Monolingual text summarization for Indic Languages using LLMs

Jothir Adithya T K, Nithish Kumar S, Felicia Lilian J, Mahalakshmi S
Department of Computer Science and Business Systems
Thiagarajar College of Engineering, Madurai.
{jothir, snithishkumar, mahalakshmis2}@student.tce.edu ,jflcse@tce.edu

Abstract

We have analyzed the growth of advanced text summarization method leveraging LLM for Indic language. Text summarization involves transforming a longer text information into a more concise version, ensuring that the most prominent information and key meanings are maintained. Our goal is to produce concise and accurate summaries from longer texts, focusing on maintaining detailed information and coherence. We utilize NLP techniques for text cleaning, keyword extraction and summarization, along with performance evaluation metrics such as ROUGE score, BLEU score and BERT Score. The results demonstrate an incremental improvement in the quality of generated summaries, with a particular emphasis on enhancing informativeness while minimizing redundancy. This research work also highlights the importance of tuning parameters and leveraging advanced models for producing high-quality summaries in diverse domains for Indic Language.

1 Introduction

The process of generating longer textual content into the shorter version while preserving the prime information and its meaning defined as text summarization. Effective summarization enhances readability and aids in quick understanding of the content (Dou et al., 2020). Text summarization is a critical tool in the digital age, allowing for the extraction of important information from large content of data (Bhatnagar et al., 2023). Text summarization in low resource language which is the most challenging one in the field of NLP.

In this paper, we focus on refine text summarization for English and Indic Tamil language by implementing our model using the Gemini LLM (Lal et al., 2023). The challenges which lies in ensuring that the generated summaries are both comprehensive and concise, covering all essential aspects of the original content while avoiding redundancy (Nayak and Timmapathini, 2021). We applied several Natural Language Processing techniques, including text cleaning and keyword extraction using the Rake algorithm, to ensure the quality of the input data. Our model utilizes the map-reduce summarization chain, designed to retain important details in the output summary (Dhar and Das, 2021). We then evaluate the generated summaries against reference summaries using multiple performance metrics including ROUGEScore, BLEUScore and BERTScore. To evaluate the model success in balancing coverage, precision and readability the performance metrics are used (Roychowdhury et al., 2022). By focusing on important sentences, we demonstrate the model's ability to generate insightful, well-rounded summaries (Steinberger and Ježek, 2012).

1.1 Key Objective of this research work

In this paper, we have aim to analyse and address the problem of tokenizing for Indic languages,

- Analyzing Tokenization Difficulties: Investigate issues related to segmenting text accurately due to the language complexity.
- Morphology and script features for Indic languages: Evaluating keyword extraction and assess how tokenization affects the accuracy of keyword extraction from Indic texts.
- Improving Stopword Extraction: Examine challenges in identifying and removing stopwords and propose enhancements based on tokenization results.

2 Methodology

2.1 Overview of Proposed Model

This research work have leverages the modular framework to building custom language model applications. The focus is on generating high-quality

summaries by chaining LLM functionalities with advanced text preprocessing techniques and incorporating evaluation metrics such as ROUGEScore, BLEUScore and BERTScore.

Our aim have to improve the comprehensiveness and informativeness of the output summaries, ensuring about the output summaries captures key aspects of the text and adheres to the factual accuracy of the original documents (Liu et al., 2020). The Figure 1 represents the work flow of text summarization.

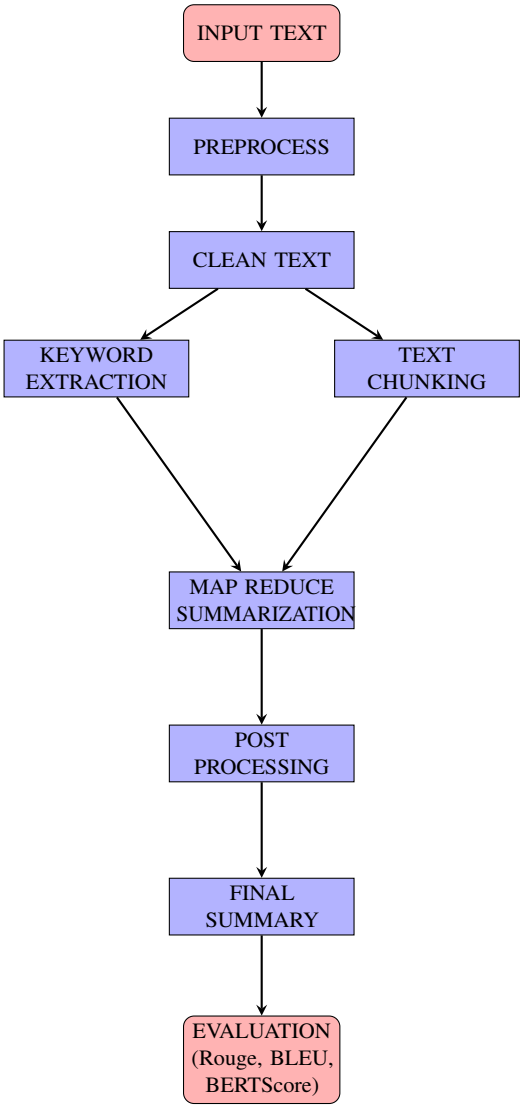

Figure 1: Workflow of Text Summarization

2.2 Training & Testing

The training process involves preprocessing text,extracting keyword and structuring input with a tailored prompt template for a pre-trained LLM to generate summaries. The testing process evaluates these summaries against reference text using ROUGE,BLEU and BERTscore metrics to ensure accuracy,relevance and coherence.

2.3 Data and Document Preparation

The documents used for summarization were primarily long-form texts related to various domain such as Medical field, Agriculture, Political, Space and Education domain. We have manually created the text summarization datasets for English and Tamil Language from authorized newspapers, such as The Hindu, Dinamalar, and Malaimalar. Upon request, we will provide access to these datasets. Table 1 provides the dataset details represents the number of paragraphs in each domain for both English and Tamil Language.

Domain	**English**	**Tamil**	**Total**
Medical	890	560	1450
Agriculture	760	610	1370
Political	660	820	1480
Space	520	520	1040
Education	910	845	1755
Total	**3740**	**3355**	**7095**

Table 1: Dataset details

2.3.1 Preprocessing

Before providing the text into the summarization LLM model, several preprocessing steps have applied. The figure 2, represents the flowchart of text cleaning process.

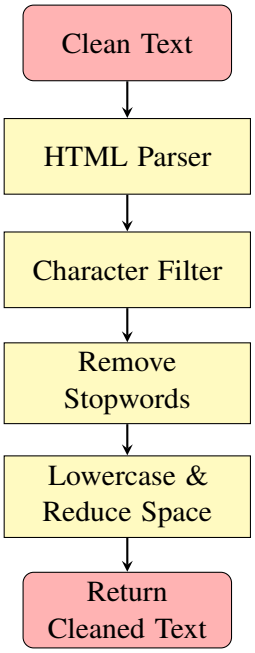

Figure 2: Flowchart of Text Cleaning Process

- **HTML Tag Removal**: For web-sourced documents, HTML tags were removed using *BeautifulSoup*.
- **Special Character Filtering**: Non-alphabetic characters were stripped, retaining only the core text content.
- **Stopword Removal**: Stopwords were removed to minimize noise in the text and thus enhancing the clarity of the summary.
- **Text Normalization**: All text have changed into lowercase and multiple spaces are reduced to attain single space for consistency.

2.4 Text Summarization Pipeline

2.4.1 Text Splitting

The original text was split into smaller, manageable chunks to make sure the input text fits in the token limit of the LLM Model and provides comprehensive summaries. Initially, a chunk size of 200 characters with a 100-character overlap was used to retain context between sections.These parameters may vary depending on adjustments made during the model training process.

Size, Overlap	Param	R-1	R-2	R-L	BERT
100,10	precision	0.361	0.277	0.361	0.734
100,10	recall	0.211	0.146	0.211	0.727
100,10	F1-score	0.386	0.211	0.386	0.730
200,50	precision	0.443	0.222	0.426	0.709
200,50	recall	0.241	0.143	0.236	0.687
200,50	F1-score	0.369	0.226	0.361	0.698
500,100	precision	0.546	0.369	0.543	0.775
500,100	recall	0.384	0.273	0.371	0.783
500,100	F1-score	0.564	0.302	0.545	0.779

Table 2: Illustration of Chunk variation and Results for Tamil

2.4.2 Summarization Chain Design

We utilized our LLM Model to construct a summarization pipeline. The figure 3, represents the flowchart of summarization pipeline. The pipeline is based on the **map-reduce** approach which operates as follows,

- **Map Step**: Each chunk of text is summarized individually by our LLM Model.
- **Reduce Step**: The partial summaries are combined into a single comprehensive summary, ensuring key points from all sections are captured.

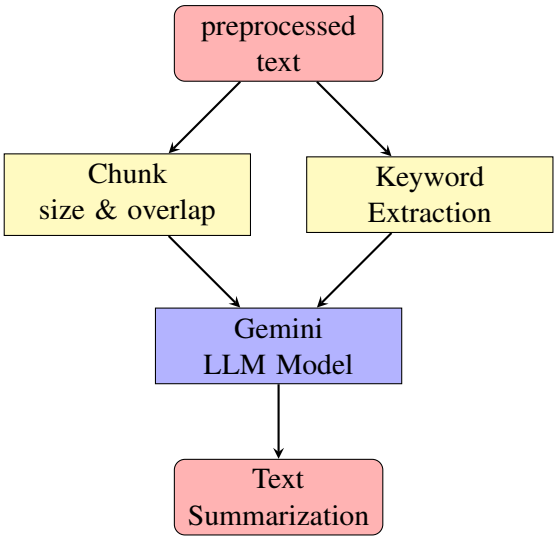

Figure 3: Flowchart of summarization pipeline

2.5 Keyword Extraction for Focused Summarization

To enhance the informativeness of the summaries, we integrated a **keyword extraction** step using the *RAKE algorithm and TF-IDF*. These keywords helped focus the LLM model on identifying the most important points of the document while ensuring that the summaries remain relevant and detailed.

2.6 Evaluation Metrics

2.6.1 ROUGE Score

We have researched the calibre of the generated summaries using **ROUGEScore** which is a common standard metric for summarization tasks. The following variants of ROUGEScore were used to evaluate precision measure, recall measure and F1 score:

- **ROUGE-1**: Overlap of unigrams between the reference and generated summaries.
  - Precision: $\frac{\text{overlapping unigrams}}{\text{unigrams in generated summary}}$
  - Recall: $\frac{\text{overlapping unigrams}}{\text{unigrams in reference summary}}$
  - F1-Score: $\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

- **ROUGE-2**: Overlap of bigrams between the reference and generated summaries.
  - Precision: $\frac{\text{overlapping bigrams}}{\text{bigrams in generated summary}}$
  - Recall: $\frac{\text{overlapping bigrams}}{\text{bigrams in reference summary}}$
  - F1-Score: $\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

- **ROUGE-L**: Longest common subsequence between reference and generated summaries.
  - Precision: $\frac{\text{LCS length}}{\text{generated summary length}}$
  - Recall: $\frac{\text{LCS length}}{\text{reference summary length}}$
  - F1-Score: $\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

2.6.2 BLEU Score

We assessed the summaries using the BLEU (Bilingual Evaluation Understudy) metric, which calculates n-gram precision while penalizing summaries that are too short.

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \cdot \log p_n\right)$$

2.6.3 BERTScore

We utilized **BERTScore**, which compares the contextual embeddings of the reference and generated summaries, providing precision, recall, and F1 scores based on meaning rather than surface-level matching. The BERTScore is computed as follows

$$\text{Precision} = \frac{1}{|c_g|} \sum_{w_g \in c_g} \max_{w_r \in c_r} \text{sim}(w_g, w_r)$$

$$\text{Recall} = \frac{1}{|c_r|} \sum_{w_r \in c_r} \max_{w_g \in c_g} \text{sim}(w_g, w_r)$$

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

where: - c_g and c_r are the contextual embeddings of the generated and reference summaries, respectively. - $\text{sim}(w_g, w_r)$ is the similarity score between words w_g from generated summary and w_r from reference summary.

2.7 Post-Summarization Improvements

Post-summarization adjustments were made based on evaluation scores:

- **Fine-Tuning Prompts**: Prompts were iteratively refined to improve the depth and clarity of the summaries.

- **Adjusting Temperature and Top-p Values**: Distinct scores for both temperature and top-p were tested to find the optimal balance between creativity and factuality.

- **Hierarchical Summarization**: In some cases, a hierarchical summarization approach was applied to produce both high-level and detailed summaries.

2.8 Automated Workflow and Scalability

The entire pipeline was automated using our LLM model batch processing capabilities, allowing for the efficient summarization of multiple documents in a scalable manner. Integration with cloud storage (Google Drive) enabled easy access and management of generated summaries (Jangra et al., 2020).

3 Result

3.1 Domain: Medical

About: Monkeypox

Lang	Param	R-1	R-2	R-L	BERT
ENG	precision	0.945	0.588	0.490	0.794
TAM	precision	0.681	0.292	0.650	0.817
ENG	recall	0.273	0.370	0.342	0.828
TAM	recall	0.230	0.254	0.226	0.817
ENG	F1-score	0.424	0.264	0.220	0.811
TAM	F1-score	0.218	0.392	0.211	0.817

Table 3: ROUGE and BERT scores - Medical Domain

Lang	Parameter	BLEU
ENG	Score	0.351
TAM	Score	0.220

Table 4: BLEU Score- Medical Domain

Lang	Param	No.of.words	No.of.Lines
ENG	Reference Summary	1420	50
ENG	Generated Summary	434	40
TAM	Reference Summary	1519	41
TAM	Generated Summary	282	38

Table 5: Difference between reference and generated summary- Medical Domain

Table 2 illustrates the R1, R2, R-L and BERT values for Medical Domain. Table 3 provides the BLUE score and Table 4 illustrates the output generated for Medical domain.

3.2 Domain: Education

About: Outcome Based Education

Lang	Param	R-1	R-2	R-L	BERT
ENG	precision	0.785	0.401	0.592	0.880
TAM	precision	0.614	0.256	0.602	0.729
ENG	recall	0.494	0.252	0.373	0.879
TAM	recall	0.240	0.355	0.236	0.745
ENG	F1-score	0.607	0.310	0.458	0.880
TAM	F1-score	0.227	0.391	0.323	0.737

Table 6: ROUGE and BERT scores - Education Domain

Lang	Parameter	BLEU
ENG	Score	0.398
TAM	Score	0.211

Table 7: BLEU Score - Education Domain

Lang	Param	No.of.words	No.of.Lines
ENG	Reference Summary	336	7
ENG	Generated Summary	224	25
TAM	Reference Summary	1146	26
TAM	Generated Summary	262	51

Table 8: Difference between reference and generated summary- Education Domain

Table 5 illustrates the R1, R2, R-L and BERT values for Education Domain. Table 6 provides the BLUE score and Table 7 illustrates the output generated for Education domain.

3.3 Domain: Agriculture

About: Technological Improvement in Agriculture

Lang	Param	R-1	R-2	R-L	BERT
ENG	precision	0.868	0.413	0.494	0.852
TAM	precision	0.800	0.250	0.600	0.946
ENG	recall	0.331	0.357	0.389	0.851
TAM	recall	0.243	0.237	0.307	0.946
ENG	F1-score	0.480	0.228	0.273	0.851
TAM	F1-score	0.242	0.265	0.382	0.946

Table 9: ROUGE and BERT scores - Agriculture Domain

Lang	Parameter	BLEU
ENG	Score	0.346
TAM	Score	0.144

Table 10: BLEU Score - Agriculture Domain

Lang	Param	No.of.words	No.of.Lines
ENG	Reference Summary	609	22
ENG	Generated Summary	285	18
TAM	Reference Summary	673	19
TAM	Generated Summary	273	40

Table 11: Difference between reference and generated summary - Agriculture Domain

Table 8 illustrates the R1, R2, R-L and BERT values for Agriculture Domain. Table 9 provides the BLUE score and Table 10 illustrates the output generated for Agriculture domain.

3.4 Domain: Space

About: Chandrayaan 1,2

Lang	Param	R-1	R-2	R-L	BERT
ENG	precision	0.854	0.486	0.540	0.888
TAM	precision	0.833	0.520	0.833	0.817
ENG	recall	0.554	0.315	0.351	0.881
TAM	recall	0.306	0.263	0.306	0.812
ENG	F1-score	0.672	0.382	0.425	0.884
TAM	F1-score	0.447	0.348	0.448	0.817

Table 12: ROUGE and BERT scores - Space Domain

Lang	Parameter	BLEU
ENG	Score	0.328
TAM	Score	0.158

Table 13: BLEU Score - Space Domain

Lang	Param	No.of.words	No.of.Lines
ENG	Reference Summary	415	18
ENG	Generated Summary	267	27
TAM	Reference Summary	1107	37
TAM	Generated Summary	337	15

Table 14: Difference between reference and generated summary - Space Domain

Table 11 illustrates the R1, R2, R-L and BERT values for Space Domain. Table 12 provides the BLUE score and Table 13 illustrates the output generated for Space domain.

3.5 Domain: Political

About: Political Development in India

Lang	Param	R-1	R-2	R-L	BERT
ENG	precision	0.894	0.507	0.672	0.869
TAM	precision	0.404	0.391	0.442	0.719
ENG	recall	0.269	0.252	0.303	0.853
TAM	recall	0.203	0.338	0.298	0.705
ENG	F1-score	0.414	0.234	0.311	0.861
TAM	F1-score	0.269	0.360	0.261	0.712

Table 15: ROUGE and BERT scores - Political Domain

Lang	Parameter	BLEU
ENG	Score	0.344
TAM	Score	0.257

Table 16: BLEU Score - Political Domain

Lang	Param	No.of.words	No.of.Lines
ENG	Reference Summary	1169	43
ENG	Generated Summary	355	17
TAM	Reference Summary	1119	19
TAM	Generated Summary	224	20

Table 17: Difference between reference and generated summary - Political Domain

Table 14 illustrates the R1, R2, R-L and BERT values for Political Domain. Table 15 provides the BLUE score and Table 16 illustrates the output generated for Political domain.

4 Error Analysis

In the initial stages of our work, errors were encountered during the keyword extraction and chunking processes. Specifically, our pre-processing approach did not adequately handle certain delimiters, leading to the loss of contextually significant separators, which negatively impacted downstream tasks. To address this issue, we introduced a text-cleaning method that preserved delimiters such as commas, periods, semicolons, and exclamation marks. This modification significantly improved the performance of both the keyword extraction and chunking processes.

For Tamil keyword extraction, our initial approach utilized the RAKE algorithm. However, RAKE struggled to identify semantically meaningful keywords in Tamil text due to its reliance on co-occurrence patterns, which lacked adaptation for the nuances of the Tamil language.

To overcome this limitation, we incorporated a TF-IDF-based approach. This model emphasized the statistical importance of terms by considering their frequency relative to the document and the overall corpus. The TF-IDF approach demonstrated a marked improvement in the precision and recall of extracted keywords, yielding better alignment with expected results.

5 Conclusion

In this research paper, we have studied and analysed the LLM methodology for text summarization. Our proposed method involves preprocessing the input text, performing keyword extraction and using a Map-Reduce summarization technique to generate concise and relevant information summaries. The process are designed to handle large scale texts by splitting them into manageable chunks and applying summarization to each chunk, followed by a combining step to produce a coherent final summary (Nayak and Timmapathini, 2021).

Evaluation metrics such as ROUGEScore, BLEUScore and BERTScore are applied to evaluate the quality and accuracy of the generated output summaries. The scores demonstrate that the proposed method have effectively captures the key information. Our results show promising improvements in ROUGE-1, BLEU and BERTScore highlighting the effectiveness of our approach.

Additionally, the integration of sustainability efforts, recent regulatory changes and the usage of hybrid power units are effectively summarized from a larger body of text (Kamal et al., 2022). This framework can be further enhanced by exploring other LLMs and fine-tuning for domain-specific texts, making it applicable for real-world summarization needs in various fields.

Limitations

The Reference summary or Dataset for English and Tamil languages which we have collected from various domain-specific news articles from standard newspapers such as The Hindu, Dinamalar, Malaimalar. So, the generated summaries are somewhat not equal in length due to some minimal differences in the dataset content of various domain.Our model has effectively handled ambiguous or polysemous words in most cases, though there are occasional instances where its performance could be improved. This highlights an area for further improvement, particularly in dealing with complex language nuances

Acknowledgement

This research work is supported by MuthirAI-A Global Research Center for Tamil and AI and funded by the Thiagarajar Research Fellowship (TRF) by Thiagarajar College of Engineering, Madurai, India

References

Isa M. Apallius de Vos, Ghislaine L. van den Boogerd, Mara D. Fennema, and Adriana Correia. 2021. Comparing in context: Improving cosine similarity measures with a metric tensor. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 128–138, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Nikhilesh Bhatnagar, Ashok Urlana, Pruthwik Mishra, Vandan Mujadia, and Dipti M. Sharma. 2023. Automatic data retrieval for cross lingual summarization. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 822–827, Goa University, Goa, India. NLP Association of India (NLPAI).

Purnima Bindal, Vikas Kumar, Vasudha Bhatnagar, Parikshet Sirohi, and Ashwini Siwal. 2023. Citation-based summarization of landmark judgments. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 588–593, Goa University, Goa, India. NLP Association of India (NLPAI).

Rudra Dhar and Dipankar Das. 2021. Leveraging expectation maximization for identifying claims in low resource Indian languages. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 307–312, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Zi-Yi Dou, Sachin Kumar, and Yulia Tsvetkov. 2020. A deep reinforced model for zero-shot cross-lingual summarization with bilingual semantic similarity rewards. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 60–68, Online. Association for Computational Linguistics.

Sunil Gundapu and Radhika Mamidi. 2020. Multichannel LSTM-CNN for Telugu text classification. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON): TechDOfication 2020 Shared Task*, pages 11–15, Patna, India. NLP Association of India (NLPAI).

Anubhav Jangra, Raghav Jain, Vaibhav Mavi, Sriparna Saha, and Pushpak Bhattacharyya. 2020. Semantic extractor-paraphraser based abstractive summarization. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 191–199, Indian Institute of Technology Patna, Patna, India. NLP Association of India (NLPAI).

Pratik Joshi, Christain Barnes, Sebastin Santy, Simran Khanuja, Sanket Shah, Anirudh Srinivasan, Satwik Bhattamishra, Sunayana Sitaram, Monojit Choudhury, and Kalika Bali. 2019. Unsung challenges of building and deploying language technologies for low resource language communities. In *Proceedings of the 16th International Conference on Natural Language Processing*, pages 211–219, International

Institute of Information Technology, Hyderabad, India. NLP Association of India.

Ashraf Kamal, Padmapriya Mohankumar, and Vishal K Singh. 2022. IMFinE:an integrated BERT-CNN-BiGRU model for mental health detection in financial context on textual data. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON)*, pages 139–148, New Delhi, India. Association for Computational Linguistics.

Daisy Monika Lal, Paul Rayson, Krishna Pratap Singh, and Uma Shanker Tiwary. 2023. Abstractive Hindi text summarization: A challenge in a low-resource setting. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 603–612, Goa University, Goa, India. NLP Association of India (NLPAI).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro{bert}a: A robustly optimized {bert} pretraining approach.

Jash Mehta, Deep Gandhi, Naitik Rathod, and Sudhir Bagul. 2021. IndicFed: A federated approach for sentiment analysis in indic languages. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 487–492, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Anmol Nayak and Hari Prasad Timmapathini. 2021. Using integrated gradients and constituency parse trees to explain linguistic acceptability learnt by BERT. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 80–85, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Rajendra Roul. 2021. Multi-document text summarization using semantic word and sentence similarity: A combined approach. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 423–430, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Sohini Roychowdhury, Kamal Sarkar, and Arka Maji. 2022. Unsupervised Bengali text summarization using sentence embedding and spectral clustering. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON)*, pages 337–346, New Delhi, India. Association for Computational Linguistics.

Aafiya S Hussain, Talha Z Chafekar, Grishma Sharma, and Deepak H Sharma. 2022. Event oriented abstractive summarization. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON)*, pages 99–108, New Delhi, India. Association for Computational Linguistics.

Numair Shaikh, Jayesh Patil, and Sheetal Sonawane. 2023. Query-based summarization and sentiment analysis for Indian financial text by leveraging dense passage retriever, RoBERTa, and FinBERT. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 398–407, Goa University, Goa, India. NLP Association of India (NLPAI).

Josef Steinberger and Karel Ježek. 2012. Evaluation measures for text summarization. *COMPUTING AND INFORMATICS*, 28(2):251–275.

Raji Sukumar, Hemalatha N, Sarin S, and Rose Mary C A. 2021. Text based smart answering system in agriculture using RNN. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 663–669, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).