Leveraging LLaMa for Abstractive Text Summarisation in Malayalam: An Experimental Study

Hristo Tanev
Joint Research Centre, European Commission
Italy
htanev@gmail.com

Anitha S Pillai
Department of Computer Applications
Hindustan Institute of Technology and Science, Chennai, India
anithasp@hindustanuniv.ac.in

Revathy V R
Department of Computer Science
Kristu Jayanti (Deemed to be) University, Bengaluru, India
revathyvrajendran@gmail.com

Abstract

Recent years witnessed tremendous advancements in natural language processing (NLP) because of the development of complex language models that have automated several NLP applications, including text summarisation. Despite this progress, Malayalam text summarisation still faces challenges because of its unique grammatical structures. This research paper explores the potential of using a large language model, specifically the LLaMA (Large Language Model Meta AI) framework, for abstractive text summarisation of Malayalam language. In order to assess the performance of LLaMA for text summarization, for the low- resource language Malayalam, a dataset was curated with reference text and summaries. The evaluation showed that the LLaMA model could effectively summarize lengthy articles while maintaining important information and coherence. The generated summaries were compared with the reference summaries generated by human writers to observe how well aligned the model was with a human level of summarisation. The results proved that LLM can deal with the Malayalam text summarisation task, but more research is needed to understand the most relevant training strategy.

1 Introduction

Being popular among India's 22 officially recognised languages, Malayalam is primarily spoken in the South Indian state of Kerala. Connected through its roots with the ancient Dravidian language family, Malayalam has developed its unique alphabet by adapting linguistic elements from Sanskrit. Malayalam is uniquely characterised by its ability to have constructive word structuring acquired through its xtensive vocabulary and linguistic and agglutinative nature (Nambiar et al., 2023). The rapid accumulation of digital content has necessitated the need for effective text summarisation models. Especially in Malayalam, a low-resource language with multiple dialects, there is a need for text summarisation for various NLP applications. Though there is a demand for text summarization models in Malayalam, the problem in developing them is due to inflectional structure and limited availability of annotated data. Researchers show extensive interest in developing text summarisation models that can generate concise, coherent, and contextually significant Malayalam text summarisation. Text summarisation is an NLP task that condenses lengthy documents into brief versions without losing significant information.

The substantial daily generation of content such as news articles on digital platforms and the unavailability of online tools for processing them necessitate a robust model for text summarisation. The demand for automated text summarisation has become greater than for manual text summarisation, as it requires a lot of time and labour. Automated Malayalam text summarisation tools will also be an essential resource for journalists, scholars, and readers to bring rapid accessibility to the information of their choice. Most summarisation research articles use abstractive summarisation approaches over extractive types (Nambiar et al., 2023; Shakil, Farooq, and Kalita., 2024). While generating novel summaries, abstractive methods keep originality in essence, and extractive

Proceedings of Recent Advances in Natural Language Processing,pages 1248–1255
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-144

approaches directly generate sentences by joining the important phrases and constructing sentences. Other complexities with abstractive summarisation models for the

Malayalam is a morphologically rich and agglutinative language, with suffixes being extensively used for indicating grammatical functions such as tense, case, mood, and number. These properties enhance exponentially the vocabulary space and make the standard text processing techniques less effective. Besides, this language has a relatively free word order (normally Subject-Object-Verb but has flexibility) which turns syntactic parsing more complex and hinders training in sequence models that depend on fixed word positions. Compared to other Indian languages, there are less resources in Malayalam concerning available annotated corpora and tools for NLP tasks.

Malayalam language includes the requirement of fully labelled datasets and intricacies because of the richness and copious morphology of Malayalam syntax. Further, constructing a model that is capable of comprehension and fluent generation of Malayalam text is difficult. Malayalam language has nominally rich Dravidian morphological features, such as extensive inflection, agglutination, and compounding, which impose peculiar constraints on language modeling and text generation. These challenges are typically addressed by transformer-based models such as LLaMA through subword tokenisation techniques like Byte Pair Encoding or SentencePiece, which decompose the intricate word forms into simple subword units so that they can generalise to morphologically diverse forms supported by that model.

Current developments seen in sophisticated language models such as Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformer (GPT) and their successors have made improvements in addressing some of the most challenging tasks in NLP (Chhibbar and Kalita., 2024; Doss., 2024; Hemamou and Debiane., 2024; Mudigonda et al., 2024; Zhang, Yu, and Zhang., 2024). These Large Language Models (LLMs) are deep learning algorithms with a large number of parameters, billions or trillions, trained on large volumes of data. LLMs use the transformer architecture (Ravaut et al., 2024). It is equipped to acquire context and meaning by breaking words into sequences within a sentence.

Recent work in (Deroy and Maity., 2024) has highlighted the subword-based modelling efficacy in low-resource languages like many other Dravidian languages. Their findings showed how, through prompt learning and subword tokenisation, one can significantly improve performance, especially for morphologically complex and code-mixed text.

Inspired by this, for this study, LLaMA has employed for Malayalam summarisation using similar prompt methods, with LLaMA's multilingual pre-training and vocabulary flexibility. This enables the model to generate linguistically coherent and contextually accurate summaries. Additionally, models like the LLaMA-based MalayaLLM, which incorporates a custom vocabulary of 18,000 Malayalam tokens, have shown that adapting LLaMA for morphologically rich languages offers improved performance, even in low-resource or one-shot learning scenarios.

2 Recent Work

The announcement of the first generative large language model (LLM) by OpenAI marked a revolutionary milestone in the development of natural language processing technologies (Radford et al., 2018). These models are trained on vast datasets and consist of billions of parameters, enabling them to engage in a wide range of conversations and perform numerous tasks. These include sentiment analysis, named entity recognition, and various forms of language understanding and generation (Sindhu et al., 2024). One of the most powerful capabilities of generative LLMs is their ability to produce human-like text, often indistinguishable from content written by actual people (Touvron et al., 2023).

Models like BERT and RoBERTa are used for applications, such as text classification, sentiment analysis and outcome prediction (Prasanthi et al., 2023). Some of the existing LLMs are multilingual: supporting many languages, performing Machine Translation, and performing tasks in cross-language contexts (Mujadia et al., 2023)

According to (Ilanchezhiyan et al. 2023), transformer models can be effectively applied to text summarization in various Indian languages. Fine-tuning these models on Indian language datasets was found to significantly improve the quality of the generated summaries.

The authors of (Munaf et al. 2023) investigated how to apply models, generally pre-trained, to perform summaries for under-resourced languages. They were able to prove that fine-tuning of such models brought better summary quality. Generative LLMs tend to be applied mainly in summarising texts, translating languages, or conversing via chatbots. Unlike these models, the discriminative LLMs aim for classifying and predicting purposes. Conversational LLMs are models that can talk like a human through the understanding of context; numerous chatbots and virtual assistants currently utilise such models (Sindhu et al., 2024). Such studies include (Mujadia et al., 2023), who researched the extent to which LLMs translate between English and different Indian languages. They also looked into how these language models account for cultural complexity.

Another study presented a novel way to summarise PDF documents making use of LLMs (Ramprasad and Sivakumar., 2024). Their work underlines the importance of context relevance in the summarisation process.

An overview of how these LLMs were changing was given by (Patil and Gudivada., 2024) in 2024. They highlighted key improvements, like better training techniques. The study by (Zhang et al. 2024) checked various LLMs to see how well they make news summaries. The work by (Pradhan and Todi, 2023) looked at metrics from LLMs to evaluate summarization quality.

(Mullick et al., 2024) explored ways to improve aspect-based summarisation by refining LLMs. They focused on producing summaries that hit on specific parts of the text. Their work showed that fine-tuning models boosts accuracy and relevance.

(Parmar et al., 2024) aimed to improve how coherent summaries are. They created a special dataset to test and enhance how well LLMs keep summaries logical.

Another study by (Ravaut et al., 2024) examined how LLMs utilise context in summarization tasks, emphasizing the role of contextual information in improving summary quality. Sindhu et al. analyzed the evolution of LLMs, their applications, and emerging challenges (Sindhu et al., 2024). They discussed how advancements in architecture have made LLMs more efficient. Suleiman and Awajan provided insights into extractive text summarization techniques and their evaluation measures (Suleiman and Awajan, 2019).

For this research, Llama 3.1 70B Instruct, which Meta released in July 2024 was used (Touvron et al., 2023). LLaMA, or Large Language Model Meta AI, is a series of models developed by Meta AI since February 2023. Its transformer-based architecture enables it to process and understand language in a manner that closely resembles human comprehension. When initially released, LLaMA shared its model weights with researchers for non-commercial use. This made it a handy model for the academic and research community.

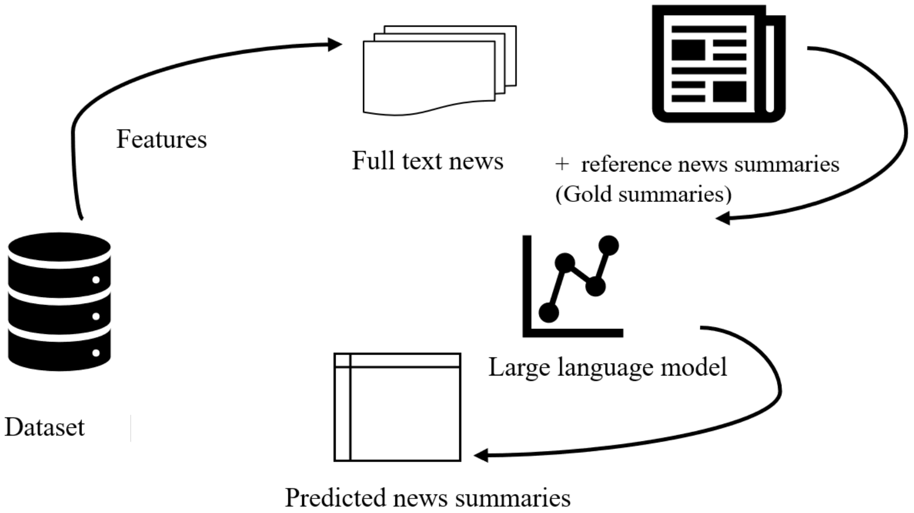

Figure 1. Proposed System Architecture

3 Data Collection and Model development

For this research work, a dataset of dimensionality 2 with 500 samples was created. The sample blog articles were collected from several news websites such as Malayalam.news18.com, reporterlive.com and Malayalam.samayam.com. The first column is Malayalam full length article and the second column is the summary. This dataset is useful for building and testing models that work with the Malayalam language. It can help with tasks like summarising text, sorting articles by topic, and looking at trends in Malayalam text articles over different categories of architecture.

Figure 1 shows the architectural diagram for summarising these articles. LLaMa 3.1 is used to summarise the Malayalam articles in this study.

The reference summaries are the gold standard for training the model. They help the model learn how the articles are structured and what the important parts are. Once the model is trained, it can take new text articles and make concise summaries. These summaries focus on the key details from the original text. The setup is built to handle the unique features of the Malayalam language. This way, the summaries stay accurate and relevant. By automating the summarization, it is possible to handle lots of text content quickly and easily, making it simple to share important information.

3.1 Model Training and Testing

The full text and its abstractive summary from the dataset are used to train and test the proposed model. The one- shot learning strategy takes one example and uses it for training. This learning approach is used to improve the summation ability of the LLaMA model.

Several robust features have motivated for considerations for usage of the LLaMA (the Large Language Model Meta AI) framework towards preference over other large language models. Unlike commercial models such as GPT-3 or PaLM, the LLaMA is open-weight and allows fine-tuning, thus making it more suitable for academia and experimentation. The smaller versions have a favorable trade-off between computational efficiency and performance: extremely important when working with languages that have limited resources like Malayalam. Preliminary investigations and benchmarks also indicate that LLaMA performs well on multilingual tasks, especially when fine-tuned with domain-specific training. Hence, the good accessibility, customizability, and efficiency of the model made it an appropriate choice for this task, which relates to low-resource summarization without the constraints of API access, licensing, or computing expenses.

A single Malayalam article along with its reference summary were randomly selected from the curated Malayalam text summary corpus. The article-summary pair then used as an in-context example to prompt and enhance the model's performance. The prompt text used for the The entire dataset was fed into the Llama 3.1 model, which created summaries.

The ability of the model understands different languages and generate summaries proved here also; the model generated summaries that were clear and relevant. They closely matched the gold summaries. The model also understood Malayalam well, which helped it provide short summaries that kept the important aspects of the articles. This showed that Llama 3.1 can work well with languages that have low resources, like Malayalam.

4 Experimental Setup

. To evaluate how closely the model's summaries matched gold summaries, ROUGE, a standard benchmark for measuring summarisation quality, was used. Specifically, the measured ROUGE-1, ROUGE-2, and ROUGE-L scores assess the overlap of unigrams, bigrams, and longest common subsequences between the machine- generated and human summaries.

The scores were computed using a library called PyRouge (Heinzerling.B., 2018) to evaluate their performance in terms of how well the machine's outputs are able to match individual ones. ROUGE-1 considers the single words, ROUGE-2 the pairs, and ROUGE- L checks the longest matching sequence between both summaries (Lin, C.Y., 2004). Each gold summary was provided with one LLaMA- generated summary. The process for obtaining the ROUGE score was done on each summary separately. Once all the summaries are ready, to evaluate them, the scorers were used. The scorer considered F1 scores for ROUGE-1, ROUGE-2, and ROUGE-L, showing how well the two summaries matched in precision and recall. If either summary was missing or not valid, then the ROUGE scores were set to None. This kept the dataset accurate. After calculating the scores, they were stored in lists for each ROUGE type. In the next step, these scores were added as new columns in the original dataset. This setup allowed us to analyze how well the model performed in summarising Malayalam text across different scoring methods. It gave us a solid way to see how the LLaMa-made summaries overlap with the human ones.

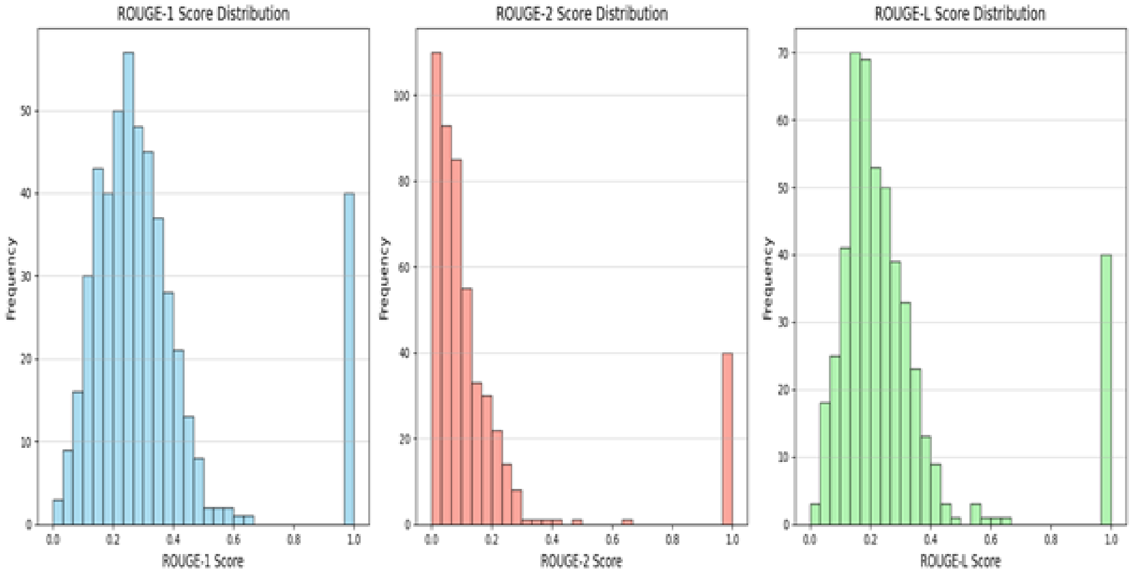

Figure 2. Rouge score distribution for Malayalam text dataset

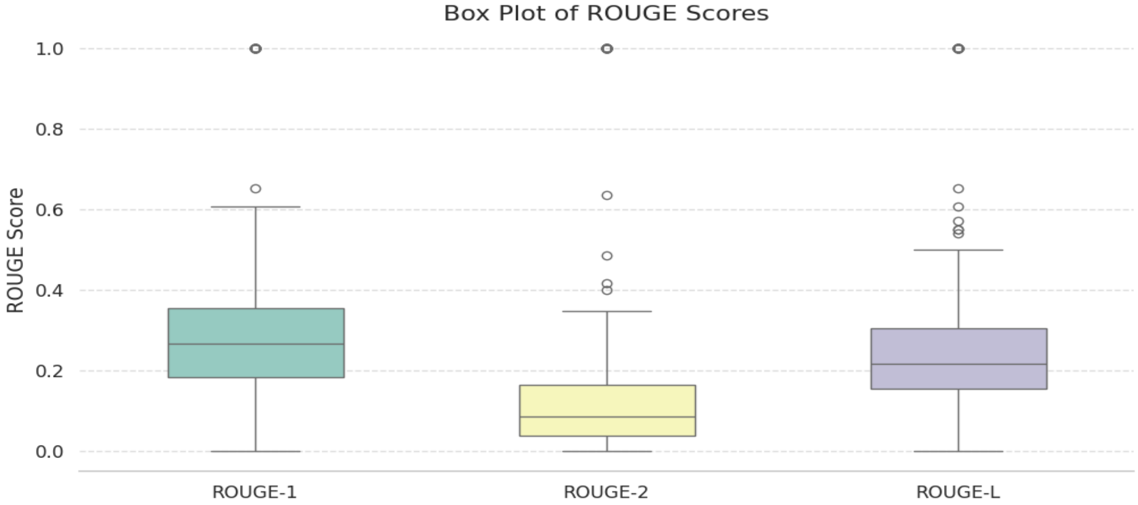

Figure 3. Box plot showing Rouge score distribution for Malayalam text dataset

5 Results And Discussion

The graphs (Figure 2) display the distribution of ROUGE-1, ROUGE-2, and ROUGE-L scores, which are metrics used to evaluate the quality of text summaries against reference summaries. In the ROUGE-1 score distribution graph, the majority of the scores are concentrated between 0.2 and 0.4, indicating that many of the generated summaries have moderate overlap with the reference summaries at the unigram level. However, there is also a noticeable spike for the ROUGE-1 score of 1.0, suggesting that a subset of summaries perfectly matched their reference summaries. The ROUGE-2 and ROUGE-L score distribution graphs show similar patterns, with most scores clustering towards the lower end of the spectrum, particularly between 0.0 and 0.2. This implies that the generated summaries generally have lower bigram and sequence overlap with the reference summaries (Krishnaprasad et al., 2016).

Like the ROUGE-1 graph, there is a notable spike at a score of 1.0 in both ROUGE-2 and ROUGE-L, which indicates that some summaries matched perfectly in terms of bigrams and sequence with their references.

The box plot (Figure 3) displays the distribution of ROUGE-1, ROUGE-2, and ROUGE-L scores, providing a summary of the central tendency, spread, and outliers for each metric. The box in each plot shows the interquartile range (IQR), with the line enclosed in the box indicating the median score. The ”whiskers” stretch to the minimum and maximum values that lie within 1.5 times the IQR, with any points beyond this range considered outliers and shown as individual dots. For the ROUGE-1 scores, the median is around 0.2, and the interquartile range spans from about 0.1 to 0.35. There are some outliers above 0.6, indicating a few summaries that performed significantly better than most (Steffes et al., 2023).

The ROUGE-2 scores have a lower median, around 0.1, and a narrower IQR, suggesting that most summaries had lower bigram overlap with the reference summaries. ROUGE-L scores have a distribution similar to ROUGE-1 but slightly lower, with a median close to 0.2. The presence of several outliers in the ROUGE-L plot suggests that while most summaries had moderate sequence overlap, a few achieved near-perfect matches.

A small qualitative analysis of the output of the LLM was conducted, and for the analysis, 10 randomly selected summaries were selected. It was found that in all of them, the generated summary by LLaMA is relevant, and in only 3 cases were there small factual inaccuracies, which were born out of small differences in the semantics of a statement in the given text and the semantics of its corresponding statement in the LLaMA LLM abstractive summary. It was also inspected the outliers, who scored 1.0, and it was found out that none of them were standard articles but food recipes or health advice related to foods (Table 1 and Table 2).

6 Conclusion

In this study, the potential of the LLaMA framework for text summarisation in the Malayalam language, an area that has remained relatively under-researched despite the language's complex structure and limited digital resources was explored. Also, it showcased the LLaMA 3.1 model's capacity to generate concise summaries that are coherent and relevant when compared to reference summaries created by human writing, developing a special Malayalam text dataset. The ROUGE scores obtained are in line with other summarisation experiments which use, however, deep neural networks specifically trained for the purpose (still on different data sets). The results clearly show that the LLM can provide a baseline solution for summarization of low-resourced languages with moderate accuracy.

Exactly matched gold summaries and LLaMa Summaries
Graampu chertha vellam, anti-inflammatory gunangalnalkkukayum, inflammation kuraykkanum, panjasaraavum rakthachapavum niyanthrikkayum sahayikkunnu.
Vavarile kozhuppine kuraykkan inji, kurumullak, sherappu eniva sahayikkunnu. Inji metabolism mechappettukkayum, kozhuppu kuraykkan sahayikkukayum cheyyunnu.
Vaazhapazham dietil Cherkkunnathu fiber, potassium, vitamin C, B6 eniva nalkunnu. Ith madhumehathe niyanthrikkayum, Kosha santhulitham nilanirthanum sahayikkunnu.
Salmonmatsyam dietil uppeeduthunathu Hridaya arogavyum mechappettuttanum, cholesterolkuraykkayum sahayikkunnu.

Table 1: Exactly matched gold summaries and LLaMa Summaries

Exactly matched gold summaries and LLaMa Summaries
Adding cloves to water helps reduce anti-inflammatory properties, lowers inflammation, and helps control blood sugar and blood pressure.
Boiled okra helps reduce fat. Ginger, lemon, and jaggery help in reducing fat. Ginger boosts metabolism and helps reduce fat.
Adding banana to the diet provides fiber, potassium, vitamin C, and B6. This helps control d
Including salmon in the diet improves heart health, boosts metabolism, and helps reduce cholesterol.

Table 2: English Translation of Table 1 summaries

References

Chhibbar, N. and Kalita, J. 2024. Automatic summarization of long documents. Proc. 21st Int. Conf. on Natural Language Processing (ICON), 607– 615.

Deroy, A. and Maity, S. 2024. Prompt engineering using GPT for word-level code-mixed language identification in low-resource Dravidian languages. arXiv preprint arXiv:2411.04025. DOI: 10.48550/arXiv.2411.04025.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. Proc. NAACL-HLT (Vol. 1), 4171–4186. DOI: 10.18653/v1/N19-1423.

Doss, S. 2024. Comparative analysis of news articles summarization using LLMs. Proc. 2024 Asia Pacific Conf. on Innovative Technologies (APCIT), 1–6.

Heinzerling. ., "pyrouge: A Python wrapper for the ROUGE summarization evaluation package," GitHub repository, 2018. [Online]. Available: https://github.com/bheinzerling/pyrouge

Hemamou, L. and Debiane, M. 2024. Scaling up summarization: Leveraging large language models for long text extractive summarization. arXiv preprint arXiv:2408.15801. Available at: https://arxiv.org/abs/2408.15801.

Ilanchezhiyan, V., Darshan, R., Dhitshithaa, E.M. and Bharathi, B. 2023. Text summarization for Indian languages: Finetuned transformer model application. FIRE (Working Notes), 22(2), 766–774.

Krishnaprasad, P., Sooryanarayanan, A. and Ramanujan, A. 2016. Malayalam text summarization: An extractive approach. Proc. 2016 Int. Conf. Next Generation Intelligent Systems (ICNGIS), 1–4. DOI: 10.1109/ICNGIS.2016.7854070.

Mudigonda, K.S.P., Anand, A., Reddy, R.V. and Kumar, S.D. 2024. Extractive text summarization on medical insights using fine-tuned transformers. Int. J. Comput. Appl. 46, 11, 957–973. DOI: 10.1080/1206212X.2024.2401081.

Mujadia, V. et al. 2023. Assessing translation capabilities of large language models involving English and Indian languages. arXiv preprint arXiv:2311.09216. Available at: https://arxiv.org/abs/2311.09216.

Mullick, A. et al. 2024. Leveraging the power of LLMs: A fine-tuning approach for high-quality aspect- based summarization. arXiv preprint arXiv:2408.02584. DOI: 10.48550/arXiv.2408.02584.

Munaf, M., Afzal, H., Mahmood, K. and Iltaf, N. 2023. Low resource summarization using pre-trained language models. ACM Trans. Asian Low-Resour. Lang. Inf. Process. DOI: 10.1145/3675780.

Nambiar, K.S., Peter, S.D. and Idicula, S.M. 2023. Abstractive summarization of text document in Malayalam language: Enhancing attention model using POS tagging feature. ACM Trans. Asian Low- Resour. Lang. Inf. Process. DOI: 10.1145/3696107.

Parmar, M. et al. 2024. Towards enhancing coherence in extractive summarization: Dataset and experiments with LLMs. Proc. 2024 Conf. Empirical Methods in Natural Language Processing (EMNLP), 19810– 19820. DOI: 10.18653/v1/2024.emnlp-main.1106.

Patil, R. and Gudivada, V. 2024. A review of current trends, techniques, and challenges in large language models (LLMs). Appl. Sci. 14, 5, 2074. DOI: 10.3390/app14052074.

Pradhan, A. and Todi, K. 2023. Understanding large language model based metrics for text summarization. Proc. 4th Workshop on Evaluation and Comparison of NLP Systems (Eval4NLP), 149–155. DOI:

10.18653/v1/2023.eval4nlp-1.12.

Prasanthi, K. N., Madhavi, R. E., Sabarinadh, D. N. S., and Sravani, B. 2023. A novel approach for sentiment analysis on social media using BERT & ROBERTA transformer-based models. In Proceedings of the 2023 IEEE 8th International Conference for Convergence in Technology (I2CT). IEEE, 1–6. DOI: 10.1109/I2CT57861.2023.10100057

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In Text Summarization Branches Out: Proceedings of the ACL-04 Workshop (pp. 74–81). Barcelona, Spain.

Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I. 2018. Improving language understanding by generative pre-training. NeurIPS. Available at: https://www.semanticscholar.org/paper/Improving- Language-Understanding-by-Generative-Radford- Narasimhan/cd18800a0fe0b668a1cc19f2ec95b5003d 0a5035.

Ramprasad, A. and Sivakumar, P. 2024. Context-aware summarization for PDF documents using large language models. Proc. 2024 Int. Conf. Expert Clouds and Applications (ICOECA), IEEE, 186–191.

Ravaut, M., Sun, A., Chen, N. and Joty, S. 2024. On context utilization in summarization with large language models. Proc. 62nd Annu. Meeting of the ACL (Vol. 1: Long Papers), 2764–2781. DOI:

10.18653/v1/2024.acl-long.153.

Shakil, H., Farooq, A. and Kalita, J. 2024. Abstractive text summarization: State of the art, challenges, and improvements. arXiv preprint arXiv:2409.02413. Available at: https://arxiv.org/abs/2409.02413.

Sindhu, B., Prathamesh, R.P., Sameera, M.B. and KumaraSwamy, S. 2024. The evolution of large language models: Models, applications and challenges. Proc. 2024 Int. Conf. Current Trends in Advanced Computing (ICCTAC), IEEE,1-8.

Steffes, B., Rataj, P., Burger, L. and Roth, L. 2023. On evaluating legal summaries with ROUGE. Proc. 19th Int. Conf. on Artificial Intelligence and Law, 457–461. DOI: 10.1145/3594536.3595167.

Touvron, H. et al. 2023. LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971. Available at: https://arxiv.org/abs/2302.13971.

Zhang, H., Yu, P.S. and Zhang, J. 2024. A systematic survey of text summarization: From statistical methods to large language models. arXiv preprint arXiv:2406.11289. Available at: https://arxiv.org/abs/2406.11289.

Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K., and Hashimoto, T. B. 2024. Benchmarking large language models for news summarization. Trans. Assoc. Comput. Linguistics 12, 39–57. DOI: 10.1162/tacl_a_00632.