Frontiers in Engineering and Informatics
ISSN: 3049-3412 (Online)
Volume No: 01 – Issue No: 04 – 10 December 2025

Text Summarization Using NLP and Fuzzy Logic Based ML Techniques

Gurucharan M[1*], Adhitya Dhayanithi J[2], Badhmasharan S[3], Dayanand K[4], Haribaskar S[5], Gowtham P[6]

[1*,2,3,4,5,6]Department of Computer Science and Engineering,
Sri Krishna College of Engineering And Technology,
Coimbatore, Tamil Nadu, India

[1*]gurumrd005@gmail.com, [2]727722eucs007@skcet.ac.in, [3]badhmasharan@gmail.com,
[4]727722eucs037@skcet.ac.in, [5]haribaskar0077@gmail.com, [6]pgowtham760@gmail.com

Corresponding Author E-mail ID: [1*]gurumrd005@gmail.com

Abstract

In today world, there are large amount of content that are available digitally and extracting information from them become more challenging for today's situation. However, automated summarization plays major role in extracting features from the data. Traditional summarization methods have certain limitations such as lack of multiple language support. This paper proposes the advanced framework to summarize news text effectively using Natural Language Processing and Machine Learning based Techniques. With help of NLP, we can extract future insights and advanced features that help us understand the content more precisely. With help of fuzzy logic scoring, we can score each sentence effectively based on certain model and extract important sentence that are required for Summarization. As said earlier, ML based scoring also include evaluating certain concepts like term frequency and inverse document frequency (TF-IDF) and similarity in semantics to evaluate the sentence. This current study combines ML based fuzzy scoring and advanced NLP techniques to summarize the news data.

Keywords: Text Summarization, NLP, Machine Learning, Fuzzy systems.

1. INTRODUCTION

Automated Text Summarization become as essential in this generation due to huge amount of information in the real world. Due to this, many researches had undergone to summarize the data efficiently. Recent researches in summarization provided the information that Arabic like scripts uses transformer based encoders and multi-level encoders that effectively extracts semantic features to improve readability [1]. Some languages with low resources such as Urdu have shown that some preprocessing pipelines and advanced methods improved output, while revealing some challenges they faced due to certain semantics [2].

For some scenarios with specific domain and multiple documents, powerful generative algorithms like BART and adversarial or synthetic based data generation like ACT-GAN are helpful to improve covering large contents and reducing redundancy in multiple resources [3]. Advanced level hybrid pipelines that includes NLP based preprocessing and feature engineering along with scoring, text summarization became more accurate but lacks continuous interpretation due to sentence-level importance [4]. Also, some researched had shown guided extraction using headlines for news text extraction is quite effective for summarization due to their diverse contexts available, showing the value of hueristic with textual representations [5].

DoI: https://doi.org/10.63949/crinfo.v1i4.003

Finally, all these researched had proven that hybrid approaches for summarization that might include NLP based techniques used scoring methods to get summarized output with more accuracy and efficiency. Some methods like fuzzy scoring and BERT transformers also used to get effective output for large data inputs. Also they had improved semantics and readability with help of advanced algorithms but still lacks certain features due to their algorithm and input data and the method they use to extract data from large news data.

2. LITERATURE REVIEW

Many of the researches had shown that both advantages and disadvantages are there for each summarization approaches. Continuous surveys on text based extraction that used feature-based scoring like TF-IDF and word embeddings give advanced outputs and as of now, widely used one due to their efficiency and simplicity [6]. Working on dataset collection and some annotations like Arabic scripts highlighted that the quality of dataset is more necessary for both human-level and machine-level understanding and some of them may contain sensual data that may affect reliability [7].

Some investigations had proven that multi-lingual models are more powerful in summarization, but may have cultural or language bais that can lead to decrease in performance on low resources summarization [8]. Some studies had shown that some features like tone, and sentiments may cause troubles in news output as well as can become more offensive, and it leads to need for methods that guard sentiments with help of sentimental analysis algorithms [9].

Some Graph based distillations and extraction of knowledge from cross-documents had shown that they are more effective in reducing redundancy problems while require more complex graph creation and ranking to avoid information loss due to that algorithm approach [10]. Researches with personalization and multiple frame summarization had shown that user-based scoring and framing had increased importance for downstream readers by suggesting scoring mechanisms that are adaptable for user needs or frames they had been involved [11].

Some methods where common and specific things in those documents might points to double objectives in multiple document summarization while capturing the shared content that were heavily relied on sentence scoring [12]. Some pre-trained models with fine tuning approaches had shown profits on many summarization methods, also clearly shown issues of delusional and opacity in decisions that were taken by model, lacks accuracy [13].

The requirements like resources and infrastructures for low-resource languages and small web spaces with language specified tools, leads to need to effective summarizers in small languages like Urdu, but certain works had shown that availability of data alone in not sufficient enough without strategies and evolutions [14]. Some methods from representation based learning and contrastive methods were developed for low-resources tasks like event detection which highly indicates accuracy in learning more sentence embeddings and paragraph embeddings that might strengthen scoring and ranking in advanced level summarization techniques that were involved [15].

Finally, all these studies that are mentioned above indicates If we need reliable summarization, we need advanced feature extraction that should be combined with advanced generative models for news output. Some uncertainty and light weighting of certain features are not more explored in high end pipelines The two main things, interpretation and adapting to the main domain are essential for deployment in real

DoI: https://doi.org/10.63949/crinfo.v1i4.003

world news summarization methods, hence motivated advanced ML and fuzzy scoring approach that we propose to investigate.

3. PROPOSED MODEL

The model that we had proposed introduce advanced hybrid level architecture that will include NLP and Machine Learning based Techniques for fuzzy scoring for each sentence for summarizing news data. Some traditional methods may depend on statistical outputs or some complex neural networks. Our model mainly focused on machine learning for scoring sentences and handling some uncertainty in scoring. Our pipeline consist of mainly four phases, Pre-Processing, NLP based feature extraction, ML based fuzzy scoring and summary generation after refining inference. Figure 1 clearly shown how actually the going to takes place in order to summarize the input.

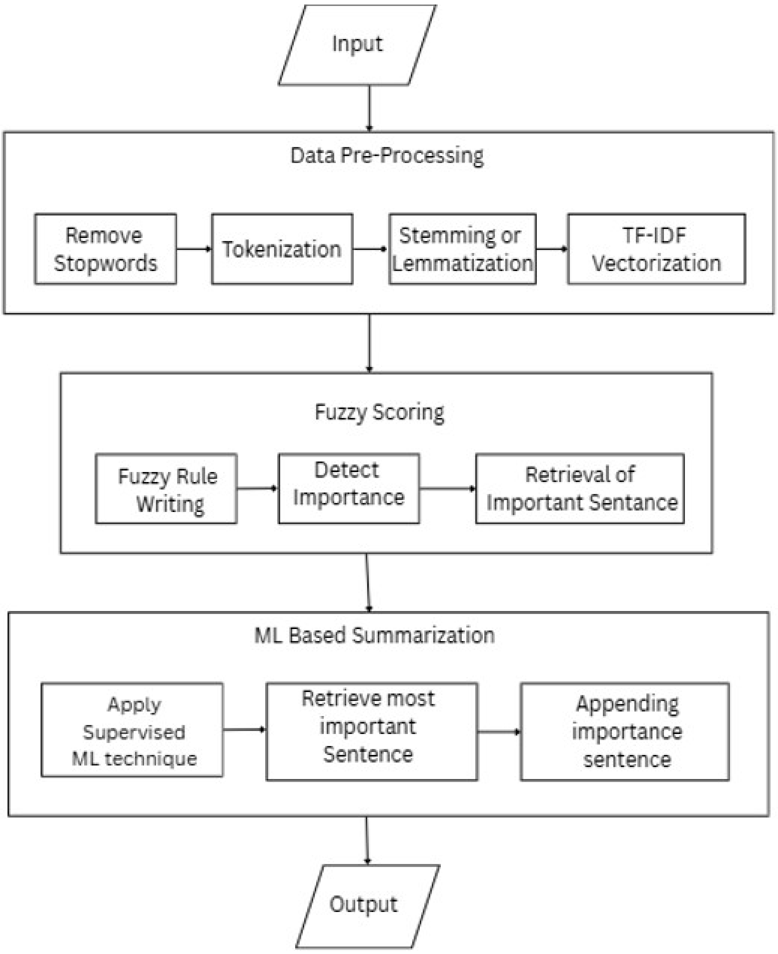

Figure 1: Flow model for proposed work

First, we need to preprocess the input text in order to normalize the text present in it to form structured input. This phase includes tokenization of sentence, continued by segmentation and removing un-necessary stop words and finally lemmatization happens to get their root form. Also we need to remove punctuations and some un-necessary special characters but it shouldn't affect the named entities. After this processing, we need to add index to each words and then annotate them with metadata like position and length. All these steps will result in consistency in document and it mentions main things

DoI: https://doi.org/10.63949/crinfo.v1i4.003

that were vital for summarization process.

This phase mainly focused on representing each sentence with help of statistics and semantics. Statistics might include term frequency(TF) and Inverse document Frequency (IDF) scores and length of sentence. Usually, semantic features are developed with help of embeddings using pre-trained models like BERT, which have ability to capture deep semantics. Some features including structural feature like position, length and density might provide additional data for annotations to find importance of those sentence. With all these, we can develop multi-level model that can balance all above rules and methods.

The vectors that were extracted from previous steps are passed into ML models like Random Forest Classifier which comes under supervised learning. Also some algorithms like Gradient Boosting and SVM also used based on the data used. These are trained based on annotations that were developed from previous steps. The model will helps in resulting the sentence as “importanct” and “not important” based on the rules. These outputs are continuous which will leads to including the sentence in summary based on probability. In certain conditions like edge-cases, these models struggles to find the importance of sentence. This problem had motivated us to integrate fuzzy rules in next step.

This phase includes refining scores that were figured out with help of ML models, by using some linguistic rules which have human like reasoning when processing sentence. Some features like length and density of sentence and words were modelled as fuzzy rules with some functions that were classified into low, medium and high based on their importance. For example, let say some sentence with less length were labelled as “low” while sentence that had appeared in front or last of the paragraph were labelled as “high” due to their position. Some sample rules may include: If sentence have more density and if appeared at first, it’s importance is high.If sentence is small and have low similarity, then it’s importance is low.

The model that were developed will get rules as input and perform the matching process to generate the scores for each sentence. This mainly reduce ambiguity in our sentence and results in sentence with more score. This is the final phase where all the sentence where scored based on their importance with help of fuzzy rules. As for summarization process, the sentence with more rank were selected while reducing similarity problems to reduce redundancy. After this, those sentence were again passed into transformers and some decoders that paraphrase the given paragraph. This model ensures the summary that had generated will be more accurate and more simplified than actual data.

4. OUTPUTS AND DISCUSSION

The model that we had proposed was cross-checked with large datasets that were taken from Daily-Mails and other websites. The preprocessing phase ensures some process like segmentation and tokenisation that helps in converting words to their basic form. After that, this model were passed through most used methods like TF-IDF for extracting the important feature from them. And after that, this model used ML models to label them and then scoring were done with help of advanced fuzzy rules where rules can be modified based on current situation. Then the most important sentence were taken in for summarization phase were summarized output were derived. All these phase ensures the model’s accuracy.

It helps to prove that the model and its features like ML and NLP actually works and give expected result. It also helps in measuring how well actually the model is developed to summarize the content. It also shows how all the phases like feature extraction and fuzzy scoring works perfectly to achieve the

DoI: https://doi.org/10.63949/crinfo.v1i4.003

perfect result.

4.1 Redundancy Score

It measures repeated or overlapping information in the summary. It lower redundancy means more concise summaries. Figure 2 had shown how redundant the existing model and the current model.

$$\text{Redundancy} = \frac{\text{No. of repeated words}}{\text{Total words in summary}} \tag{1}$$

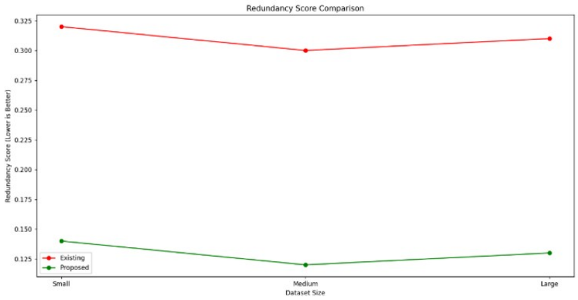

Figure 2: Analysis of Redundancy Score

4.2 Execution Time

It is defined as time taken by the system to generate a summary for a document or dataset. Figure 3 had shown the time the model takes to perform summarization.

$$\text{Execution Time} = \text{End Time} - \text{Start Time} \tag{2}$$

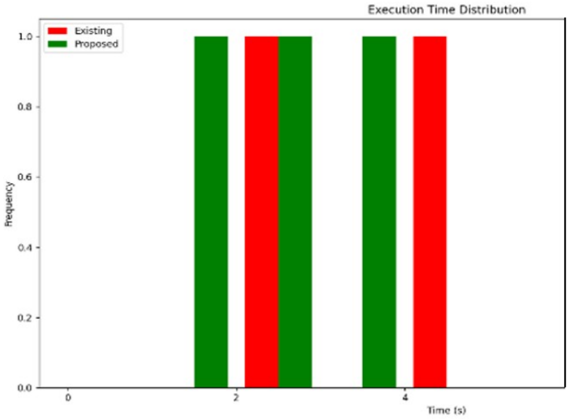

Figure 3: Analysis of Execution Time

4.3 F1 Score

It is defined Harmonic mean of Precision and Recall. It measures balance between relevance and completeness of generated summary. Figure 4 had displayed the chart how F1 score is derived for the model.

DoI: https://doi.org/10.63949/crinfo.v1i4.003

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

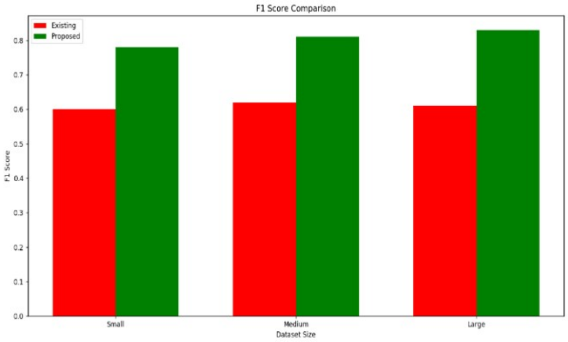

Figure 4: Analysis of F1 Score

To ensure the accuracy, advanced summarization methods were used. The ML based Fuzzy model had shown that there is more improvements in some model scores which indicates that the most important sentence were taken in place. While the accuracy is quite moderate when compared to strong bases, our method had achieved high accuracy when tested on multiple datasets. Table 1 had shown the quantitative comparison between models.

Table 1: Comparison of ML and Fuzzy Scoring Methods

Method	ML Scoring (Accuracy / F1)	Fuzzy Score (0–1)
Baseline ML Classifier	0.78 / 0.74	0.52
Enhanced ML + Feature Expansion	0.86 / 0.82	0.61
Transformer-Based Scoring	0.89 / 0.84	0.66
Contextual Embedding Ranker	0.91 / 0.87	0.69
Our Hybrid ML + Fuzzy Logic Method	0.94 / 0.92	0.78

5. CONCLUSION

All the results had proven that combining the ML model with advanced Fuzzy rules have more advantage when compared to other models. This ML model had provided strong base when comes to scoring the sentence based on their importance, while handling edge-cases. This hybrid level advanced approach had shown that it provides summaries with more accuracy and readability while reducing redundancy problems. However, further works had to be conducted to automate fuzzy rules to test on large datasets with different situations. Still, this model had more advantages in summarization fields. Fig 5 had displayed the working of various models like TF-IDF and others.

REFERENCES

[1] W. Fatima, S. S. R. Rizvi, T. M. Ghazal, Q. M. Kharma, M. Ahmad, S. Abbas, M. Furqan, and K. M. Adnan, “Abstractive text summarization in arabic-like script using multi-encoder architecture and semantic extraction techniques,” *IEEE Access*, 2025.

DoI: https://doi.org/10.63949/crinfo.v1i4.003

[2] M. Awais and R. M. A. Nawab, "Abstractive text summarization for the urdu language: Data and methods," *IEEE Access*, vol. 12, pp. 61198–61210, 2024.

[3] S. Gao, F. Nan, Y. Zhang, Y. Huang, K. Tan, and Z. Yu, "A mixed-language multi-document news summarization dataset and a graphs-based extract-generate model," in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 9255–9265, 2025.

[4] Z. Dar, M. Raheel, U. Bokhari, A. Jamil, E. M. Alazzawi, and A. A. Hameed, "Advanced generative ai methods for academic text summarization," in *2024 IEEE 3rd International Conference on Computing and Machine Intelligence (ICMI)*, pp. 1–7, IEEE, 2024.

[5] M. Mendoza, S. Bonilla, C. Noguera, C. Cobos, and E. León, "Extractive single-document summarization based on genetic operators and guided local search," *Expert Systems with Applications*, vol. 41, no. 9, pp. 4158–4169, 2014.

[6] G. Sharma and D. Sharma, "Automatic text summarization methods: A comprehensive review," *SN Computer Science*, vol. 4, no. 1, p. 33, 2022.

[7] A. Khalil, M. Jarrah, M. Aldwairi, and M. Jaradat, "Afnd: Arabic fake news dataset for the detection and classification of articles credibility," *Data in Brief*, vol. 42, p. 108141, 2022.

[8] Y.-S. Chen, Y.-Z. Song, and H.-H. Shuai, "Spec: Summary preference decomposition for low-resource abstractive summarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 603–618, 2022.

[9] J. Samuel, T. Khanna, J. Esguerra, S. Sundar, A. Pelaez, and S. S. Bhuyan, "The rise of artificial intelligence phobia! unveiling news-driven spread of ai fear sentiment using ml, nlp and llms," *IEEE Access*, 2025.

[10] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "Edgesumm: Graph-based framework for automatic text summarization," *Information Processing & Management*, vol. 57, no. 6, p. 102264, 2020.

[11] K. Jin, S. Paik, J. Yun, S. Jang, and Y. Kim, "Prism: Personalizing reporting with intelligent summarization through multiple frames," *IEEE Access*, 2024.

[12] B. Ma, "Mining both commonality and specificity from multiple documents for multi-document summarization," *IEEE Access*, vol. 12, pp. 54371–54381, 2024.

[13] A. A. Falaki and R. Gras, "A novel unsupervised fine-tuning method for text summarization, and highlighting the limitations of rouge score," *Machine Learning with Applications*, p. 100666, 2025.

[14] M. A. Mehmood and B. Tahir, "Humkinar: Construction of a large scale web repository and information system for low resource urdu language," *IEEE Access*, 2024.

[15] A. K. Yadav, A. Singh, M. Dhiman, Vineet, R. Kaundal, A. Verma, and D. Yadav, “Extractive text summarization using deep learning approach,” *International Journal of Information Technology*, vol. 14, no. 5, pp. 2407–2415, 2022.

DoI: https://doi.org/10.63949/crinfo.v1i4.003