Text Summarization and Translation of Summarized Outcome in French

T. Vetriselvi[1], *Mihir* Mathur[2]

[1]Assistant Professor, Dept. of IOT, SCOPE, Vellore Institute of Technology, Vellore, India, vetriselvi.t@vit.ac.in
[2]Dept. of CSE with Data Science, SCOPE, Vellore Institute of Technology, Vellore, India, mihir.mathur2020@vitstudent.ac.in

Abstract. Automatic text summarization is increasingly required with the exponential growth of unstructured text through increasing internet and social media usage across the globe. The various approaches are outcomes of extraction-based and abstraction-based. In Extraction-based summarization, the extracted content from the original data, is typically presented in the same or slightly modified form without significant paraphrasing or restructuring. Abstractive methods involve building an internal representation of the original content and then using that representation to generate a summary that may not be present in the original text verbatim. They employ natural language processing techniques, such as language generation models, to paraphrase and rephrase sections of the source document to create a more human-like summary. Abstraction in abstractive summarization is indeed a challenging task because it requires not only linguistic and syntactic understanding but also a deeper semantic understanding of the original content. The evaluation of automatically summarized text through criteria of precision and recall, and various evaluation methods, datasets used for domain based and generic text summarization. This paper proposes the Sentence Length Impact (SLI) algorithm to summarize English text content, which gives 92% accuracy and translating the same in French.

Keywords: Automatic text summarization, automatic text translation, Neural Network, Machine Learning, ROGUE evaluation, POS tagging, Natural Language Processing, Sentence Length Impact

1 Introduction

The vast amount of information available on the internet and social media platforms has made it more challenging to extract useful and relevant information for effective understanding and use by human beings. Also, the abundance of information on online forums and social networks can make it difficult and time-consuming for users to read through numerous articles that contain redundant and repetitive information [1],[2]. In such scenarios, an automated summarization and translation system can be immensely helpful in

identifying and presenting the most important and critical information in this age of information explosion, as well as making this possible to be understood by the multi-lingual global population.

This paper describes various ways in which automatic text summarization has been classified, based on ongoing developments [2], [3], [4], [5], like extractive and abstractive approaches are commonly used in both single-document and multi-document summarization tasks. [6], [7]; considering that abstractive summarization is considerably harder. Some of the most used methods, such as topic representation approaches, frequency-driven methods, graph-based and machine learning techniques for text summarization have been summarized with their advantages and disadvantages. It is also difficult to evaluate the outcomes of such summarization, and various evaluation methods are described like ROGUE [8].

Extractive summarization techniques like Intermediate Representation, Summary Sentences Selection are mentioned along with Topic Representation approaches like Topic words, Frequency driven approaches [word probability and TFIDF Term Frequency Inverse Document Frequency, SumBasic system based on greedy strategy],) along with unsupervised Extraction methods like Latent Semantic Analysis (LSA), and Probabilistic Topic models like Latent Dirichlet Allocation (LDA) model of unsupervised extraction of thematic topics and Bayesian Topic models including Bayesum, multi-document summarization by Celikyilmaz [3], [6], [7], [9-13] , Indicator representation approaches like Graph- based methods, Machine Learning Summarization through classifiers like naïve-Bayes, decision trees, support vector machines, Hidden Markov models and Conditional Random Fields are briefly mentioned [4].

There are specific user-based applications also, like summarization of research papers, medical journals, economic trends, court-case proceedings etc. which are used by scientists, doctors, economists, and law professionals [5].

Text translation involves conveying the meaning of a source-language text[English Text Summarization Outcome] into a target-language text[example, English to French]. With the global internet economy, remote collaboration and multi-lingual population which travels internationally, the various applications of automatic text translation are increasing every year [14-16].

There are many methods which have not been used often in text summarization. Some of these are include Restricted Boltzmann Machine (RBM), Analytical Hierarchy Process (AHP), Abstract Meaning Representation (AMR), abstractive summarization (AS), CLA, Recurrent Neural Network (RNN), Sentiment Memory (SM), Hierarchies Agglomerative Clustering (HAC), Multi- Objective Artificial Bee Colony (MOABC), Lowest Common Sub-summer (LCS), Non-Negative Matrix Factorization (NMF), rule-based, N-rank, Lex-rank, Text-rank, Narrative Abstractive Summarization (NATSUM), Rank-Biased Precision Summarization (RBP-Sum), Decay Topic Model (DTM). One of these methods is a new method, namely NATSUM. NATSUM is an abstractive summarization method that uses a machine learning approach. The results of the NATSUM evaluation excel in ROUGE, grammatical, non-redundancy, and coherence evaluations.

Chapter 2 presents a literature survey of some papers to gain insights about their scope and various text summarization and translation methods. Chapter 3 illustrates the proposed architecture of Text Summarization and Translation in French. Following it is Chapter 4,

which uses evaluation metrics like precision, recall, f-score, and compression ratio, retention ratio to determine the proficiency of the summarizer and French translator and shows the corresponding successful implementation. In the end chapter 5, presents conclusion and future scope. The proposed model of Sentence Length Impact (SLI) algorithm to summarize English text content gave the highest accurate results of 92%.

2 Literature Survey

In the first paper [1], some of the key approaches to automatic text summarization are described, considering explosion of text data available from multiple sources, and conclude by highlighting need of continued research in this age of information overload, on the topic of Automatic Text Summarization.

Text summarization [2] means compressing the source text into a reduced version while conserving its information content and meaning. Since most of the current automated text summarization systems use extraction methods, the paper [2] describes single and multiple document Extractive Summarization techniques. More research in NLP, Machine Learning, Neural Networks, Fuzzy Logic and Abstractive Summarization will be needed to have human like summarization done by Automatic Text Summarizers.

Paper [3] presents a new methodology of crowd-sourcing. It involves distributing translation tasks to a large number of individuals, often referred to as workers or annotators, who provide their own assessments based on predetermined criteria.

Abstractive summaries are generally considered more complex and challenging [4] to generate compared to extractive summaries. Abstractive summarization often requires advanced natural language processing (NLP) techniques, including syntactic and semantic analysis, understanding context and discourse, paraphrasing, and language generation. By conducting a Systematic Literature Survey (SLR) on text summarization, researchers have analyzed and summarized the findings of multiple studies and provided a comprehensive overview of the current state of research on text summarization techniques.

Paper [5] seems to be a valuable resource for researchers in the ATS field. It not only provides a comprehensive understanding of the domain but also assists in identifying research gaps and inspiring new approaches. By synthesizing existing knowledge and presenting it in a structured manner, the survey contributes to the advancement of ATS methods and supports further research in the field.

Paper [6] focuses on using Twitter data to provide opinions on specific products and offer decision-making support to customers. By leveraging sentiment analysis techniques, the paper likely aims to extract valuable insights from user-generated content and provide feedback to companies for product improvement. This implementation can contribute to enhancing customer experiences, facilitating informed decision-making, and fostering a feedback loop between customers and businesses in the era of online communication and social networking.

Paper [7] proposes an approach to address the challenge of training models for languages without publicly available labelled corpora. By utilizing pre-trained sentence vectors, positional encoding, and self-attention mechanisms, the model can capture the semantic meaning, positional information, and dependencies between sentences in the document. This methodology provides a way to overcome the data scarcity issue and

develop effective document summarization models for languages with limited labelled corpora.

The paper [8] focuses on improving keyword detection. It is important to acknowledge that summarizing different types of content, such as news stories, financial reports, or medical reports, requires domain-specific approaches and adaptations. Understanding the unique characteristics and challenges of each domain helps in developing more accurate and relevant summarization methods. Keyword detection in clusters is shown by the authors in this paper.

The paper [9] concentrates on two important aspects of sign language processing: sign language recognition and sign language translation. These tasks aim to recognize continuous signs and map them to sign glosses, as well as generate spoken language translations based on the recognized sign glosses.

In paper [10], the authors emphasize the vital role of text summarization in saving users' time and resources, particularly considering the abundance of available data. They highlight the importance of evaluating and comparing different summarization methods based on accuracy scores to identify the most effective and concise summarization techniques. By leveraging these methods, users can access the key information from large volumes of text efficiently.

3 Proposed Architecture

The following Figure 1 explores the steps involved in automatic extractive text summarization in English, along with the translation of summarized English text in French language. It has been divided into 5 modules of – Text Pre-processing, Text Summarization, Translation Pre-processing, NLP Stages in Translation, Summarized Translation in French.

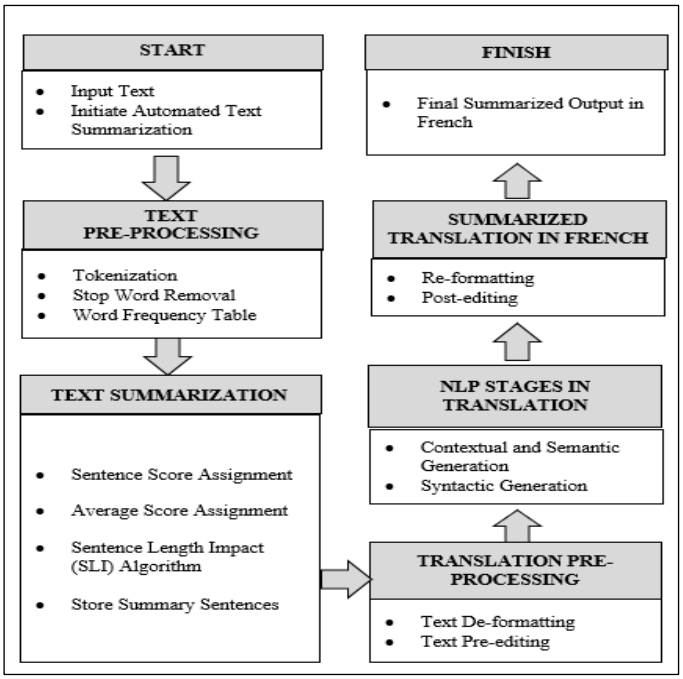

Fig. 1. Architecture of Automatic Text Summarization and Translation

As mentioned in the architecture diagram, below is the proposed **Sentence Length Impact (SLI) Algorithm**

Algo 1: Sentence Length Impact Algorithm

Input: Dataset

Data:
//para=Each paragraph in dataset
//cnt=Number of words
//N=Total number of sentences in dataset
//sent=Individual sentence
//impct=Impact Factor of each sentence based on its length
//summ=Summarized Outcome

Output: Summarized Outcome

```
for i=1 to N do
      for each para
        sent←sent+1 on reaching delimiter like .
       end
      for each sent
        cnt←cnt+1
      end

end

for each sent
       Calculate length and max(cnt)
       Calculate impct

end

if(impct>0.5)
      summ←summ+sent
end

print(summ)
```

Figure 2 depicts the various ways of classification of Automatic Text Summarization and is created based on various papers referenced in this article.

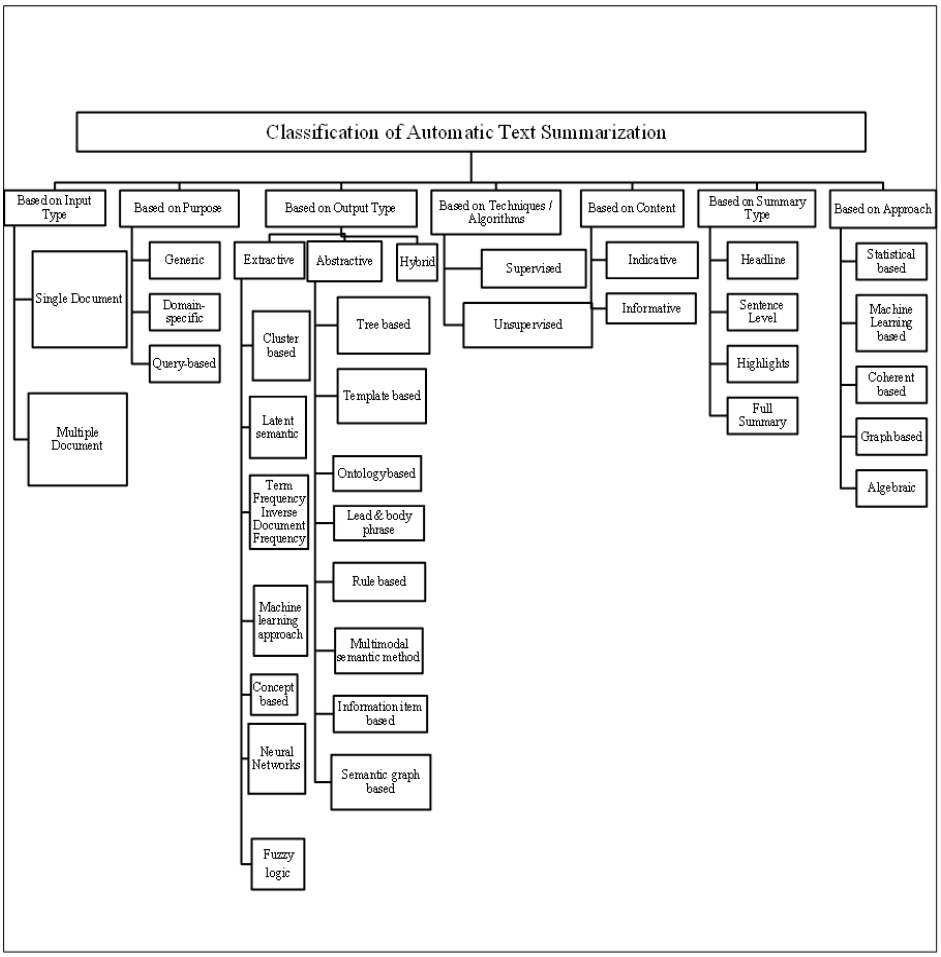

Fig. 2. Classification of Automatic Text Summarization

Figure 3 below, illustrates the architecture diagram of Automatic Text Summarization based on Sentence Length Impact Algorithm and Translation of Summarized Outcome in French.

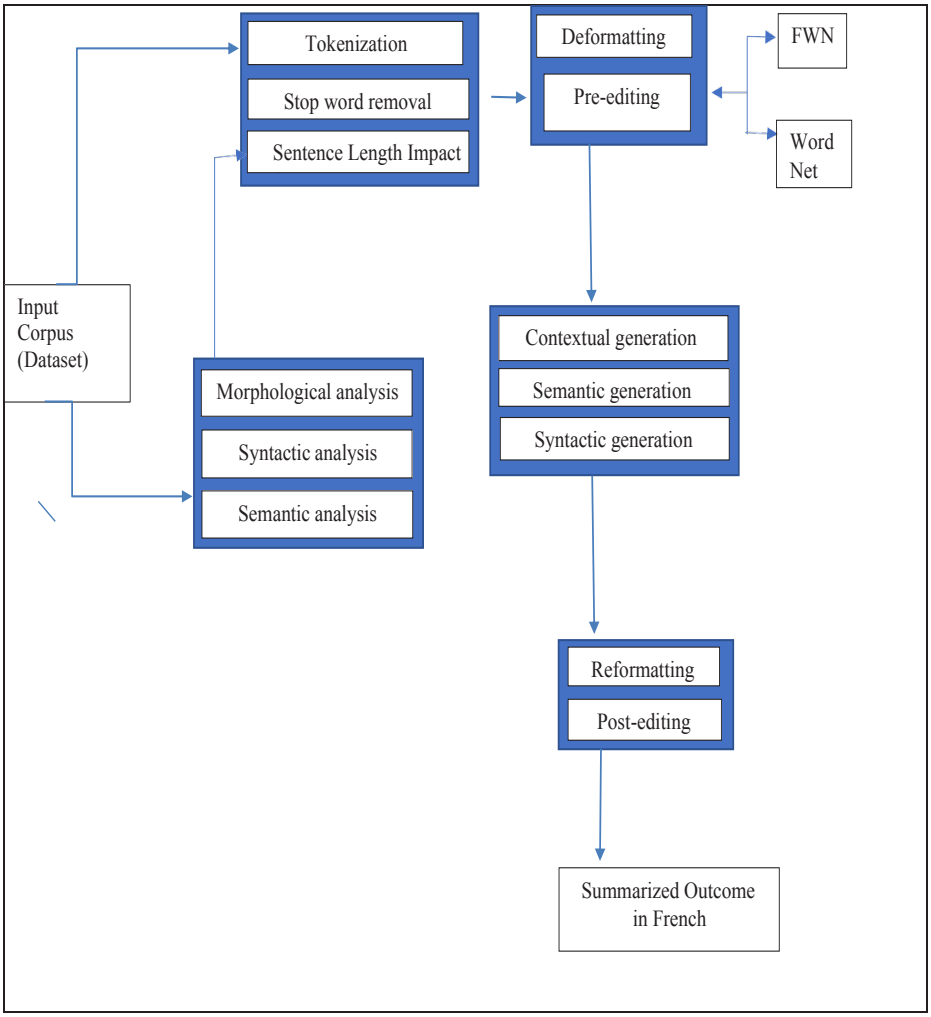

Fig. 3. Automatic Text Summarization based on Sentence Length Impact (SLI) Algorithm and Translation of Summarized Outcome in French

4 Experimental Evaluation

4.1 Dataset 1 Description

The development of logical models with legal reasoning has taken place within the field of AI & law. While AI has made significant contributions to the logical study of law, there is potential for the reverse influence, i.e., learning from AI. AI has introduced innovative approaches and techniques that can enhance the logical study of law in several ways.

4.2 Dataset 2 Description

The analysis of research work is related to fake news detection results in feature extraction and vectorization. It explores traditional machine-learning models for creating a supervised machine-learning algorithm that can classify the news articles as true or false.

4.3 Dataset 3 Description

This sheds light on the problem of plagiarism in research. Plagiarism involves using someone else's intellectual work without proper attribution, either intentionally or unintentionally. It is important to address this issue through awareness, education, and the use of plagiarism detection tools to promote academic integrity and uphold the principles of responsible research.

4.4 Experimental Evaluation Methods

Evaluation methods are crucial for comparing the results of different text summarization methods.

Precision and recall are evaluation metrics, used in information retrieval and text summarization.

- Precision measures the proportion of relevant information in the system-generated summary.
- Recall measures the proportion of relevant information from the reference summary that is present in the system-generated summary.

The terms are defined in the following equations:

$$Precision = Correct / (Correct + Wrong) \quad (1)$$

$$Recall = Correct / (Correct + Missed) \quad (2)$$

In these formulae:

- Correct = the number of sentences that are the same in both summary which are generated by the system and human.

- Wrong = the number of sentences presented in summary and produced by system but is not included in the summary generated by human.

Certain critical measures worked, based on size of the summary are:

- Compression Ratio measures how much shorter the summarized outcome is compared to the original document.
- Retention Ratio measures how much of the central or important information is retained in the summarized outcome.

$$Compression\ Ratio\ (CR) = Length(S) / Length(T) \quad (3)$$

$$Retention\ Ratio\ (RR) = Information\ in\ S / Information\ in\ R \quad (4)$$

- Length(S): Length of Summarized Outcome
- Length(T): Length of Original Document

- Information in S: Central Information retained in Summarized Outcome
- Information in R: Information in Original Dataset

Table 1. Metric wise Evaluation on Different Datasets

Dataset	Precision	Recall	F-Measure
Dataset 1	0.882	0.764	0.818
Dataset 2	0.853	0.789	0.819
Dataset 3	0.794	0.823	0.808
Average	0.843	0.792	0.815

Metric wise evaluation done with three different data sets, the performance of them is an average around 0.8, which is represented in Table 1.1.

Table 1.2: Average Text Compression Ratio and Retention Ratio for Different Frequent Number of Terms

Dataset	Number of Terms	1	2	3	4	5	10	15	25
Dataset 1	Compression Ratio	21	37	52	67	79	85	91	93.8
	Retention Ratio	41	45.7	57.2	64.3	87.5	91	93.5	95
Dataset 2	Compression Ratio	23	40	55	69	81	87	92	94.1
	Retention Ratio	43	48	59	66.2	88.3	92	93.2	96
Dataset 3	Compression Ratio	25	42	53	64	78	86	90	92.6
	Retention Ratio	45	47	60.4	65.6	89.6	92.5	93.4	95.7

The two measures are compression and retention % over the given document results, compared in Table 1.2. The model performance on these measures is validated over the same dataset.

Table 1.3: Algorithm wise Accuracy computation

Algorithm	Accuracy
Fuzzy SVM (Abs)	0.758
Random Forest (Ext)	0.775
Decision Tree (Abs)	0.765
Naïve Bayes (Ext)	0.762
KNN (both Abs and Ext)	0.668
ANN (Ext)	0.801
Proposed Model	**0.921**

After giving the dataset as input, and applying all small pre-processing like removing square brackets, extra spaces, special characters and digits, this is followed by printing the URL of dataset.

Text pre-processing techniques employed over a text article are tokens splitting, stop word removal and updating word-frequency table. Now sentence score is assigned to the individual sentences based on the maximum appearance of the sentence or word in word

frequency table. The word with maximum frequency will be assigned maximum score and thus be added in the summary.

Based on Sentence Length and word with maximum frequency, Sentence Length Impact (SLI) factor is calculated which decides whether a particular sentence will be included in summarized outcome or not.

The summarized output is taken to French translation. After doing the translation pre-processing and applying NLP stages of translation – Contextual, Semantic and Syntactic Generation, and post the reformatting and post editing, the summarized French translation is generated.

Here URL of Dataset 1 is given as input and based on Sentence Length Impact Algorithm, the summarized outcome is generated which is passed to French translator to give the French translation.

Figure 4 shows the designed user interface using Flask web application framework where on giving the required text, to get its summarized outcome using the above procedure.

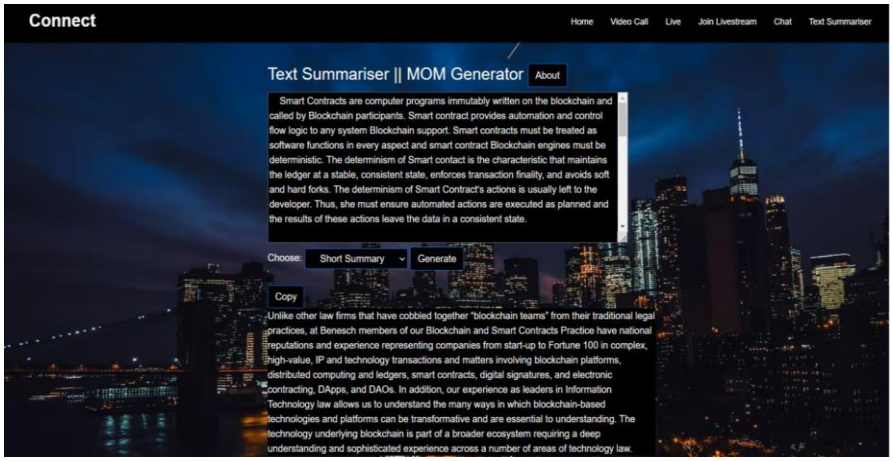

Fig. 4. User Interface with Flask web application to implement Automatic Text Summarization based on Sentence Length Impact (SLI) Algorithm

5 Conclusions and Future Directions

The Text summarization is implemented by applying text pre-processing techniques on dataset like tokenization, stop word removal and updating word-frequency table. Based on sentence score value, sentences or words with maximum frequency are arranged and summarized outcome is generated. After doing the translation pre-processing on summarized English outcome and applying NLP stages of translation – Contextual, Semantic and Syntactic Generation, and post the Reformatting and Post-editing, the summarized French translation is generated. The proposed NLP based Sentence Length Impact (SLI) model gave 92% accurate summarized results and it can be further optimized to get better precision and recall.

References

1. M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, K. Kochut, Proceedings of arXiv, USA, July 2017, 9 pages (2017).
2. O. Tas, F. Kiyani, Press Academia Procedia, **5(1)**, pp.205-213 (2007).
3. Y. Graham, T. Baldwin, A. Moffati, J. Zobel, Natural Language Engineering, Volume **23**, Issue 1 , January 2017 , pp. 3 – 30 (2017).
4. A.P. Widyassari, S. Rustad, G. F. Shidik, E. Noersasongko, A. Syukur, A. Affandy. Journal of King Saud University-Computer and Information Sciences (2020).
5. M. F. Mridha, A. A. Lima, K. Nur, S. C. Das, M. Hasan, M. M. Kabir, IEEE Access, vol. **9**, pp. 156043-156070 (2021).
6. Fernandes, R. D'Souza, IEEE Annual India Conference (INDICON), pp. 1-5 (2016).
7. M. Jang, P. Kang, IEEE Access, **9**, 14358-14368 (2021).
8. Á. Hernández-Castañeda, R. A. García-Hernández, Y. Ledeneva. C. E. Millán-Hernández, IEEE Access, vol. **8**, pp. 49896-49907 (2020).
9. H. Zhou, W. Zhou, Y. Zhou, H. Li, IEEE Transactions on Multimedia, vol. **24**, pp. 768-779 (2022).
10. Rahul, S. Adhikari, M. Siwaliya, *NLP based Machine Learning Approaches for Text Summarization*, Fourth International Conference on Computing Methodologies and Communication (ICCMC), pp. 535-538 (2020).
11. V. Gupta, G. S. Lehal, Web Intelligence **2, 3**, pp 258–268 (2010).
12. G. Erkan, D. Radev, *LexRank,* J. Artif. Intell. Res.(JAIR) **22, 1** (2004), pp 457–479 (2004).
13. E. Baralis, L. Cagliero, S. Jabeen, A. Fiori, S, Shah*,* Expert Systems with Applications **40, 17** (2013), pp 6976–6984 (2013).
14. R. Mihalcea, P. Tarau. *A language independent algorithm for single and multiple document summarization*, (2005).
15. Chin-Yew Lin, *Rouge: A package for automatic evaluation of summaries*, in Text Summarization Branches Out: Proceedings of the ACL-04Workshop, pp 74– 81 (2004).
16. S. Karmakar, T. Lad, H. Chothani, International Research Journal of Computer Science (IRJCS) Issue **1**, Volume **2** (2015).