OPEN ACCESS
International Journal of Innovations in Science & Technology
IJIST
P ISSN : 2618-1630
E ISSN : 2909-6130
Recognized Journal

What Have You Read? Based Multi-Document Summarization

Original Article

Sabina Irum[1], Jamal Abdul Nasir [2], Zakia Jalil[3]
[1] Faculty of Engineering and Computer Science National University of Modern Languages Islamabad, Pakistan
[2] Department of Computer Science Business Information Systems NUI Galway, Ireland
[3] Faculty of Basic and Applied Sciences International Islamic University, Islamabad, Pakistan
* Correspondence: Sabina Irum, Email**: sairum@numl.edu.pk**

Citation: J. Z. Irum. S, Nasir. A. J, "What Have You Read?' Based Multi-Document Summarization," Int. J. Innov. Sci. Technol., Special Issue , pp. 94-102, 2022.
Received | June 6, 2022; **Revised |** June 20, 2022; **Accepted |** June 28, 2022; **Published |** June 30, 2022.
DOI| https://doi.org/10.33411/IJIST/2022040508

Due to the tremendous amount of data available today, extracting essential information from such a large volume of data is quite tough. Particularly in the case of text documents, which need a significant amount of time from the user to read the material and extract useful information. The major problem is identifying the user's relevant documents, removing the most significant pieces of information, determining document relevancy, excluding extraneous information, reducing details, and generating a compact, consistent report. For all these issues, we proposed a novel technique that solves the problem of extracting important information from a huge amount of text data and using previously read documents to generate summaries of new documents. Our technique is more focused on extracting topics (also known as topic signatures) from the previously read documents and then selecting the sentences that are more relevant to these topics based on update summary generation. Besides this, the concept of overlapping value is used that digs out the meaningful words and word similarities. Another thing that makes our work better is the Dice Coefficient which measures the intersection of words between document sets and helps to eliminate redundancy. The summary generated is based on more diverse and highly representative sentences with an average length. Empirically, we have observed that our proposed novel technique performed better with baseline competitors on the real-world TAC2008 dataset.

Keywords: Data mining, Text mining, Text summarization, Topic Signature, Density peak, Update Summarization.

Introduction

With the recent increase in the content available in text documents, discovering relevant documents is a tedious task. There comes great difficulty in taking out the required information from large amount of data. The user's main concern is finding the most relevant text that may not have overlapping content; it must have a clear structure, precise length, and good readability. The user's major concerns are relevant papers, the most crucial pieces of information, document relevance, and removing unnecessary information. To overcome this issue, Text summarization was introduced, which identifies the most important meaningful information in a document. Before going to the text summarization, first, we should know what a summary is. Although there are numerous techniques for completing the summary process, the exponential rise of text documents makes it difficult to determine whether a text document meets the demands of a user. These methods also overlook the themes in the sentences, as well as the theme and field of study. Moreover, the previous summaries ignore the proper identification of significant contents and avoid the primary constraint considered as length. They ignore the sentences that are highly representative and novel from the summary, so eliminating redundancy becomes a crucial task. Another issue mentioned in Li et al.[1] and M. et al.[2] has mentioned that different words frequently express the same topic in a text. Moreover, Bing et al. [3] highlighted the Multi-Document Summarization by the use of Phrase Selection and Integration. Still, the abstractive summarization method is difficult to handle as it generates summaries near to human language. So, these issues should be kept in view to generate a fluent and diverse summary.

Topic signatures methodology is employed in this paper, a statistical method for distilling the subjects buried in a batch of documents. The underlying premise of this approach is that the phrases in a pre-classified corpus that represent the target notion are intimately related. In this study, we used a novel technique to assess the novelty of unigram and bigram phrases.

This section gives an overview of Novel (technique based on unigrams and bigrams) -based Text Summarization and the current study. In Section 2, we discussed existing and relevant work. Section 3 delves more into the process and its many steps. The experimentation on various data sets [4] and the findings are analyzed and discussed in Section 4. Similarly, the conclusion and future research directions were offered at the end.

Related Work

Several models are proposed to rate the most appropriate and significant sentence to generate the best summaries. Most previous methods have used clustering-based approaches where the researchers take clusters and rank them in two independent steps mentioned by [5] and Song et al. [6]. The rest of the researchers manage the process of clustering in a sharing manner, assuming that clustering mends the sentence ranking properly. Some of the methods described by Bind et al. [3] include centroid-based methods, which employ clustering algorithms to construct sentence groupings based on phrase similarity and then pick the best representative sentences from these clusters. The graph-based strategies select sentences from their neighbors' using concepts like PageRank algorithms[7]. In the previously existing methods, there are a few drawbacks i.e., extra processing is required to find out the clusters among documents, and ranking is required between and within the clusters. Zhang et al. [8] and Tong et al. [9] suggested an automatic clustering approach that can automatically find the final cluster centers and clusters. Srivastava's et al. [10] develops an action plan on Latent Dirichlet Allocation, a topic modeling technique that uses K-Medoids clustering to produce summaries. Some other methods based on density defined by X. Tao et al. [11], [12] are based on clustering techniques with local and global consistencies. Some of the authors, Sindhu [13], describe Intrinsic and extrinsic as the two different

evaluation methods for evaluating summaries. With the machine-generated summary with the human-generated summary, an intrinsic method evaluates the summary's quality and the summary's effectiveness in carrying out activities like information retrieval, question-and-answer sessions, and text classification is how extrinsic techniques gauge its quality. Moreover, the previous summaries ignore the proper identification of significant contents; they ignore high representative and novel sentences that have a proper length for the generated summary.

Methodology

In order to save time and effort, the summarization work should be done so that the constructed analysis contains the most representative lines and the information, in summary, is conveyed to the user acceptably. This information should be diverse and not redundant. To make the summary more representative and diverse, the density peak clustering approach is applied. The density-peak clustering (DPC) method already introduced by [8], [14], [11], [12], [15], [16] makes the data clusters efficiently by fast searching density peaks. This algorithm takes the sentences as input and generates cosine similarity. After generating the cosine similarity, the data is used to create a similarity matrix. By taking the sentences as input, each sentence's representativeness, diversity, and length is calculated, ranked, and selected for a summary generation.

Update Summarization Framework

Update summarization was first presented as a sub-task in 2007 at the Document Understanding Conference (DUC), and it was further refined in the year 2011 at the Text Analysis Conference. This research aims to construct two summaries for two chronologically arranged documents newswire item sets using Update summarization. The summary of the material of a new article set should not be repeated from a set of former articles and tell readers new information on the issue. Figure 1 depicts the four primary phases of the Density Peak Clustering-Topic Signature update summarization system. The following methodology is used in order to carry out the summarization process.

Text preprocessing:

Text preprocessing is the basic and most important step that includes stopping word removal, tokenization, breaking the text into sentences, counting words, and Bag of Word Transformations (subitem). Word vectors terms that appear three times or more are set aside and used to create new term-sentence matrices for the old documents set A (background) and the new documents set B (input) (the lexicon perceptions regarding in both set A and set B and Bigram terms that appear three times or more are set aside and used to create new term-sentence matrices). This step is performed to transform raw data into an understandable format and helps remove incorrect or irrelevant data from a data set.

Topic Signatures using overlap values:

Topic Signature plays an important role in text summarization as it is used to identify the important hidden concept of a text. The Topic Signatures technique is a statistical method for identifying underlying subjects in a collection of publications. The novelty of unigram and bigram sentences is determined using this approach. By changing the pre-existing hypotheses, two previously held hypotheses were modified.

Hypothesis 1: $(H_1) : P(D_{old}|uts) = p = P(D_{old}|uts)$

i.e. the uts has the same chances in Dold and Dnew.

Hypothesis 2: $(H_2) : P(D_{old}|uts) = p1P\ 2p2 = P(D_{old}|uts)$

i.e. the UTS means identical things happen in both the old and new worlds Dold and Dnew, representing that uts has high update property.

Overlap Value

All the sentences selected after applying the topic signature are passed through one additional filter in which the Overlap value calculation for each meaningful word is calculated. Following applying a topic signature, we get sentences based on the hidden topic in the document set. So, the next step is calculating the overlap value for each meaningful word in a document. The overlapping similarity between the words in one sentence is compared with those in another. The same is the case for other sentences. After that, the sum of all overlapping values is calculated and presented as the weight of a particular sentence. In the end, only those sentences selected have a high value. By performing this task, topic identification can be easily don't, increasing the efficiency and accuracy of text summarization.

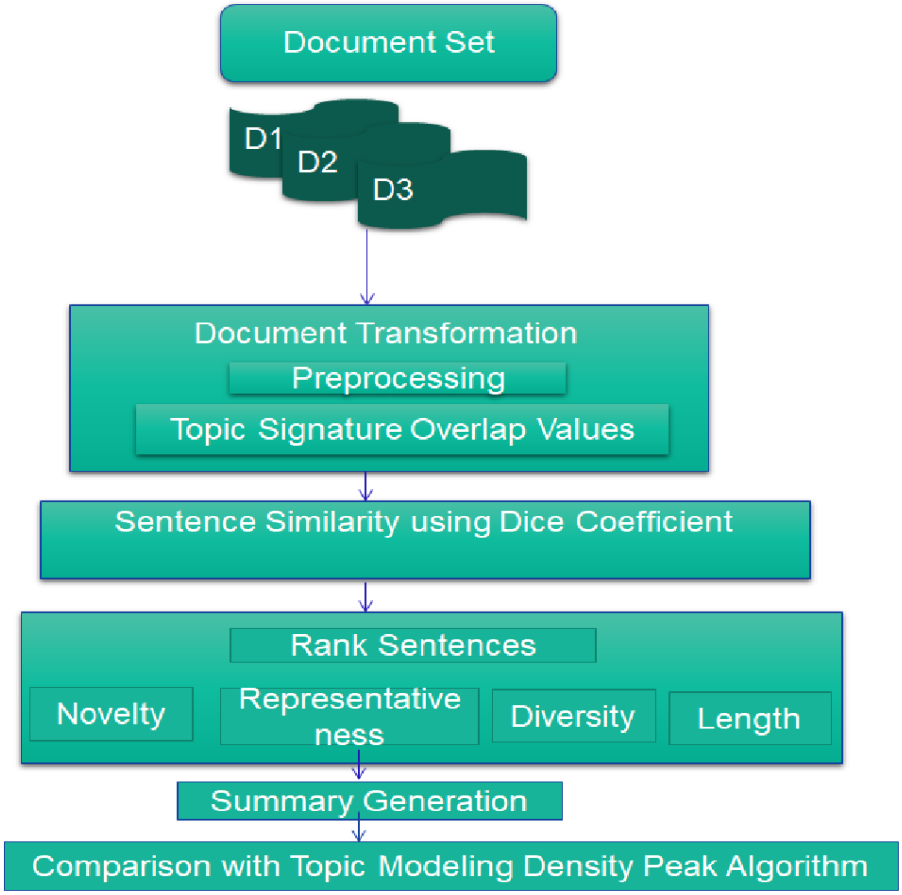

Figure 1. Explains the Flow Diagram of the Methodology

Sentence similarity using Dice Coefficient:

Sentence similarity is calculated using Dice Coefficient, which measures set intersection or existence. The dice coefficient works by checking if the intersection words are present in both A and B document sets. There is a range 0 to 1 for the value of the dice coefficient. Predefined threshold 0.65 is compared with the similarity, which shows that if the similarity value exceeds the criterion, the sentence is similar to the other selected sentences and will be eliminated from the selected sentences. Only Novel sentences are added to the summary; this way, the user can get diversified sentences in the summary. This module is very important as it focuses on the elimination of redundancy. By selecting Novel sentences, proper sentences can be selected to make a good summary in text summarization.

Rank sentences:

Sentence scoring is based on the Density Peak algorithm. Sentences present in a document are considered as objects which are preprocessed properly. After preprocessing following calculations are performed to dig out the best sentences. This module results in the proper selection of sentences as it is based on four important components: Novelty, representativeness, Diversity, and length.

a. Novelty: The candidate's novelty attribute should be quantified, and only those phrases are considered novel if they contain more novel unigram and bigram terms. In the update summarization job, the uniqueness of a sentence is based on Topic Signature values.

$$Nov_s(i) = \frac{nl(si)}{max\ \sum_{i=1}^{N} nl(sj)} X\ log\ \frac{max\ \sum_{i=1}^{N} rl(sj)}{rl(si)} \quad \text{eq.(1)}$$

b. Representativeness: Representativeness is used to select the sentences in a summary that reflect a certain piece of information. The representative score is used to judge whether a sentence is significant to the document. A particular sentence with a greater number of similar sentences is thought to be more representative.

$$Repre_s(i) = 1/N \sum_{j=0,j!=i}^{N} X(sim(i,j) - \delta) \quad \text{eq. (2)}$$

$$where\ X(x) = \begin{cases} 1\ if\ x > 0 \\ 0\ otherwise \end{cases}\ and\ sim(i,j)\ value\ s\ are\ based\ on\ SxT\ Matrix.$$

c. Diversity Score: Many previous algorithms use diversity score as a post-processing feature after the sentences have been rated [17]. A sentence diversity score is used in the documentation procedure for ranking to remove repeated and redundant phrases.

$$Div_s(i) = 1 - max_{j:Reps(j)>Reps(i)}(sim(i,j)) \quad \text{eq.(3)}$$

The following equation is used for a sentence with higher density.

$$Div_s(i) = 1 - min_{j!=i}(sim(i,j)) \quad \text{eq.(4)}$$

d. Length: Document summarization aims to extract the most significant facts from a given collection of papers; hence, length is crucial in this endeavor.

$$L_s(i) = \frac{el(si)}{max\ \sum_{i=1}^{N} el(sj)} X\ log\ \frac{max\ \sum_{i=1}^{N} rl(sj)}{rl(si)} \quad \text{eq.(5)}$$

5) Summary generation using Unified Sentence Score: The unified sentences score is used to combine the novelty, representativeness, diversity, and length that make a chance of the sentence to be considered as a summary sentence.

$$USS_s(i) = R_{S(i)} X\ D_{S(i)} X L_{S(i)}\ X\ N_{S(i)} \quad \text{eq.(6)}$$

The methodology generates a collective score, which is used to construct the summary. For the summary, only sentences with a high rank are considered.

Experiments And Their Outcomes

This section demonstrates the usefulness and proficiency of the Topic Signatures-based Density Peak Clustering method. Validation of the proposed methodology is performed to find the updated summarization task and compared with up to the mark tasks. In our method, we have used rapid miner to calculate the cosine similarity, and the rest of the work is done in Python. Java is used to calculate the Topic Signatures that are further used in python code. All the graphs are made in tableau, and figures are made using Visio. All the results are compared against baseline [16] methods to check the accuracy of the proposed system.

Data Set for Evaluation

The data set used for the proposed methodology is TAC 2008. The TAC 2008 Update Summarization assignment aims to produce concise, fluent summaries of several documents from news sources, assuming the user has already read several prior items. Each

update summary updates the reader on fresh data regarding a specific subject. To encourage research on text-based systems that perform summarization on the collection of documents, TAC 2008 is used. About 48 subjects make up the test dataset. 20 significant documents have been separated into two sets, Document Set A and Document Set B, and each issue has a topic statement (title and narrative). There are 10 documents total in each set, with Set A's documents all coming before Set B's chronologically. The papers are taken from the news item collection AQUAINT-2.

As the proposed methodology works on the update summarization task, it is assumed that the user has already accessed the text Set A, and only the summary of Set B is generated.

Metric for Evaluation

The ROUGE toolbox is employed in the TAC 2008 summary evaluation. It counts overlapping units like the n-gram between the selected i.e., system-generated summary, and the reference i.e., human-generated summary, to determine summary quality.

Results:

To measure the effectiveness of Topic Signature multi-document summarization, TAC2008 dataset is used, which is from Text Analysis Conference (TAC), for generic summarization evaluation. The dataset comprises 48 document clusters, with 10 English documents in each cluster. To evaluate each cluster, 10 summaries generated by human authors are compared against system-generated summaries. Topic Signatures Based summarization method is compared with baseline Density Peak Clustering Algorithm method to analyze content using different sets of related multiple documents. Initially, Rouge 1 score is computed, and the result shows the improvement in multi-document summarization by using Topic Signatures.

Topic signatures are based on each meaningful word's overlapping value. Despite not relying on complex algorithms or outside sources, experiments show that this strategy performs better than Baseline methods. Additionally, it heavily relies on embedded themes, overlap values, and dice coefficient values to only identify keywords and summaries that contain information from various sources, raising the level of concern and producing facts with substantial value. Only highly frequent bigrams that convey more information than unigram terms are chosen utilizing an update summary generation approach.

The following performance table shows the experimentation performed using Unified sentences score topic signature. Table 1 contains an interpretation of ROUGE 1 summary results and a comparison of the suggested approach to the Baseline data. The f-measure, the ratio of recall and precision and expressed as a percentage value, is used to express the rouge measure. As we can see, there is a difference of 0.023 percent between the proposed methodology and baseline methods. Results for ROUGE 2 0.016 percent better as compared to the baseline method. We prefer the topic signatures method because it brings out the topics from the documents that help us to generate the summary easily and precisely.

Besides the proposed methodology, we have done this work by taking cosine similarity and overlap value with topic Signature density peak clustering algorithm and got 0.0174 percent improved Rouge 1 results compared to the baseline methods. The results obtained by using cosine similarity are 0.341572 whereas, by changing cosine similarity with Overlapping values, we can get better results i.e. 0.359064. The improvement shows that the summary generated using the Density Peak Clustering Algorithm is more fluent, Less Redundant, and more accurate than the previously generated summaries. Similar is the case for Rouge 2 value, where the difference of 0.012 percent can be seen.

Table 1. Text analysis conference, 2008 data set performance comparison

Methods	Rouge 1	Rouge 2
TS-DPCA	0.359064	0.09453
Cosine Similarity	0.341572	0.08213
Baseline	0.3361	0.0782

The topic signatures-based Density Peak Clustering Algorithm, coupled with the dice coefficient and overlap value approach, outperforms the best results, or 0.023 percent, as opposed to the baseline method, according to experimental results on the dataset.

Discussions and outcomes:

The ROUGE 1 F-measure score is used to compare system-generated and reference summaries' similarities. Table 1 shows that the topic signature density peak clustering algorithm based on overlapping values and insertion of words performs 0.023 percent better than baseline results. In this experiment, we have also computed results by applying the same methodology but different similarity measures. For different similarity measures i.e., cosine similarity, 0.0054 percent improved results can be achieved, which shows that by using an overlapping value for the meaningful words in a document, we can achieve more improved summaries for the summarization task. The greater the resemblance to the reference summary, the great results will be. The following graph shows the summary vise comparison of all the summaries generated for 48 folders for both Topic Signatures and Density Peak clustering algorithm [6]. An experimental result on the dataset shows that different optimizing techniques can be used to achieve better results for summarization tasks.

In the various Rouge score versions, the Topic Signature strategy with overlap value and dice coefficient surpassed the alternative methods regarding Recall and F-Score. The density peak clustering method outperforms TextRank, LexRank, and other methods, as discussed by Zhang et al. [8]. The summaries created using the Unified Sentence Score density peak clustering algorithm performs better than those created using older techniques like text rank [18] and clustering [16]. Previous summaries missed more examples of sentences and clear topic identification. Contrarily, the summaries that are produced without overlapped values are less fluid than those that do, as figures 2 and 3.

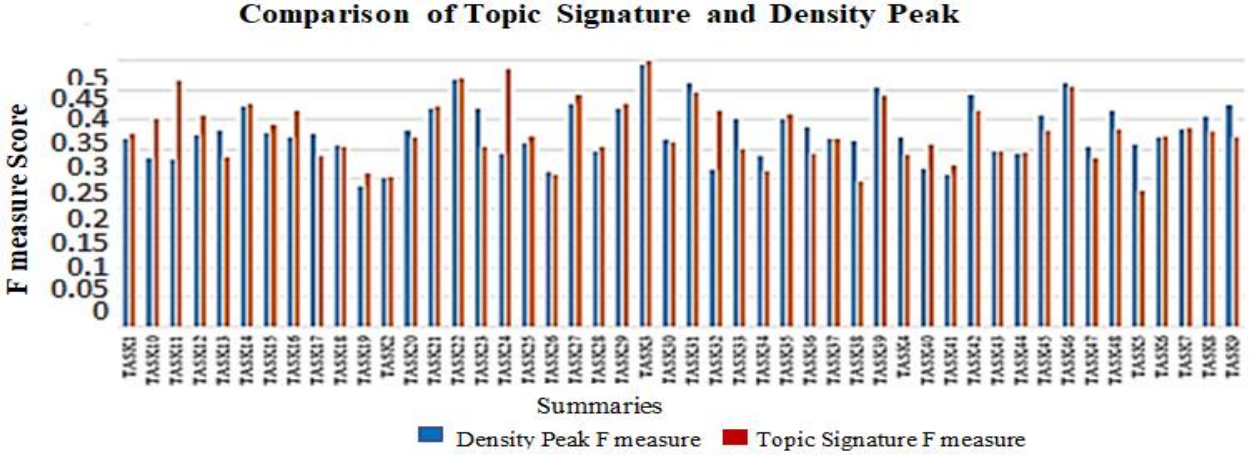

Figure. 2. Visualization of Results generated on TAC 2008 dataset

The Unified Sentence Score density peak clustering algorithms summaries perform better than those made using older techniques. Previous summaries missed more examples of sentences and clear topic identification. Contrarily, the summaries that are produced without overlapped values are less fluid than others.

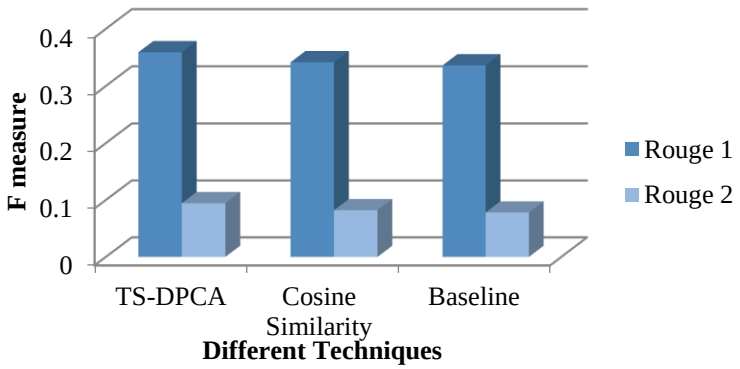

Figure 3. Comparison of Topic Signature with Density Peak Algorithm

Conclusion

To complete the revised summarization task in this research, the Density Peak Clustering technique is applied with overlapping values and the Dice Coefficient. The algorithm uses the topic signature method to assess the novel phrases and novel features in a sentence. This algorithm also focuses on the summary's diversity because the summary should contain the least possible redundancy and adequately cover the important information. The findings show that, when tested on the TAC 2008 data set, the Density peak clustering based on the Topic signatures overlap value technique outperformed the best update Multi-document summarization method. It demonstrates how update summarization duties can be effectively handled by density peak clustering with Topic signatures. However, future abstractive summarization projects could benefit from our work as the process of summarizing is writing a brief, focused summary that highlights the key points of the original material. The ability to extract condensed meanings from lengthy texts has a wide range of possible uses in real-world settings. By enhancing the quality of summaries and researching how different BERT techniques affect summarization, we hope to contribute to the advancement of this discipline in the future.

References

[1] R. Li and H. Shindo, "A hierarchical tree model for update summarization," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9022, pp. 660–665, 2015, doi: 10.1007/978-3-319-16354-3_72.

[2] K. M. Svore, L. Vanderwende, and C. J. C. Burges, "Enhancing single-document summarization by combining RankNet and third-party sources," *EMNLP-CoNLL 2007 - Proc. 2007 Jt. Conf. Empir. Methods Nat. Lang. Process. Comput. Nat. Lang. Learn.*, no. June, pp. 448–457, 2007.

[3] L. Bing, P. Li, Y. Liao, W. Lam, W. Guo, and R. J. Passonneau, "Abstractive Multi-Document Summarization via Phrase Selection and Merging," *ACL-IJCNLP 2015 - 53rd Annu. Meet. Assoc. Comput. Linguist. 7th Int. Jt. Conf. Nat. Lang. Process. Asian Fed. Nat. Lang. Process. Proc. Conf.*, vol. 1, pp. 1587–1597, Jun. 2015, doi: 10.48550/arxiv.1506.01597.

[4] R. O. and S. W. Anjum. M. S, Mumtaz. S, "Heart Attack Risk Prediction with Duke Treadmill Score with Symptoms using Data Mining," *I nternational J. Innov. Sci. Technol.*, vol. 3, no. 4, pp. 174–185, 2021.

[5] C. Li, Y. Liu, and L. Zhao, "Improving update summarization via supervised ILP and sentence reranking," *NAACL HLT 2015 - 2015 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Proc. Conf.*, no. August 2016, pp. 1317–1322, 2015, doi: 10.3115/v1/n15-1145.

[6] Y. Song, W. Ng, K. W. T. Leung, and Q. Fang, "SFP-Rank: significant frequent pattern analysis for effective ranking," *Knowl. Inf. Syst.*, vol. 43, no. 3, pp. 529–553, 2015, doi: 10.1007/s10115-014-0738-y.

[7] Q. M. A. and M. . Kiran. I, Siddique. Z, Butt. A. R, Mudassir. A. I, "Towards Skin Cancer Classification Using Machine Learning and Deep Learning Algorithms: A Comparison," *I nternational J. Innov. Sci. Technol.*, vol. 3, no. special issue, pp. 110–118, 2021.

[8] Y. Zhang, Y. Xia, Y. Liu, and W. Wang, "Clustering sentences with density peaks for multi-document summarization," *NAACL HLT 2015 - 2015 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Proc. Conf.*, no. January, pp. 1262–1267, 2015, doi: 10.3115/v1/n15-1136.

[9] W. Tong, S. Liu, and X. Z. Gao, "A density-peak-based clustering algorithm of automatically determining the number of clusters," *Neurocomputing*, vol. 458, pp. 655–666, Oct. 2021, doi: 10.1016/J.NEUCOM.2020.03.125.

[10] R. Srivastava, P. Singh, K. P. S. Rana, and V. Kumar, "A topic modeled unsupervised approach to single document extractive text summarization," *Knowledge-Based Syst.*, vol. 246, Jun. 2022, doi: 10.1016/J.KNOSYS.2022.108636.

[11] X. Tao, R. Wang, R. Chang, C. Li, R. Liu, and J. Zou, "Spectral clustering algorithm using density-sensitive distance measure with global and local consistencies," *Knowledge-Based Syst.*, vol. 170, pp. 26–42, Apr. 2019, doi: 10.1016/J.KNOSYS.2019.01.026.

[12] X. Tao *et al.*, "Density peak clustering using global and local consistency adjustable manifold distance," *Inf. Sci. (Ny).*, vol. 577, pp. 769–804, Oct. 2021, doi: 10.1016/J.INS.2021.08.036.

[13] K. Sindhu and K. Seshadri, "Text Summarization: A Technical Overview and Research Perspectives," *Handb. Intell. Comput. Optim. Sustain. Dev.*, pp. 261–286, Feb. 2022, doi: 10.1002/9781119792642.CH13.

[14] D. Cheng, J. Huang, S. Zhang, X. Zhang, and X. Luo, "A Novel Approximate Spectral Clustering Algorithm with Dense Cores and Density Peaks," *IEEE Trans. Syst. Man, Cybern. Syst.*, vol. 52, no. 4, pp. 2348–2360, Apr. 2022, doi: 10.1109/TSMC.2021.3049490.

[15] Z. Liang and P. Chen, "An automatic clustering algorithm based on the density-peak framework and Chameleon method," *Pattern Recognit. Lett.*, vol. 150, pp. 40–48, Oct. 2021, doi: 10.1016/J.PATREC.2021.06.017.

[16] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science (80-.).*, vol. 344, no. 6191, pp. 1492–1496, Jun. 2014, doi: 10.1126/SCIENCE.1242072/SUPPL_FILE/RODRIGUEZ.SM.PDF.

[17] N. R. Mabroukeh and C. I. Ezeife, "A taxonomy of sequential pattern mining algorithms," *ACM Comput. Surv.*, vol. 43, no. 1, Nov. 2010, doi: 10.1145/1824795.1824798.

[18] C. Mallick, A. K. Das, M. Dutta, A. K. Das, and A. Sarkar, *Graph-based text summarization using modified TextRank*, vol. 758, no. August. Springer Singapore, 2018. doi: 10.1007/978-981-13-0514-6_14.