certain text aspects that affect the text query results like critical phrases, frequent words, and sentencing impact. Data structures like graphs and trees help serve text aspects as linked information paths to represent these elements. These interlinked paths enable to manage the corpus diversity and generate a convenient corpus traversal during information processing and extraction [2], [3]. The traversing edges and nodes can use summarized text with adequate accuracy. Text summarization techniques extract the crucial part of the given text, and it dramatically helps assess the actual theme of unstructured text. It also helps

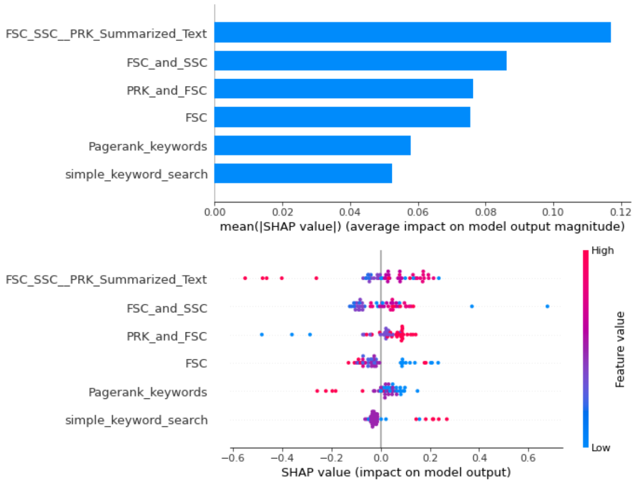


FIGURE 22. SHAP machine learning evaluation to explain individual predictions.

to judge the quality and effectiveness of text query results, with a reduced cost. There exist various forms of text summarizations, and their selection depends on the processing requirements [4]. Automatic text assessment and corpus processing give better results with reduced cost and time [4], [5]. This section shows the relevance of the current technique with other studies over corpus processing. Section-III contains a brief description of all those techniques, which we have applied in the current study to generate metadata and text extraction. Every text analysis initiates from the text pre-processing and mostly include the dictionary analysis. This step remains intact in many applications of natural language processing [4], [7], [11]. Assessing the existence of the desired text in unstructured text documents requires a multi-document review. All the webpages and informal documents have keywords or page rank keywords to represent the actual text document's theme. TF-IDF and PRK generally help to assess the contents of relevant text documents [7], [16], [26]. However, for in-depth text assessments, multiple text processing techniques require systematically joined to relate text segments [9], [14]. This section has briefed about word embeddings, cosine similarities, K-Means clustering, and adjacency matrix. Our corpus processing technique has hierarchically joined these techniques to perform information retrieval by path-oriented corpus subgrouping. Section-IV presents a way to convert the miscellaneous mesh of textual documents into the metadata-based interlinked path-oriented corpus. It has two fundamental concerns: first generating the path-oriented mechanism, and the second is text query processing. At first, we presented the large corpus by three different reduced corpuses by applying text extraction. Through experimentation, Word2Vec reduced text selected to represent the actual corpus and to perform KMC. We have divided the actual-corpus into multiple sub corpuses by applying first stage KMC. Then we have interconnected these clusters into dense AM & AG as a path-oriented corpus representation, and we have further applied KMC over a dictionary of sub-corpus-keywords. We have interconnected the obtained clusters with sparse AM & AG. This procedure provided weighted connected paths during text processing. At the second stage, the query processing starts with comparing the first-stage clusters to identify the relevant clusters. For this, the dense adjacency graph with weights paths facilitates the selection of the appropriate sub-corpuses. The related sub-corpus processing utilizes the sparse adjacency matrix—finally, the related actual documents identified as the processing outcome. Table 2 presents this process in the form of an algorithm. We applied Algorithm 1 by using Python Language (NLTK and Gensim packages). Table 3 shows the sample results of this algorithm. Section-V contains the implementation results of our technique. It shows that joining the metadata with reduced pre-extracted text gives better text processing results. Using this form of metadata, our DCC converted a broad set of documents into small manageable subgroups. We presented these subgroups/sub-corpuses as nodes and joined by weighted paths in AG. These weighted paths contain the CS values to explain the extent of the relation between any two subgroups of documents. These weighted paths provide a traversal approach for the corresponding portions of the corpus. This form of interlinked processing dynamically identifies the strongly related to less related parts of the corpus. This technique works efficiently for the text query by systematically applying the existing text processing approaches. This section presented the results of our technique by tables and graphs. The results presented for the path-oriented corpus processing, proven better than the approaches that do not necessarily apply the path-oriented text processing.

VII. CONCLUSION

The current study's contribution is a technique to overcome the diversity of unstructured text with reduced cost, as the proposed technique provides a language-structure independent mechanism. It utilizes several techniques to identify the corpus' related parts. One of our contributions is an improved corpus visualization instead of dividing the corpus into disjoint sets. We proposed an interrelated-nodes approach that combines homogenous parts of text-document based on words & sentence similarity estimations. Our other contribution is the constructive implementation of the K-Means clustering approach to reduce the redundancy in information processing. Our experiment results show that the proposed technique gives improved performance. The page rank keywords, sentence score algorithm, and Word2Vec model jointly give better similarity estimation. This suggested combination effectively produce more coherent and informative query results and reflects the salient documents.

This technique is practically advantageous and supports all the dictionary-based text processing algorithms. The suggested text processing mechanism can highlight the given text is more and less critical parts during query processing. This technique is an approach that search engine can utilize for any form of unstructured text documents. This proposed work targets the corpus' homogenous parts and identifies the related text for better text query processing. It follows a practical approach by focusing on documents, more related documents, and the most related contents.

Although we have presented different improvements in our presented technique, it also has limitations as well. First, this technique ignores text-summary & sentence positions. It may cause less reading ease for generated summaries. Moreover, we have not considered the structural similarity & features of the language stylometry in our text similarity estimation. Every text engineering & data science research has some inherent limitations of text biasedness, polarity, incomplete text, and spam. Therefore, these issues require more interest in related further research.

This presented technique suggests various future implementations to generate corpus visualization with node & weighted-paths to perform knowledge mining, multiple document processing, and other contents analysis. The research over text summarization has critical applications like information-coverage with reduced text, query processing with redundancy-reduction, topic assessment, and improving readability with short-text. For all these areas of interest, text summarization serves as a multi-objective problem-solving approach. It is our futuristic perspective of later studies in the field of text mining.

ACKNOWLEDGMENT

The authors are thankful to Government College University Faisalabad (GCUF), Pakistan, to provide these research resources.

REFERENCES

[1] K. Liu and N. El-Gohary, "Ontology-based semi-supervised conditional random fields for automated information extraction from bridge inspection reports," *Autom. Construction*, vol. 81, pp. 313–327, Sep. 2017.

[2] Y. Ying, T. Qingping, X. Qinzheng, Z. Ping, and L. Panpan, "A graph-based approach of automatic keyphrase extraction," *Procedia Comput. Sci.*, vol. 107, pp. 248–255, 2017.

[3] A. Gupta, I. Banerjee, and D. L. Rubin, "Automatic information extraction from unstructured mammography reports using distributed semantics," *J. Biomed. Informat.*, vol. 78, pp. 78–86, Feb. 2018.

[4] M. Moradi, "CIBS: A biomedical text summarizer using topic-based sentence clustering," *J. Biomed. Informat.*, vol. 88, pp. 53–61, Dec. 2018.

[5] M. Nasr Azadani, N. Ghadiri, and E. Davoodijam, "Graph-based biomedical text summarization: An itemset mining and sentence clustering approach," *J. Biomed. Informat.*, vol. 84, pp. 42–58, Aug. 2018.

[6] T. Uçkan and A. Karci, "Extractive multi-document text summarization based on graph independent sets," *Egyptian Informat. J.*, vol. 21, no. 3, pp. 145–157, Sep. 2020.

[7] J. M. Sanchez-Gomez, M. A. Vega-Rodríguez, and C. J. Pérez, "Experimental analysis of multiple criteria for extractive multi-document text summarization," *Expert Syst. Appl.*, vol. 140, Feb. 2020, Art. no. 112904.

[8] Z. Deng, F. Ma, R. Lan, W. Huang, and X. Luo, "A two-stage Chinese text summarization algorithm using keyword information and adversarial learning," *Neurocomputing*, to be published.

[9] M. Bidoki, M. R. Moosavi, and M. Fakhrahmad, "A semantic approach to extractive multi-document summarization: Applying sentence expansion for tuning of conceptual densities," *Inf. Process. Manage.*, vol. 57, no. 6, Nov. 2020, Art. no. 102341.

[10] M. Mohd, R. Jan, and M. Shah, "Text document summarization using word embedding," *Expert Syst. Appl.*, vol. 143, Apr. 2020, Art. no. 112958.

[11] B. Mutlu, E. A. Sezer, and M. A. Akcayol, "Candidate sentence selection for extractive text summarization," *Inf. Process. Manage.*, vol. 57, no. 6, Nov. 2020, Art. no. 102359.

[12] P. Wang, B. Xu, J. Xu, G. Tian, C.-L. Liu, and H. Hao, "Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification," *Neurocomputing*, vol. 174, pp. 806–814, Jan. 2016.

[13] N. Yadav and V. Kumar, "A novel technique for automatic retrieval of embedded text from books," *Optik*, vol. 127, no. 20, pp. 9538–9550, Oct. 2016.

[14] Y.-C. Wu, "Language independent Web news extraction system based on text detection framework," *Inf. Sci.*, vol. 342, pp. 132–149, May 2016.

[15] S. Al-Anazi, H. AlMahmoud, and I. Al-Turaiki, "Finding similar documents using different clustering techniques," *Procedia Comput. Sci.*, vol. 82, pp. 28–34, 2016.

[16] M. Peng, B. Gao, J. Zhu, J. Huang, M. Yuan, and F. Li, "High quality information extraction and query-oriented summarization for automatic query-reply in social network," *Expert Syst. Appl.*, vol. 44, pp. 92–101, Feb. 2016.

[17] A. Trabelsi and O. R. Zaïane, "Extraction and clustering of arguing expressions in contentious text," *Data Knowl. Eng.*, vol. 100, pp. 226–239, Nov. 2015.

[18] T. Wei, Y. Lu, H. Chang, Q. Zhou, and X. Bao, "A semantic approach for text clustering using WordNet and lexical chains," *Expert Syst. Appl.*, vol. 42, no. 4, pp. 2264–2275, Mar. 2015.

[19] H. Hashimi, A. Hafez, and H. Mathkour, "Selection criteria for text mining approaches," *Comput. Hum. Behav.*, vol. 51, pp. 729–733, Oct. 2015.

[20] S. A. Babar and P. D. Patil, "Improving performance of text summarization," *Procedia Comput. Sci.*, vol. 46, pp. 354–363, 2015.

[21] I. Alsmadi and I. Alhami, "Clustering and classification of email contents," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 27, no. 1, pp. 46–57, 2015.

[22] J. M. Sanchez-Gomez, M. A. Vega-Rodríguez, and C. J. Pérez, "Parallelizing a multi-objective optimization approach for extractive multi-document text summarization," *J. Parallel Distrib. Comput.*, vol. 134, pp. 166–179, Dec. 2019.

[23] O. Rouane, H. Belhadef, and M. Bouakkaz, "Combine clustering and frequent itemsets mining to enhance biomedical text summarization," *Expert Syst. Appl.*, vol. 135, pp. 362–373, Nov. 2019.

[24] B. Mutlu, E. A. Sezer, and M. A. Akcayol, "Multi-document extractive text summarization: A comparative assessment on features," *Knowl.-Based Syst.*, vol. 183, Nov. 2019, Art. no. 104848.

[25] A. Joshi, E. Fidalgo, E. Alegre, and L. Fernández-Robles, "SummCoder: An unsupervised framework for extractive text summarization based on deep auto-encoders," *Expert Syst. Appl.*, vol. 129, pp. 200–215, Sep. 2019.

[26] R. Adelia, S. Suyanto, and U. N. Wisesty, "Indonesian abstractive text summarization using bidirectional gated recurrent unit," *Procedia Comput. Sci.*, vol. 157, pp. 581–588, 2019.

[27] M. Mohamed and M. Oussalah, "SRL-ESA-TextSum: A text summarization approach based on semantic role labeling and explicit semantic analysis," *Inf. Process. Manage.*, vol. 56, no. 4, pp. 1356–1372, Jul. 2019.

[28] F. B. Goularte, S. M. Nassar, R. Fileto, and H. Saggion, "A text summarization method based on fuzzy rules and applicable to automated assessment," *Expert Syst. Appl.*, vol. 115, pp. 264–275, Jan. 2019.

[29] M. A. Mosa, A. S. Anwar, and A. Hamouda, "A survey of multiple types of text summarization with their satellite contents based on swarm intelligence optimization algorithms," *Knowl.-Based Syst.*, vol. 163, pp. 518–532, Jan. 2019.

[30] N. Alami, M. Meknassi, and N. En-nahnahi, "Enhancing unsupervised neural networks based text summarization with word embedding and ensemble learning," *Expert Syst. Appl.*, vol. 123, pp. 195–211, Jun. 2019.

[31] J. M. Sanchez-Gomez, M. A. Vega-Rodríguez, and C. J. Pérez, "Comparison of automatic methods for reducing the Pareto front to a single solution applied to multi-document text summarization," *Knowl.-Based Syst.*, vol. 174, pp. 123–136, Jun. 2019.

[32] H. Van Lierde and T. W. S. Chow, "Query-oriented text summarization based on hypergraph transversals," *Inf. Process. Manage.*, vol. 56, no. 4, pp. 1317–1338, Jul. 2019.

[33] H. Van Lierde and T. W. S. Chow, "Learning with fuzzy hypergraphs: A topical approach to query-oriented text summarization," *Inf. Sci.*, vol. 496, pp. 212–224, Sep. 2019.

[34] X. Qian, M. Li, Y. Ren, and S. Jiang, "Social media based event summarization by user–text–image co-clustering," *Knowl.-Based Syst.*, vol. 164, pp. 107–121, Jan. 2019.

[35] P. Janaszkiewicz and P. Różewski, "The method of multidimensional approach to text summarization," *Procedia Comput. Sci.*, vol. 159, pp. 2189–2196, 2019.

[36] C. Gulden, M. Kirchner, C. Schüttler, M. Hinderer, M. Kampf, H.-U. Prokosch, and D. Toddenroth, "Extractive summarization of clinical trial descriptions," *Int. J. Med. Informat.*, vol. 129, pp. 114–121, Sep. 2019.

[37] R. Y. Rumagit, N. Setiyawati, and D. H. Bangkalang, "Comparison of graph-based and term weighting method for automatic summarization of online news," *Procedia Comput. Sci.*, vol. 157, pp. 663–672, 2019.

[38] D. Patel, S. Shah, and H. Chhinkaniwala, "Fuzzy logic based multi document summarization with improved sentence scoring and redundancy removal technique," *Expert Syst. Appl.*, vol. 134, pp. 167–177, Nov. 2019.

[39] N. Saini, S. Saha, A. Jangra, and P. Bhattacharyya, "Extractive single document summarization using multi-objective optimization: Exploring self-organized differential evolution, grey wolf optimizer and water cycle algorithm," *Knowl.-Based Syst.*, vol. 164, pp. 45–67, Jan. 2019.

[40] P. Verma and H. Om, "MCRMR: Maximum coverage and relevancy with minimal redundancy based multi-document summarization," *Expert Syst. Appl.*, vol. 120, pp. 43–56, Apr. 2019.

[41] J. Chambua, Z. Niu, and Y. Zhu, "User preferences prediction approach based on embedded deep summaries," *Expert Syst. Appl.*, vol. 132, pp. 87–98, Oct. 2019.

[42] M. Rajangam and C. Annamalai, "Extractive document summarization using an adaptive, knowledge based cognitive model," *Cognit. Syst. Res.*, vol. 56, pp. 56–71, Aug. 2019.

[43] G. Fuentes-Pineda and I. V. Meza-Ruiz, "Topic discovery in massive text corpora based on min-hashing," *Expert Syst. Appl.*, vol. 136, pp. 62–72, Dec. 2019.

[44] R. McCreadie, S. Rajput, I. Soboroff, C. Macdonald, and I. Ounis, "On enhancing the robustness of timeline summarization test collections," *Inf. Process. Manage.*, vol. 56, no. 5, pp. 1815–1836, Sep. 2019.

[45] M.-T. Nguyen, V. C. Tran, X. H. Nguyen, and L.-M. Nguyen, "Web document summarization by exploiting social context with matrix co-factorization," *Inf. Process. Manage.*, vol. 56, no. 3, pp. 495–515, May 2019.

[46] X. Mao, H. Yang, S. Huang, Y. Liu, and R. Li, "Extractive summarization using supervised and unsupervised learning," *Expert Syst. Appl.*, vol. 133, pp. 173–181, Nov. 2019.

[47] V. Suárez-Paniagua, R. M. Rivera Zavala, I. Segura-Bedmar, and P. Martínez, "A two-stage deep learning approach for extracting entities and relationships from medical texts," *J. Biomed. Informat.*, vol. 99, Nov. 2019, Art. no. 103285.

[48] T. Hou, B. Yannou, Y. Leroy, and E. Poirson, "Mining customer product reviews for product development: A summarization process," *Expert Syst. Appl.*, vol. 132, pp. 141–150, Oct. 2019.

[49] S. Gerani, G. Carenini, and R. T. Ng, "Modeling content and structure for abstractive review summarization," *Comput. Speech Lang.*, vol. 53, pp. 302–331, Jan. 2019.

[50] Y. Wu, Y. Li, and Y. Xu, "Dual pattern-enhanced representations model for query-focused multi-document summarisation," *Knowl.-Based Syst.*, vol. 163, pp. 736–748, Jan. 2019.

[51] U. P., V. K. Govindan, and S. D. Madhu Kumar, "Enhanced sparse representation classifier for text classification," *Expert Syst. Appl.*, vol. 129, pp. 260–272, Sep. 2019.

[52] (2018). *NEWS SUMMARY: Generating Short Length Descriptions of News Articles*. [Online]. Available: https://www.kaggle.com/sunnysai12345/news-summary

[53] (2020). *Implementing Word2Vec With Gensim Library in Python*. https://stackabuse.com/implementing-word2vec-with-gensim-library-inpython/

[54] (Sep. 5, 2019). *Python | Extractive Text Summarization Using Gensim*. [Online]. Available: https://www.geeksforgeeks.org/python-extractive-text-summarizationusing-gensim/

[55] (May 26, 2018). *Text Summarisation With Gensim (TextRank Algorithm)*. [Online]. Available: https://medium.com/@shivangisareen/text-summarisation-with-gensimtextrank-46bbb3401289

[56] R. Glauber and D. Barreiro Claro, "A systematic mapping study on open information extraction," *Expert Syst. Appl.*, vol. 112, pp. 372–387, Dec. 2018.

[57] X. Xie, Y. Fu, H. Jin, Y. Zhao, and W. Cao, "A novel text mining approach for scholar information extraction from Web content in chinese," *Future Gener. Comput. Syst.*, vol. 111, pp. 859–872, Oct. 2020.

[58] F. Jenhani, M. S. Gouider, and L. B. Said, "Hybrid system for information extraction from social media text: Drug abuse case study," *Procedia Comput. Sci.*, vol. 159, pp. 688–697, 2019.

[59] Z. Nasar, S. W. Jaffry, and M. K. Malik, "Textual keyword extraction and summarization: State-of-the-art," *Inf. Process. Manage.*, vol. 56, no. 6, Nov. 2019, Art. no. 102088.

[60] T. T. Nguyen, P. Krishnakumari, S. C. Calvert, H. L. Vu, and H. van Lint, "Feature extraction and clustering analysis of highway congestion," *Transp. Res. Part C: Emerg. Technol.*, vol. 100, pp. 238–258, Mar. 2019.

[61] A. M. Thomas and M. G. Resmipriya, "An efficient text classification scheme using clustering," *Procedia Technol.*, vol. 24, pp. 1220–1225, 2016.

MUHAMMAD YAHYA SAEED is currently working as a Lecturer with the Department of Software Engineering, Government College University Faisalabad (GCUF).

MUHAMMAD AWAIS is currently working as an Assistant Professor and an Incharge of the Department of Software Engineering, Government College University Faisalabad (GCUF).

RAMZAN TALIB is currently working as a Professor and the Chairman of the Department of Computer Science, Government College University Faisalabad (GCUF).

MUHAMMAD YOUNAS (Member, IEEE) received the Ph.D. degree from the Faculty of Engineering, School of Computing, Universiti Teknologi Malaysia (UTM). He is currently working as an Assistant Professor with the Department of Computer Science, Government College University Faisalabad, Pakistan. His research interests include software engineering, agile software development, cloud computing, and code clone detection.