EXTRACTIVE TEXT SUMMARIZATION OF ONLINE SCIENTIFIC ARTICLES FOR DIGITAL LIBRARY REPOSITORY

R.D.R.M Gunathilaka, MAC Akmal Jahan

Department of Computer Science, Faculty of Applied Sciences
South Eastern University of Sri Lanka, Sri Lanka

ABSTRACT*: In a scientific community including researchers and students who are dedicated in reading, experimenting and writing ideas to the research world, they must refer to a significant number of articles on a daily basis. Digital libraries play a key role in supplementing scientific articles within online platforms. Due to the much abundance quantities of such articles presented in various platforms, the searching process tends toward time consuming, and identifying much related resources among them becomes again difficult. On the other hand, the majority of the scientific articles are available on a subscription-based and the online archives show only a document abstract but users necessitate extra material to determine if the article is extremely relevant or not, even if it merely provides a quick description. Therefore, this work aims to introduce an alternative approach to simplify the searching and sorting of the scientific articles for a digital library where a short subunit of sentences of the subscribed articles will be provided before purchasing it. To find a best algorithm for the process of summarisation of scientific articles within a short time and provide a good comprehension of the scientific document are to be investigated. Summaries from the publicly available SumPubMed dataset of scientific articles are evaluated using supervised and unsupervised approaches and manual summaries from them are compared. Text Rank algorithm performed better than the TF-IDF and K-Means algorithms, and the system achieved a better result when increasing the content size of the article.*

Keywords: K-Means, ROUGE, extractive summarization, Text-Rank, TF-IDF.

1. INTRODUCTION

Summarization in the sense of corpus is retrieving the significant facts from any format of documents and presenting them in such a way that people can grasp as much information as possible within a short time. Mainly, there exist two types of summarization techniques such as extractive and abstractive approaches (Allahyari, 2017). The former summarization chooses highest important data from a document in which few sentences are grouped together as a subunit of sentences in the given articles. The later summarization approach produces a summary by taking the main idea of the document/s and generates new sentences. In this work, the process of extractive text summarization is focused for scientific articles.

When we focus on the scientific community, they need to refer a significant number of scientific articles on daily basis. Every year, a large number of research and studies are conducted across the research globe. Due to the much abundance quantities of such articles presented in various platforms, the process of intelligent searching tends toward time consuming, and therefore, identifying much related resources among them becomes again a trivial task. The bulk of the top scientific articles is only available through membership or subscription basis. The majority of websites and other online archives give a document abstract during user search. However, users actually require additional materials or content related to the specific article to determine if the particular article is extremely relevant or not before purchasing it, even if it merely provides a quick description. The motivation of this work is if there are online digital archives in university libraries and other educational institutes, that should provide a text summarisation outline for subscripted and indexed scientific articles before getting

subscription by a researcher or institute. Therefore, this will help to minimise time spent for a search, find more relevant articles and ease the process of paying for the particular article without wasting money.

Therefore, in this work we have proposed an idea which focuses a centralized platform, a part of a digital library archival that can be accessed globally. Any digital library can connect through this platform with least amount of subscription fee instead of paying institutional full yearly subscription. When the user provides the title of the search area, this platform can provide the relevant articles with an extractive summary, that covers the summary of both the abstract, introduction and body of the complete document. Then the user can comprehensively understand its higher frequency of relevancy before purchase and download it. Therefore, this work tends toward to find a best algorithm for the process of summarisation of scientific articles within a short time and provide a good comprehension of the scientific document.

In this study, two objectives are experimentally investigated: i) how text summarisation algorithms perform with scientific articles, which leads to select a better one with a reasonable time complexity; ii) how content size of the complete article influences the summarisation performance. For the first objective, state-of-the-art word processing algorithms which possible to be used in the text summarization process are investigated with the articles. We have experimented Text Rank, TF-IDF, and K-Means algorithms in independent platform. For the second objective, a different dataset with different content types is investigated.

2. LITERATURE REVIEW

Extractive (highlighting) and abstractive (paraphrasing) summarization are two types of summarizations. The former chooses the most important data from the document while the later produces a summary by taking key ideas of the document/s and generating new sentences (Allahyari et al., 2017). For the summarization, three steps have been followed: i) construct intermediate sentences from the input; ii) process sentence scores and iii) generate a document summary.

In the literature, Rahimi S. R. et al., (2017) presents a connection between text summarization and text mining. The latter concept discovers hidden patterns and extracts new information from a document by connecting words and sentences. Therefore, summarisation of text in a document is a subset of text mining. Different categories of summary systems are explored in the past. They are: i) extractive and abstractive summary based on the output; ii) indicative and informative summary based on the details; iii) generic and query-based summary based on the content prior knowledge; iv) single and multi-document summary based on the number of input texts and v) mono and multi-language summary based on the language.

There are various strategies carried out for the text summarization in the Natural Language Processing. From the overview of existing algorithms, manual and automatic evaluation methods such as supervised and unsupervised methods have been used for extractive summarization. Mihalcea, R. (2004) presented an unsupervised graph-based ranking algorithm where the process of hyperlinked induced topic search is followed by positional power function and PageRank algorithms. K-Nearest Neighbour algorithm was also used for text summarization based on feature similarity (Jo, T., 2017) where KNN version was modified

by defining a similarity that considers both feature value similarity and feature similarity.

In the sense of single and multi-document summarisation, Sharaff, A. et al., (2022) presented K-means based cluster ranking approach where features from single document were used to compute similarity scores obtained from the sentences. Highly ranked sentences from both clusters were ranked for the final summary. On the other hand, in the sense of multi document summarisation, Pasunuru, R. et al., (2021) presented two Query-focused Multi-Document Summarization (QMDS) training datasets. In 2022, Mishra, S. et al., (2022) presented a multi-objective clustering framework for summarisation of scientific document where abstract and citation contextualisation approaches were used. The process of citation contextualisation extracts all sentences from reference list that is mentioned in that particular article, and the important sentences are clustered using multi-objective clustering.

In the sense of text summarization in abstractive form, Wei Li et al., (2018) proposed an approach to extend the basic neural encoding-decoding framework with an information selection layer. Liwei Hou et al., (2017) introduced an approach to Chinese words using Neural Model with Joint Attention to address the problem of the attention encoder-decoder models that has shortcomings to generate repeated words or phrases. Jingyi You et al., (2022) addressed a problem with regards to neural seq2seq models and BERT that they tend to get unimportant phrases ignoring the important ones.

In 2022, Arabic text abstract summarization system is proposed by Y.M. Wazery et al., (2022) using a sequence-to-sequence model where an encoder and a decoder are functioned. In 2018, Wei Li et al., (2018) presented a method to enhance the document summarization performance by capturing the long-term structural information. This captures the structural properties of summarization such as information compression and information coverage.

On the other hand, the researchers experimented with a combination of both extractive and abstractive techniques. The work in (Yang Liu and Mirella Lapata, 2019) have reported a framework for both extractive and abstractive models and explained how the BERT can be used in the process. The work in (Meena et al., 2020) proposed a new approach, namely Text Frequency Ranking Sentence Prediction (TFRSP) which used both supervised (Sequence-to-Sequence model) and unsupervised learning algorithms.

From the analysis of the literature, comparatively a very few works have been done using a combination of supervised and unsupervised method. The motivation in this work is to analyse in what extent the both approaches can contribute independently and in a combined nature. Therefore, we have used a combination of the TF-IDF and Text-Rank algorithms and evaluated them individually to test which performs better for our task. Further, an unsupervised approach using K-Means algorithm is also evaluated with the dataset to find the best suitable algorithm for text summarization of online scientific documents in a digital library archival.

3. METHODOLOGY

3.1 Dataset

Since this work requires two different experimental modules, we have used two different

datasets where the first one was generated from the publicly available SumPubMed dataset (Gupta et al., 2021) and the second one was a user generated dataset where the text contents are acquired from online scientific articles by extracting the "Topic, Abstract and Introduction sections". We have tested the second dataset by changing the content in two directions: i) by providing Topic and Introduction sections; ii) Topic, Abstract and Introduction sections to evaluate if including the introduction along with the given abstract of the document, is a better way or not. All these datasets have their standard reference summaries for comparison.

For the input of the process, we have used raw text files from SumPubMed and the second datasets which were treated with three text summarization algorithms separately, and the output is the machine-generated summarised content. The Recall-Oriented Understudy of Gisting Evaluation (ROUGE) (Lin, C. Y, 2004) scores were used to evaluate the performance of the automated summary with the standard reference summary.

3.2 Overall Process

The overall processes of this work are outlined as follows;

A. Datasets generation
B. Automated summary generation using the given algorithms
C. Evaluate the automated summary with the reference summary and compare the algorithms and the content size.

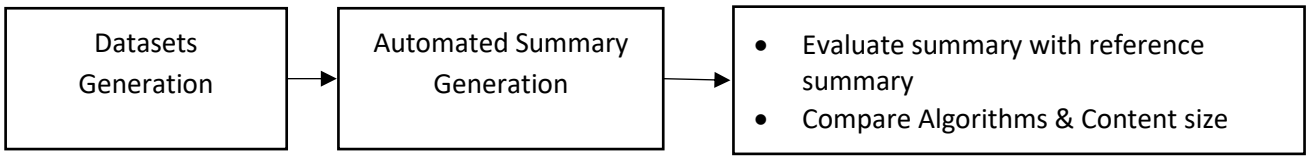


Figure 1. Overall Process.

Dataset Generation

This study has focused on scientific articles where we have generated two datasets. One is a publicly available dataset called SumPubMed which was created in 2021 (Gupta et al., 2021) where raw text files along with reference summaries are included. In the raw text files consists of background, results, and conclusions from the scientific medical articles.

The second dataset is a user generated one processed by ourselves using online scientific research articles. Here we have extracted the abstract and introduction sections along with the topics of the scientific papers. The reason for this second data set is to evaluate if we can get better results when we increase the content size of the articles, particularly to check if there is an impact of the abstract in addition to the introduction section. In this dataset, the reference summaries are generated manually (human-generated summary) for comparison.

Automated Summary Generation

The process of the summary generation of this work is outlined as shown in Figure 2.

Evaluation

There are a set of matrices for automatic evaluation, and we have selected a most popular

evaluation metric called ROUGE. This is a set of matrices that we can use to evaluate summaries. This was firstly introduced by a researcher named Chin-Yew Lin, (2004) from the Information Science Institute at the University of Southern California in 2004. It contains metrics for determining the quality of a particular summary by comparing it to another. These measurements compute how many n-grams overlapping, word pairings and sequences between the ideal reference and machine-generated summaries are found.

ROUGE-1 = Overlapping of single words (unigrams) between the reference summary and the machine-generated summary.
ROUGE-2 = Overlapping of pair of words (bigrams) between the reference summary and the machine-generated summary.
ROUGE-L = Overlapping of longest common sequences of words (LCS) between the reference summary and the machine-generated summary.

In each case, we have received three measures namely Recall, Precision, and F Measure. We can get the Recall value using the below formula. It ensures that our approach is capturing all the information included in the reference summary. It captures as many words as possible.

$$\text{Recall} = \frac{\text{Number of words overlapping}}{\text{Total word in the reference summary}}$$

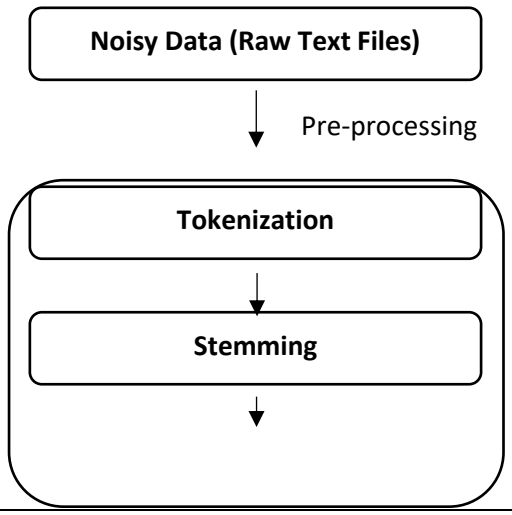
The below formula shows how to calculate precision which is used to avoid outputting irrelevant words.

$$\text{Precision} = \frac{\text{Number of words overlapping}}{\text{Total word in the generated summary}}$$

After that, we have used these two measures to calculate F-Measure as given below. F-Measure is used for the comparison.

$$\text{F-Measure} = 2x \frac{\text{Precision x Recall}}{\text{Precision+Recall}}$$

After calculating ROUGE values from two experiments, the F-Measures have been used to compare the results. Two comparisons have been performed and the first comparison is between the selected three algorithms using both datasets while the second comparison is performed by changing the content size of the text files using the second generated dataset.



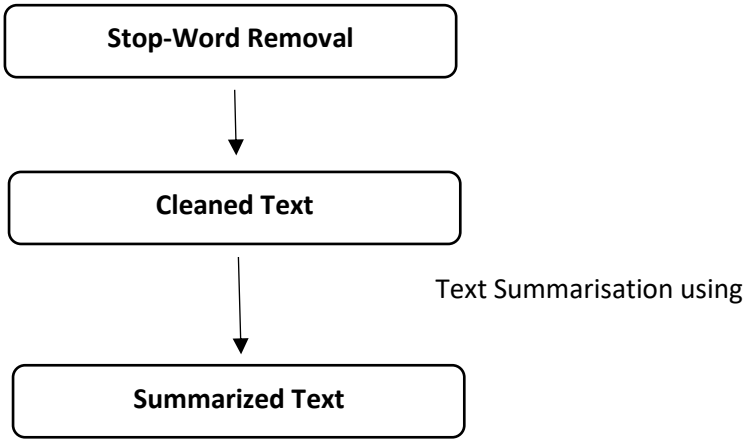


Figure 2. Overall Process of Summary Generation.

4. EXPERIMENTS AND RESULTS

In this work, we have performed two experiments using two separate datasets SumPubMed and the second-generated dataset acquired from online scientific articles. Experiment 1 targets to select a better performing algorithm among the given three text summarisation algorithms while experiment 2 focuses on how content size or type of the scientific article influences the text summarisation performance, particularly the abstract section.

4.1 Experiment 01

The first experiment has been performed using the SumPubMed dataset. Raw text files from SumPubMed dataset were fed into the given three algorithms separately. The automated summary has been compared with the given reference summary. The performance metrics of Recall, Precision, and F Measure for ROUGE-1, ROUGE-2, and ROUGE-L have been obtained. The resulted performance metrics are tabulated in Table 1 and illustrated in Figure 3.

A. ROUGE Scores for Text Rank, TF-IDF and for K-Means Algorithms

Table 1: ROUGE Scores for Text Rank, TF-IDF and for K-Means Algorithms.

	Text Rank Algorithm			**TF-IDF Algorithm**			**K-Means Algorithm**		
	ROUGE-1	**ROUGE-2**	**ROUGE-L**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
Recall	0.600848	0.271301	0.537622	0.435423	0.135076	0.382954	0.558894	0.237087	0.497151
Precision	0.223274	0.076683	0.199196	0.190643	0.050361	0.16771	0.218585	0.072571	0.193891
F-measure	0.319437	0.116757	0.285262	0.258761	0.071198	0.227625	0.306986	0.107823	0.272534

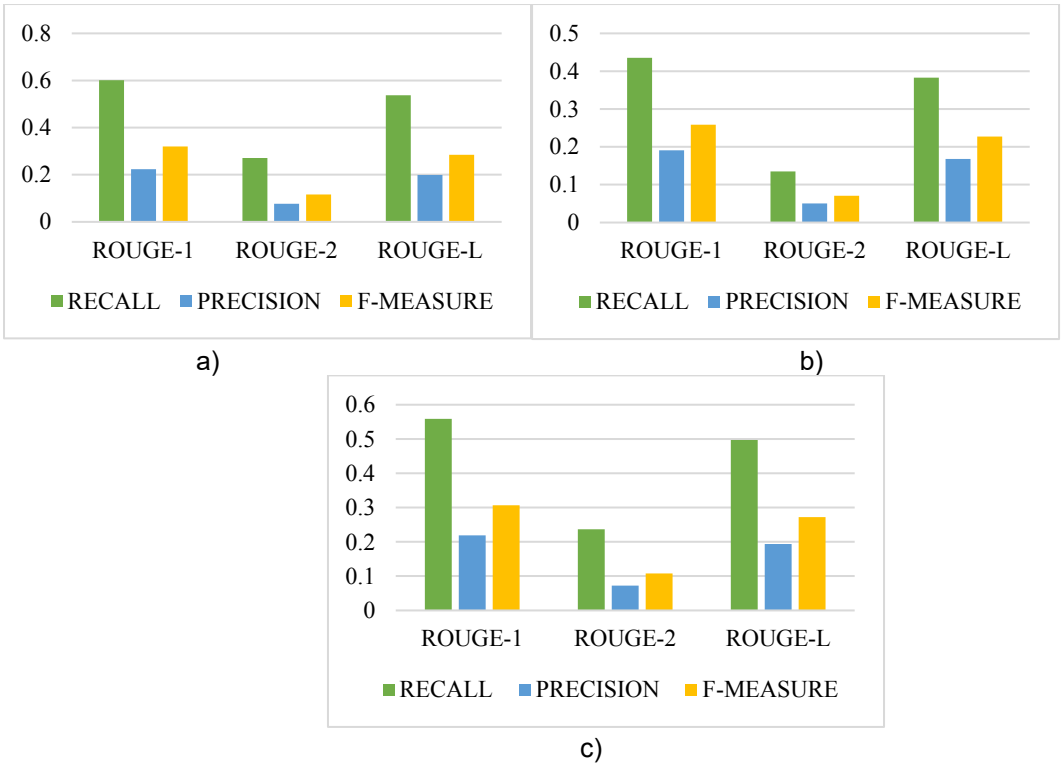


Figure 3. Rouge values of a) Text Rank; b) TF-IDF; c) K-Means algorithms.

Performance comparison of the three algorithms

To compare the results of the above three algorithms, we use the F-Measures for unigrams, bigrams, and LCS in all three cases. The below Table 2 indicates the F measures. It is observed that unigram performs better than the others.

Table 2. F-Measures using unigrams, bigram and LCS.

Algorithm	Unigrams	Bigrams	LCS
Text Rank	0.319437	0.116757	0.285262
TF-IDF	0.258761	0.071198	0.227625
K-Means	0.306986	0.107823	0.272534

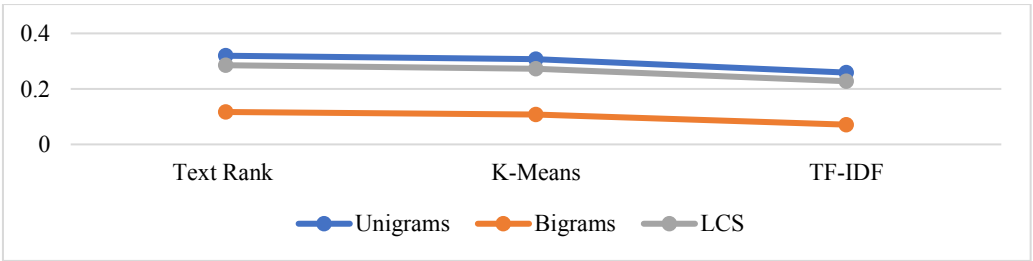


Figure 4. Overall F-Measure comparison of unigram, bigram and LCS.

4.2 Experiment 02

The second experiment has been performed using the second generated dataset, which consists of scientific articles acquired from several online platforms. The raw text files have been divided into two sets according to the content size or type. Set 01 consists of 'Topic and

Introduction' sections of the scientific papers while Set 02 consists of 'Topic, 'Abstract' and 'Introduction' sections of the scientific documents. These two sets of documents have been fed into the given algorithms. The automated summary has been treated with the reference human-generated summary, and received the Recall, Precision, and F Measure for ROUGE-1, ROUGE-2, and ROUGE-L as given in the Table 3. Figure 5 shows the graphical representation of the performance.

A. Performance of the algorithms using the content of 'Topic' and 'Introduction' sections of the articles.

Table 3. ROUGE values for the Text Rank, TF-IDF and K-Means algorithms using Introduction and Topic Sections of the articles.

	Text Rank Algorithm			TF-IDF Algorithm			K-Means Algorithm		
	ROUGE-1	**ROUGE-2**	**ROUGE-L**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
Recall	0.775992	0.673471	0.766842	0.639366	0.496895	0.621564	0.74377	0.613765	0.728703
Precision	0.500071	0.405698	0.494503	0.44351	0.330099	0.431507	0.456493	0.358826	0.447398
F-measure	0.600095	0.497627	0.593223	0.516381	0.389373	0.502171	0.559002	0.446064	0.547821

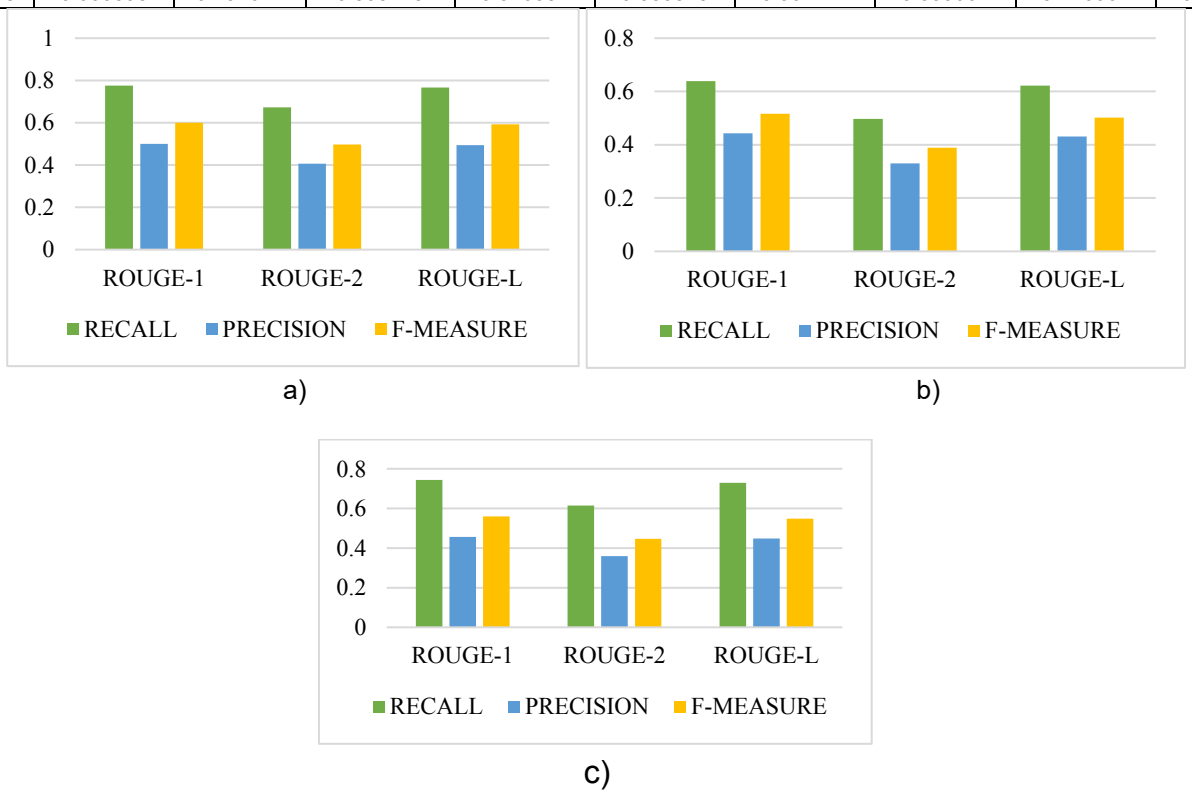


Figure 5. ROUGE values of a) Text Rank; b) TF-IDF and c) K-Means algorithms using Introduction and Topic Sections of the articles.

Performance of the algorithms using the content of 'Topic', 'Abstract' and 'Introduction' sections of the articles.

Table 4. ROUGE values for the Text Rank, TF-IDF and K-Means algorithms using Abstract, Introduction and Topic Section of the articles.

	Text Rank Algorithm			TF-IDF Algorithm			K-Means Algorithm		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
Recall	0.757143	0.65696	0.747816	0.666948	0.5216	0.651848	0.767897	0.643975	0.756479
Precision	0.53428	0.440486	0.527863	0.503193	0.386523	0.491945	0.511854	0.414286	0.504215
F-measure	0.621097	0.521516	0.613539	0.567954	0.438136	0.555164	0.609401	0.499183	0.600315

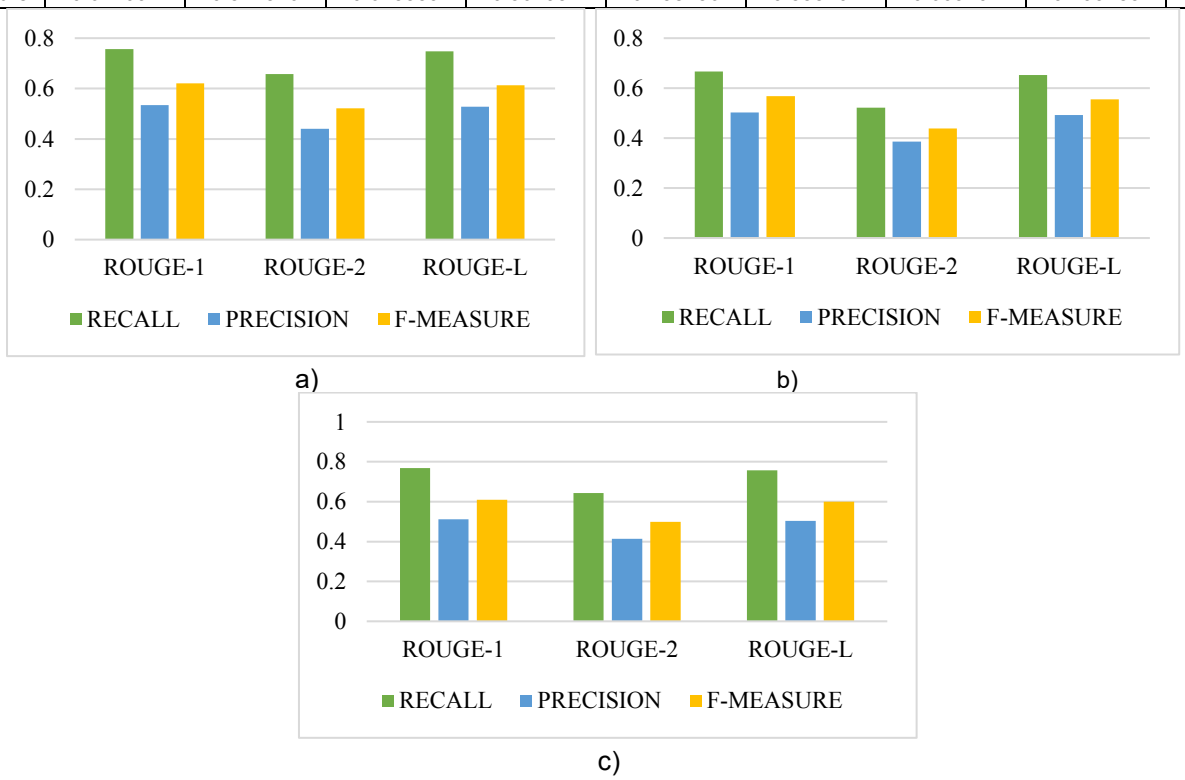


Figure 6. ROUGE values of Text Rank, TF-IDF and K-Means algorithms using Abstract, Introduction and Topic Sections of the articles.

Performance comparison of the three algorithms

To compare the results of the above three algorithms, we have used the values of F- Measures for unigrams, bigrams, and LCS in all three cases same as in Experiment 01. It is observed that unigram performs better than the others as shown in Figure 7.

Table 5. F-Measures using unigrams, bigrams and LCS.

Algorithm	Unigrams	Bigrams	LSC
Text Rank	0.610596	0.509572	0.603381
TF-IDF	0.542168	0.413755	0.528668
K-Means	0.584202	0.472624	0.574068

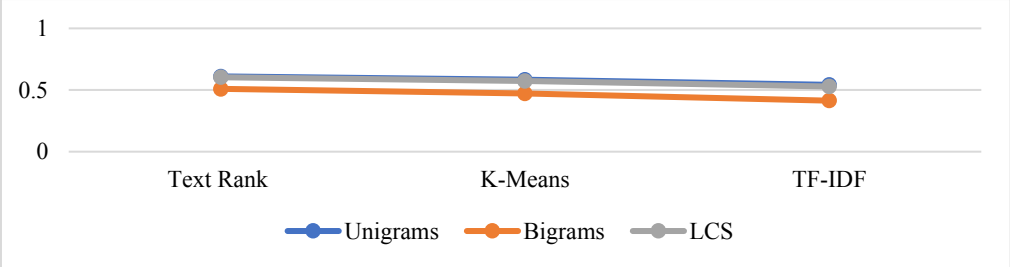


Figure 7. Overall F-Measure comparison of unigram, bigrams and LCS.

Performance comparison of content size

In this experiment, we have considered F Measures for unigrams, bigrams, and LCS in all three cases to compare the content size by altering the contents in the text files. It focuses to compare whether adding more content would be good or not, particularly including abstract of the article can give more precise results. The given Table 6 indicates the F measures of the three algorithms, and it is observed that adding the content of abstract can yield a positive impact on the performance as shown in Figure 8.

Table 6. F-Measures for different content sizes of the articles.

Content Size	**Text-Rank Algorithm**	**TF-IDF Algorithm**	**K-Means Algorithm**
Topic+Introduction	0.563648333	0.469308	0.517629
Topic+Abstract+Introduction	0.585384	0.520418	0.569633

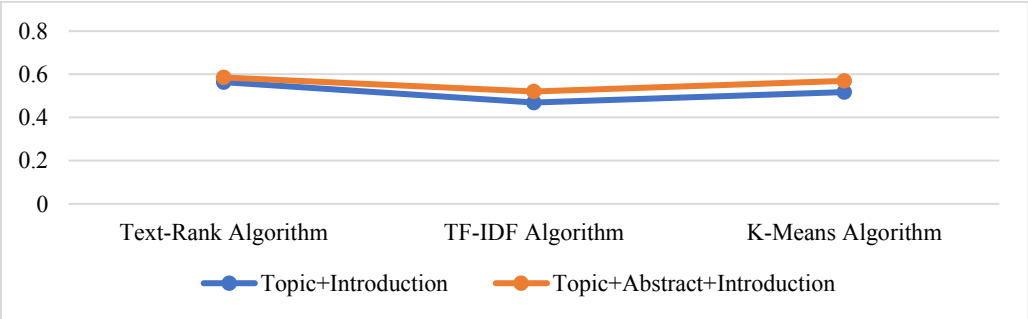


Figure 8. Overall F-Measure Comparison with the content size of the articles.

Time Comparison of the algorithms

Table 7 shows the execution time of the three algorithms where Text-Rank algorithm comparatively consumes a larger amount of time than the other two algorithms.

Table 7. Time for execution of the algorithms in seconds.

Algorithm	**Time (S)**
Text Rank	452.4629
TF-IDF	4.338225
K-Means	0.891249

5. Conclusion

In this study, we have aimed to investigate two findings in text summarisation for the scientific articles. The first one is to find a better or suitable algorithm for generating summaries from scientific articles. And the second approach is to find whether the increment of the content size particularly, including the content of abstract of the article in addition to the introduction can yield a positive impact when implement an online text summarisation platform for scientific articles. In the first case, we have performed to compare three state-of-the-art text summarisation algorithms. From the evaluation results of Text Rank, TF-IDF, and K-Means algorithms, we can observe that, the Text Rank Algorithm outperforms the others using both datasets with the highest scores. However, when we consider the time complexity of the algorithms, it is not efficient compared with the other two because it takes comparatively much larger amount of time for the execution. Since the proposed online text summarisation platform use a large number of research articles simultaneously, then the processing time of the Text-Rank algorithm will increase. In the second case, increasing the content size or adding the abstract of the article (a summary of the introduction) can raise the performance. Finally, we can conclude that it is a better way to include more details in the summary rather than just giving the abstract of the document. Further, Text Rank Algorithm performs much more accurately than the other two algorithms. However, it needs to be accelerated since it takes large amount of time to generate the summary. Therefore, in that case we can use software level parallelism and GPUs in future.

REFERENCES

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). Text summarization techniques: a brief survey. *International Journal of Advanced Computer Science and Applications (IJACSA)*, Volume 8 Issue 10, 2017.

Meena, S. M., Ramkumar, M. P., Asmitha, R. E., & G SR, E. S. (2020, September). Text summarization using text frequency ranking sentence prediction. *4th IEEE International Conference on Computer, Communication and Signal Processing (ICCCSP)*, 1-5.

Gupta, V., Bharti, P., Nokhiz, P., & Karnick, H. (2021, August). SUMPUBMED: Summarization Dataset of PubMed Scientific Articles. *In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, 292-303.

Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In Text summarization branches out, 74-81.

Rahimi, S. R., Mozhdehi, A. T., & Abdolahi, M. (2017, December). An overview on extractive text summarization. *In 2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI)*, 0054-0062.

Jo, T. (2017, January). K nearest neighbor for text summarization using feature similarity. *In 2017 IEEE International Conference on Communication, Control, Computing and Electronics Engineering (ICCCCEE)*, 1-5.

Mihalcea, R. (2004, July). Graph-based ranking algorithms for sentence extraction, applied to text summarization. *In Proceedings of the ACL interactive poster and demonstration sessions*, 170-173.

Li, W., Xiao, X., Lyu, Y., & Wang, Y. (2018). Improving neural abstractive document summarization with explicit information selection modeling. *In Proceedings of the 2018*

conference on empirical methods in natural language processing, 1787-1796.
Hou, L., Hu, P., & Bei, C. (2017, November). Abstractive document summarization via neural model with joint attention. *In National CCF Conference on Natural Language Processing and Chinese Computing*, pp. 329-338.
Li, W., Xiao, X., Lyu, Y., & Wang, Y. (2018). Improving neural abstractive document summarization with structural regularization. *In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4078-4087.
Yang Liu and Mirella Lapata. 2019. Text Summarization with Pretrained Encoders. *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740.
You, J., Hu, C., Kamigaito, H., Takamura, H., & Okumura, M. (2021, September). Abstractive Document Summarization with Word Embedding Reconstruction. *In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, 1586-1596.
Pasunuru, R., Celikyilmaz, A., Galley, M., Xiong, C., Zhang, Y., Bansal, M., & Gao, J. (2021, March). Data augmentation for abstractive query-focused multi-document summarization. *In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2021)*, pp. 13666-13674.
Sharaff, A., Jain, M., & Modugula, G. (2022). Feature based cluster ranking approach for single document summarization. *International Journal of Information Technology*, 1-9.
Mishra, S. K., Saini, N., Saha, S., & Bhattacharyya, P. (2022). Scientific document summarization in multi-objective clustering framework. *Applied Intelligence*, 52(2), 1520-1543.
Wazery, Y. M., Saleh, M. E., Alharbi, A., & Ali, A. A. (2022). Abstractive Arabic Text Summarization Based on Deep Learning. *Computational Intelligence and Neuroscience*, 2022.