Journal of Information Technology and Digital World (ISSN: 2582-418X)
www.irojournals.com/itdw/

TF-IDF Vectorization and Clustering for Extractive Text Summarization

Muthu Virumeshwaran T[1], R Thirumahal[2]

[1]PG Scholar, Department of Computer Science and Engineering, PSG College of Technology, Coimbatore, India

[2]Assistant Professor, Department of Computer Science and Engineering, PSG College of Technology, Coimbatore, India

E-mail: [1]virumeshsivakumar@gmail.com, [2]trk.cse@psgtech.ac.in

Abstract

Extractive document summarization is a vital technique for condensing large volumes of text while retaining key information. This research introduces a dynamic feature space mapping approach to enhance extractive document summarization, aiming to succinctly encapsulate key information from extensive text volumes. The proposed method involves extracting various document properties like term frequency, sentence length, and position to comprehensively describe content. By employing a mapping function, these features are projected into a dynamic feature space, enhancing summarization efficiency and feature clarity. Clustering similar phrases in this space facilitates easier sentence grouping, aiding summary creation. Leveraging TF-IDF vectorization, the most representative phrases are chosen from each cluster based on importance and diversity. This process culminates in generating a high-quality document summary quickly and systematically. The dynamic mapping method streamlines sentence grouping, systematically capturing essential document attributes. This approach addresses challenges in extractive summarization, contributing significantly to automated text summarization. Its applicability spans domains requiring rapid extraction of information from vast textual data.

Keywords: Extractive Summarization, Dynamic Feature Space Mapping, TF-IDF Vectorization, K-Means Clustering Document Preprocessing, Sentence Clustering,

DOI: https://doi.org/10.36548/jitdw.2024.1.008
Received: 08.03.2024. received in revised form: 04.04.2024, accepted: 16.04.2024, published: 29.04.2024

Summarization Efficiency, Feature Extraction, Natural Language Processing, Information Retrieval

1. Introduction

In an era where information is constantly being produced, it is critical to be able to understand, arrange, and extract meaningful information from large amounts of textual data. News stories, academic papers, social media postings, and other types of textual material abound in the digital world. The collection of written words is frequently the key source for knowledge acquisition, information sharing, and decision-making in a variety of fields, including business, academia, and media. But although the expanding textual material is a rich source of knowledge, it also breeds complexity and cognitive stress.

Information overload is a long-standing issue that predates the internet However, the more and more textual data there is available today the harder it is to seek valuable insight. Text summarization and grouping Therefore, text summary or grouping is like a potion to solve these problems. Natural Language Processing, or NLP, is the art and science of condensing large texts into concise summaries and organizing them into logical clusters. They are vital to the smooth digestion of textual data and its use for effective learning. This is the main focus of this research paper, which seeks to enable an automated text grouping and summarizing system based on advanced NLP training techniques. Our goal is to enable more effective ways to handle and understand textual data by putting these strategies into practice and evaluating their efficacy. This section, which provides a summary of the study objectives, the problem's relevance, and the suggested approach for solving it, acts as the project's entry point.

This paper will take a look at various aspects of text summarization and clustering in the following sections such as pre-processing the data, feature extraction and reducing it to vector form (e.g., using a TF-IDF weight framework), K-Means clustering etc. In addition, the paper will cover potential future developments and field expansions as well as the ROUGE metric evaluation of the generated summaries. This research could be an essential tool for the researchers to unravel the complexities of textual data and uncover the significant ramifications of these technologies for contemporary information management. The rest of the paper is organized as follows. Following the introduction in section 1, the related work is presented in

section 2. The proposed system in Section 3 The experimental setup and assessment procedure used to gauge the system's performance are described in Section 4. The paper is finally concluded in Section 5, which also identifies future avenues for this field of study.

2. Related Work

The goal of extractive document summarization, as described in [1], is to take the most important information from a document, eliminate any unnecessary information, and then organize the remaining information into a logical structure. In order to summarize documents while maintaining their essence without altering the meaning, automated document summarization is required. ExDoS uses dynamic local feature weighting to iteratively reduce the classifier's error rate in each cluster. The research paper [2] proposes the use of a deep learning model to achieve text summarization, for which it introduces an entirely new technique. The approach proposed as a solution to meeting this challenge is based on using Bi-LSTM (Bidirectional Long Short-Term Memory) machine learning model for text summary at the sentence and paragraph levels. The [3] paper's most important research application involves using deep learning models to summarize news documents related to COVID-19. The text undergoes a variety of preprocessing procedures, such as lowercase, contractions, tokenization, and GloVe word embedding. The Adam optimization approach is used by the researchers using certain settings during the training phase. To evaluate the models' performance, they make use of the ROUGE scores and loss value. In the paper [4], Zhang et al. delve into the realm of abstractive text summarization with a focus on its deep learning-based methodologies. Abstractive text summarization is the process of creating brief, logical summaries that encapsulate the main ideas of the original text. The authors introduce the significance of abstractive text summarization in processing large volumes of textual data. Abstractive summarization necessitates a deeper comprehension of the content because it creates new sentences that might not be found in the original text. The deep learning models consist of Transformers, Gated Recurrent Units (GRUs), Long Short-Term Memory (LSTM) networks, and Recurrent Neural Networks (RNNs). In paper [5], Gambhir and Gupta review the topic of deep learning-based extractive text summarization. The authors stress that extractive text summary is helpful as it aims at deeply mining relevant information out of long papers. It's also essential for verifying that a given summary fits its context and has been written clearly with no mistakes. By treating individual terms in the original text as having different degrees of

relevance, such a process lets the model select which phrases to include for use in its summary. The paper carefully explains what dataset the authors used to train and test their model. The authors of paper [6] provide a unique method for summarising lengthy texts, with an emphasis on scholarly papers. The authors argue that typical approaches that try to provide a comprehensive summary all at once can be wasteful and describe a divide-and-conquer strategy that splits the text into smaller segments and trains a neural model to summarise each component individually. In [7], a neural attention model that combines the best features of an extractor and an abstracter model is used to create abstractive summaries in a novel way. The word features such as sentence position, number of numerals, POS tags, NE tags, term weights, and number of proper nouns are all included in the word embeddings used by the proposed model. Attention layers use word and sentence attention parameters, respectively, to highlight the most important information for the abstracter and extractor models.[8] suggests an unsupervised extractive summarization model called Learning-Free Integer Programming Summarizer (LFIP-SUM), which reduces a source document into a logical summary with pertinent information. As a method of data augmentation, [9] suggests a unique way to turn Wikipedia into a sizable query-focused summarization dataset called WIKI REF. To extract summaries from documents, they create a model called Q-BERT, which is based on BERT. The DUC benchmarks are used to refine the model after it has been trained on the WIKI REF dataset. The results show that Q-BERT outperforms strong comparison systems when it comes to data augmentation and subnetwork fine-tuning. [10] This study presents a comprehensive overview of feature-based artificial text summarizing methods. The authors provide a methodical analysis of relevant research papers pertaining to different datasets, methods, and assessment criteria for text summarizing that are gathered from several sources. [11] introduces FFSUM (forward-backward forward summary) which is a novel abstractive summarizing method that integrates multi-granular multi-relational information to enhance the quality of produce output. FFSUM corrects this shortcoming by introducing a fine-grained factual graph and simulating condensed semantic relationships among facts. [12] provides a novel approach to extractive multi-document Arabic text summarization. Utilizing clustering-based methods and evolutionary multi-objective optimization techniques, the system focuses on sifting through a corpus of texts to extract pertinent and noteworthy material. Using the k-medoid clustering method with the Silhouette measure, topics of the related set of documents are extracted during the clustering stage. At the multi-objective optimization stage, three goals—

coverage, sentence relevancy, and redundancy elimination—are simultaneously optimized using an evolutionary method. [13] presents a two-stage summarization model that combines extractive and abstractive approaches to address the difficulties in document-level summarization. Using a similarity matrix or a pseudo-title that takes into account elements like sentence and paragraph position, the first stage extracts significant sentences. The extracted sentences are rewritten and restructured into a summary in the second stage using the beam search algorithm. The freshly created summary sentence functions as the subsequent round's pseudo-summary, and the globally optimal pseudo-title serves as the ultimate summary. [14] covers the broad subject of multi-document summarization, hitting key points like its history, the hurdles it experiences, and the present advances in technology. It goes into detail about different multi-document summarization methods such as those based on graphs, clusters, or optimization These include juggling various document types and languages, dealing with messy and repeating data, and addressing the quality versus quantity in summarizing.[15] offers a lively approach to put together brief extracted information along with metadata, making text clusters. This allows us to use multistage organization to study a network of many unstructured phrases and make a linked set of many nodes with different importance.

3. Proposed Work

3.1 System Design Architecture

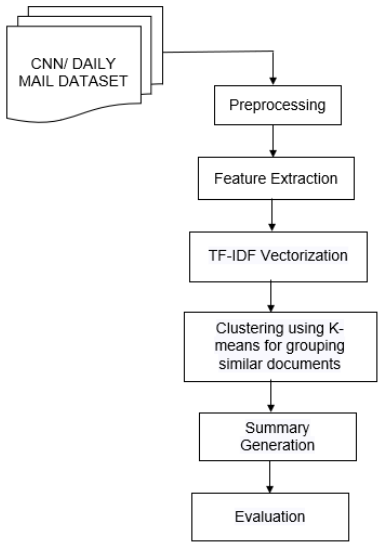

Figure 1. Conceptual Flow of Extractive Text Summarization

The conceptual flow of the suggested extractive text summarization approach is depicted in Figure.1. The process comprises many essential phases intended to convert unprocessed textual input into succinct and enlightening document synopses.

3.1.1. Data Preprocessing: The entire process begins with data preparation, an essential initial step in cleaning and preparing the text data for further examination. The text is tokenized into words and phrases at this point, and any unnecessary information is removed. Additionally, common stopwords are eliminated in order to enhance the text's quality.

3.1.2. Feature Extraction: In addition to preprocessing, the proposed methodology makes the most of feature extraction techniques to extract appropriate data from the text. Text is converted into numerical representations at this point so that it may be usable for other analysis, such as vectorization with the TF-IDF (Term Frequency-Inverse Document Frequency) technique.

3.1.3. TF-IDF Vectorization: TF-IDF (Term Frequency-Inverse Document Frequency) vectorization is used in the suggested technique. TF-IDF plays a crucial role in the process by transforming the text into a numerical representation. A vector of TF-IDF weights connected to vocabulary terms represents each document. Comparing documents and doing quantitative analysis are made possible by this transition.

3.1.4. Clustering: The next stage of the procedure is clustering, which groups papers with similar characteristics. Documents are grouped into 'k' clusters using the K-Means clustering technique, where 'k' is a user-defined value. Finding links and patterns in documents is made easier with the use of clustering.

3.1.5. Summarization: The procedure of text summarising is the last part of the technique. The algorithm finds texts that have thematic material within the same cluster by computing cosine similarity. These papers' best sentences are chosen, which helps to create succinct and coherent document summaries.

3.2 Manual Tracing: For a clear understanding on how the algorithm works, two sample article is taken and explained below.

3.2.1. Input:

Sample Documents:

Document 1: "Machine learning is interesting."

Document 2: "Machine learning is challenging."

Document 3: "Natural language processing is interesting."

3.2.2. Preprocessing

The initial step is tokenization which is to split the documents into words. Then the stopwords which are the words that carry least significance to the document such as is, was, the are removed

Document 1: Machine learning interesting.

Document 2: Machine learning challenging.

Document 3: Natural language processing interesting.

3.2.3. Feature Extraction

Extract a set of features from the input documents, such as word frequency and sentence length.

3.2.4. TF-IDF Vectorization

- **TF (Term Frequency):** This column represents the frequency of each term in each document. For example, the term "Machine" occurs once in Document 1 (TF = 1).
- **DF (Document Frequency):** This column indicates how many documents contain each term. For example, "Machine" appears in two documents (DF = 2).
- **IDF (Inverse Document Frequency):** IDF is calculated as log(N / (1 + DF(term))), where N is the total number of documents. In this case, N = 3. For example, the IDF value for "Machine" is approximately 0.405.

- **TF-IDF (Term Frequency-Inverse Document Frequency):** This is the product of TF and IDF for each term in each document. For example, the TF-IDF value for "Machine" in Document 1 is approximately 0.405.

TF-IDF values are calculated for each term in each document as shown in (Table 1), providing a numerical representation of the documents. These values capture the importance of each term within the context of the entire document collection.

3.2.5. Clustering

After clustering, document summaries are generated for each cluster. The summary is constructed by selecting sentences that are highly representative of the cluster's content.

Table 1. TF-IDF Vectorization

Term	**TF Doc 1**	**TF Doc 2**	**TF Doc 3**	**DF**	**IDF**	**TF-IDF Doc 1**	**TF-IDF Doc 2**	**TF-IDF Doc 3**
Machine	1	1	0	2	0.405	0.405	0.405	0
Learning	1	1	0	2	0.405	0.405	0.405	0
Interesting	1	0	1	2	0.405	0.405	0	0.405
Challenging	0	1	0	1	1.099	0	1.099	0
Natural	0	0	1	1	1.099	0	0	1.099
Language	0	0	1	1	1.099	0	0	1.099
Processing	0	0	1	1	1.099	0	0	1.099

Sample of three clusters is used to illustrate the manual tracing. These clusters are created based on the TF-IDF values calculated in the previous step:

Cluster 1: Documents:

Document 1: "Machine learning interesting."

Document 2: "Machine learning challenging."

Cluster Center: (0.405, 0.405, 0, 0, 0, 0, 0)

Cluster 2: Documents:

Document 3: "Natural language processing is interesting."

Cluster Center: (0, 0, 1.099, 1.099, 1.099, 1.099, 1.099, 0)

Cluster 3: No documents in this cluster.

3.2.6 Summary Generation

Calculate the cosine similarity between each document and its cluster center to identify the most representative documents. For each cluster, the document with the highest cosine similarity is selected as the most representative. The selected document's sentences are extracted to form the cluster summary.

3.2.7 Output

Cluster 1 Summary: "Machine learning interesting." "Machine learning challenging."

Cluster 2 Summary: "Natural language processing interesting."

Cluster 3 Summary: This cluster has no documents

The summary for each document is generated by selecting sentences from the same document with high TF-IDF scores, as well as sentences from other documents in the same cluster with the highest similarity scores.

4. Experimental Setup and Evaluation Process

The experimental setup and evaluation process for the proposed text summarization is described below:

4.1 Experimental Setup

4.1.1 Dataset Details

	A	B	C
1	id	article	highlights
2	92c514c913c0bdfe25341af9fd72b29db544099b	Ever noticed how plane seats appear to be getting sma	Experts question if packed out planes are
3	2003841c7dc0e7c5b1a248f9cd536d727f27a45a	A drunk teenage boy had to be rescued by security afte	Drunk teenage boy climbed into lion enclosure
4	91b7d2311527f5c2b63a65ca98d21d9c92485149	Dougie Freedman is on the verge of agreeing a new tw	Nottingham Forest are close to extending
5	caabf9cbdf96eb1410295a673e953d304391bfbb	Liverpool target Neto is also wanted by PSG and clubs	Fiorentina goalkeeper Neto has been linked with
6	3da746a7d9afcaa659088c8366ef6347fe6b53ea	Bruce Jenner will break his silence in a two-hour interv	Tell-all interview with the reality TV star, 69, will
7	5ed5e3fbd235a8046cd3b87f4a1aa51b856c8ec3	This is the moment that a crew of firefighters struggle	Giant pig fell into the swimming pool at his
8	6394f51b120ceb3da5e7b53dd5167fc4cf80b514	The amount of time people spend listening to BBC radi	Figures show that while millions still tune in they
9	98be9b2d558c17df8a13597195957a7c8587ddcd	(CNN)So, you'd like a "Full House" reunion and spinoff	Show will return with a one-hour special,
10	57f68638739c3a1de8d9922b389d6ded39977012	At 11:20pm, former world champion Ken Doherty pott	Reanne Evans faced Ken Doherty in World
11	20778c35c19d741cc182719de336d71e1a0b228e	A gang of six men have been jailed for a total of 31 yea	Gang have been jailed for a total of 31 years for
12	61a24035b8cae4ee5fab3e79b4409c33e5259565	Biting his nails nervously, these are the first pictures of	Mohammed Ali Malek, 27, has been charged
13	f8179711ef608df3b003b001441eb68495f38d17	South Korea Ambassador Lippert is now wearing a dyn	Lippert, 42, suffered deep gashes to his hand
14	67f7110c19df40e51f2a6ba92ad77356e510a4fb	England captain Alastair Cook completed a much-neec	Alastair Cook completed his century on the
15	391b4610e3a9dcefd61e428155cc5afb2683124c	(CNN)Warren Weinstein, who appears to have been th	U.S. hostage Warren Weinstein is believed to

Figure 2. Snapshot of Dataset

The CNN/DailyMail Dataset is an English-language dataset containing just over 300,000 unique news articles written by journalists at CNN and the Daily Mail. Figure 2 showcases a snapshot of the dataset. A subset of this dataset, consisting of 11,490 records with three fields each, has been used for this paper. The dataset description is provided below.

- **Data-Fields**

id: a string containing the SHA1 hash of the URL from which the story was retrieved in hexadecimal format.

article: a string containing the body of the news article

highlights: a string containing the highlight of the article as written by the article author.

4.2 Loading the Dataset

The dataset was loaded into the Pandas DataFrame to enable the manipulation of data using Pandas.

4.3 Pre-Processing: Preprocessing refers to the steps taken to clean and prepare the input data, which in the case are the original documents. These steps ensure that the data is in a suitable format for further analysis. Common preprocessing steps include: Text Cleaning: This involves

removing any irrelevant characters, symbols, or formatting from the text. It ensures that the data is uniform and free from artifacts that might interfere with the summarization process. Tokenization: Text is broken down into sentences and words. This step facilitates further analysis, as we are working with individual sentences and words to identify key content for summarization. Stop-word Removal: Stop words are common words (e.g., "the," "and," "in") that don't carry much meaning and can be safely removed to focus on more significant terms.

4.4 Feature Extraction

In this paper, feature extraction is an essential step. Here are some definitions for the traits or qualities that are crucial for document summaries. These characteristics form the foundation for evaluating the papers' structure and content. characteristics like:

- Term Frequency: determining out how frequently a given word or phrase appears in the text. This reveals which terms are more common and could be important to summarise.
- Sentence Position: Assessing where a sentence appears in the document. For instance, the first and last sentences of a document may contain essential information.
- Sentence Length: Measuring the length of sentences. Shorter or longer sentences can be indicative of important content**.**

4.5 TF-IDF Vectorization

One method for representing documents as numerical vectors is called TF-IDF (Term Frequency-Inverse Document Frequency) vectorization. This step is crucial to your research because it transforms the textual information you retrieved into a format suitable for clustering and analysis.

- Term Frequency (TF): This element measures the frequency with which a term occurs in a document. It assists in highlighting key phrases in every document.
- Inverse Document Frequency (IDF): IDF calculates a term's uniqueness or frequency throughout the whole dataset. It's employed to highlight how significant uncommon phrases are.

- Vectorization: The combination of TF and IDF values for each term results in a vector representation of the document. This allows you to compare documents and sentences based on their content

4.6 Clustering Algorithm

A crucial step in the summarization process is clustering. Finding the essential material for the summary requires grouping phrases with similar content together, which is accomplished by clustering. This is how it operates:

4.6.1 Clustering Objective: Assembling phrases with comparable TF-IDF vector representations is the aim of clustering. Conceptually similar sentences ought to be grouped together.

4.6.2 K-Means Clustering: K-Means groups sentences into clusters based on their vector representations.

4.6.3 Cluster Assignments: After clustering, each sentence is assigned to a specific cluster, enabling us to work with sentence groups rather than individual sentences as shown in Figure 3 and 4.

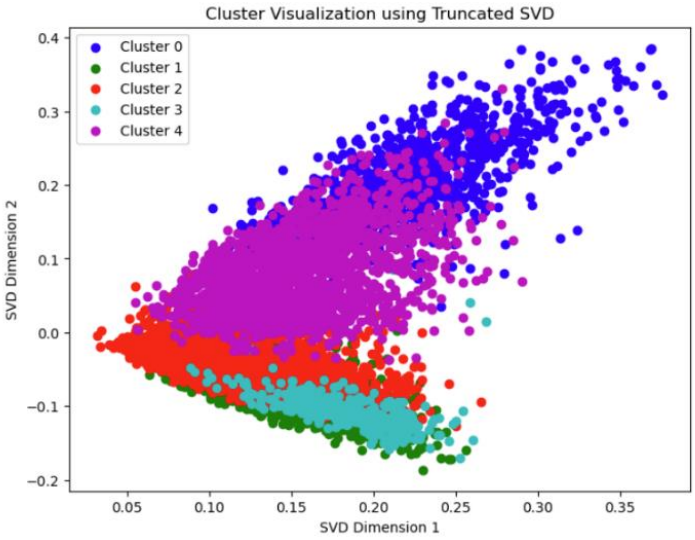

Figure 3. 2D Visualization

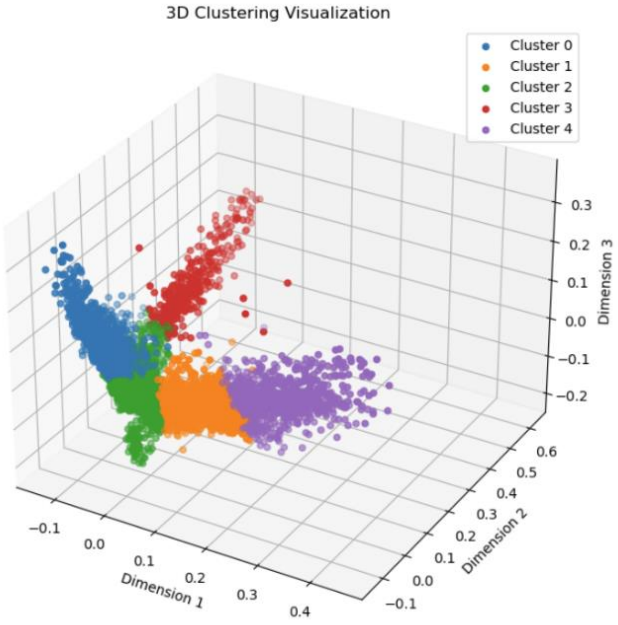

Figure 4. 3D Visualization

4.6.4 Sample Summaries

```
Cohesive Summary for Document 1:

Mr Stapelkamp said lions become adept hunting prey mature , adding : 'They learn probably better eati
ng warthogs ' Mr Stapelkamp explained pair lions mature become adept hunting prey  However , lions co
ntinue prowl , animal abruptly turns around charges towards two predators , successfully scaring retr
eating territory  Horrible scenes : Around 30 people , many children , witnessed Melvin , born zoo 20
10 named local newspaper competition , die enclosure

Cohesive Summary for Document 2:

’ Mobile phones perfect item pickpockets — ‘ ’ called Apple BlackBerry picking ’ — information owner
contained Sim card useful thief phone ’ intrinsic value contains wealth personal information possibly
even bank details  ' quick trip nearby town Terrigal , Ms Freedman family tried find power , proved f
ruitless found numerous homes lost roofs normally flat beach pounding waves Balmain resident describe
d ‘ mental ’  comes second intense low pressure system , centred coast Hunter region , predicted brin
g heavier rains large seas thunderstorms large parts northern Sydney Central Coast Wednesday


Cohesive Summary for Document 3:

First Salomon Rondon , almost entirely ineffectual front opening period , gifted ball three yards goa
l following dubious defending set piece , Venezuelan striker could n't sort feet , stumbled ball away
target  Fabricio threw shirt ground shown red card shouted ' 'm leaving , 'm leaving ' missed penalty
debut , substituted later February 36 minutes needing oxygen playing high-altitude Estadio Hernando S
iles La Paz  Zenit : Lodygin , Criscito , Neto , Lombaerts , Smolnikov ; Garcia , Witsel ; Danny , Sh
atov , Hulk , Rondon ( Kerzhakov 83 mins ) Subs used : Malafeev , Rodic , Anyukov , Tymoshchuk , Shey
daev , Arshavin

Cohesive Summary for Document 4:

`` Bruce incredibly courageous inspiration , proud entrusted deeply personal important story , '' sai
d Jeff Olde , executive vice president programming development E ! `` series present unfiltered look
Bruce boldly steps uncharted territory true first time  added : 'As family always support 100 per cen
t know see things online divide one ’ support one ’ support know ’ ridiculous ’ rumors love everyone
family support everyone equally 're excited watch tonight ’ excited Bruce tell story Diane Sawyer  ''
going OK ? know ? gon na OK ? '' ' peep trip either : Khloe Kardashian , Kendall Jenner Kylie Jenner
nothing say former Olympian attended Calvin Klein Jeans ' celebration launch # mycalvins Denim Series
Chateau Marmont Thursday
```

Figure 5. Sample Summaries

The Figure.5 shows the generated sample summaries after clustering and sentence scoring using TF-IDF Vectorization

5. Conclusion

Using machine learning algorithms, feature extraction, and natural language processing techniques, we have investigated the creation of an extractive text summarization system in this research. The research encompassed several key steps, including data preprocessing, TF-IDF vectorization, clustering, and summary generation. We have shown how these stages may be used in practice by manually tracing the algorithm and implementing the code. Data pretreatment methods included text cleaning, stop-word removal, and sentence and word tokenization. Text data could be converted into numerical representations using TF-IDF vectorization, which allowed for the capture of word significance in texts. K-Means clustering grouped texts that were comparable to each other, making it easier to organize and analyze massive datasets. By choosing representative phrases from each cluster according to cosine similarity, summaries were generated. The practical consequences of this study may be found in several disciplines, such as content summarizing, document organization, and information retrieval. The system's capacity to provide succinct summaries can improve information extraction from big document collections and save time. Researchers, journalists, and anyone working with large amounts of textual data may all benefit from it. Our study establishes the groundwork for more investigation and advancements in the text summarization and clustering domain. This work concludes by highlighting the need of text summarization and clustering for effectively managing and comprehending massive amounts of textual data. As the need for automatic text interpretation grows in an era of abundant and complicated data, it offers a solid foundation for future developments and applications.

References

[1] Ghodratnama, S., Beheshti, A., Zakershahrak, M. and Sobhanmanesh, F.,"Extractive document summarization based on dynamic feature space mapping". IEEE Access, 8, pp.139084-139095, 2020.

[2] Yadav, A.K., Singh, A., Dhiman, M., Vineet, Kaundal, R., Verma, A. and Yadav, D. "Extractive text summarization using deep learning approach". International Journal of Information Technology, 14(5), pp.2407-2415, 2022.

[3] Hayatin, N., Ghufron, K.M. and Wicaksono, G.W. "Summarization of COVID-19 news documents deep learning-based using transformer architecture". TELKOMNIKA (Telecommunication Computing Electronics and Control), 19(3), pp.754-761,2021.

[4] Zhang, M., Zhou, G., Yu, W., Huang, N. and Liu, W., "A comprehensive survey of abstractive text summarization based on deep learning". Computational intelligence and neuroscience, 2022.1-21

[5] Gambhir, M. and Gupta, V. "Deep learning-based extractive text summarization with word-level attention mechanism". Multimedia Tools and Applications, 81(15), pp.20829-20852,2022

[6] Gidiotis, A. and Tsoumakas, G. "A divide-and-conquer approach to the summarization of long documents". IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28, pp.3029-3040, 2020.

[7] Dilawari, A., Khan, M.U.G., Saleem, S. and Shaikh, F.S. "Neural Attention Model for Abstractive Text Summarization Using Linguistic Feature Space". IEEE Access, 11, pp.23557-23564.,2023.

[8] Jang, M. and Kang, P. "Learning-free unsupervised extractive summarization model". IEEE Access, 9, pp.14358-14368, 2021.

[9] Zhu, H., Dong, L., Wei, F., Qin, B. and Liu, T. "Transforming wikipedia into augmented data for query-focused summarization". IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30, pp.2357-2367, 2022.

[10] Yadav, D., Katna, R., Yadav, A.K. and Morato, J., "Feature Based Automatic Text Summarization Methods: A Comprehensive State-of-the-Art Survey". IEEE Access, 10, pp.133981-134003, 2022.

[11] Mao, Q., Li, J., Peng, H., He, S., Wang, L., Philip, S.Y. and Wang, Z. "Fact-driven abstractive summarization by utilizing multi-granular multi-relational knowledge". IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30, pp.1665-1678, 2022.

[12] Alqaisi, R., Ghanem, W. and Qaroush, A. "Extractive multi-document Arabic text summarization using evolutionary multi-objective optimization with K-medoid clustering". IEEE Access, 8, pp.228206-228224, 2020.

[13] Liu, W., Gao, Y., Li, J. and Yang, Y. "A combined extractive with abstractive model for summarization". IEEE Access, 9, pp.43970-43980, 2021.

[14] Jalil, Z., Nasir, J.A. and Nasir, M. "Extractive Multi-Document Summarization: A Review of Progress in the Last Decade". IEEE Access, 9, pp.130928-130946, 2021.

[15] Saeed, M.Y., Awais, M., Talib, R. and Younas, M. "Unstructured text documents summarization with multi-stage clustering". IEEE Access, 8, pp.212838-212854, 2020.