Character-Centric Summarization and Keyphrase Extraction with Visualization in Indian Mythology

Apurba Paul, Anupam Rana, Ishan Chattopadhaya, and Dipankar Das

Abstract With an emphasis on the **Mahabharata**, a complex narrative with many characters, events, stories, and deep relationships, this paper investigates character-centric summarization and keyphrase extraction in Indian mythology. Through refinement of transformer-based models such as BART, PEGASUS, and T5 and using the dataset, $DS_{character}$, the study generates character-specific summaries. In the dataset, the character-centric sentences emphasize the roles, actions, and significance of individual characters in the text. In case of capturing the coherence and consistency of character-specific summaries when the models were assessed using ROUGE and BLEU scores, PEGASUS performed the well overall. This study also applied KeyBERT for the extraction of character-specific keyphrases which achieves high recall and precision in recognizing important themes linked to characters. To understand the significance of each character in the story, later on we focused on to implement the visual representation of the character-specific summaries and keyphrases. This research offers a detailed understanding of how character-centric mythological sentences are processed and makes recommendations for future research areas to enhance model performance in character-centric tasks.

Keywords Summarization · Keyphrase extraction · Indian mythology · Mahabharata · Transformer-based models

A. Paul et. al—Authors contributed equally.

A. Paul (✉) · A. Rana · I. Chattopadhaya
Institute of Engineering and Management, University of Engineering and Management, Kolkata, India
e-mail: apurba.saitech2021@gmail.com

A. Rana
e-mail: anupam.rana2022@iem.edu.in

A. Paul · D. Das
Jadavpur University, Kolkata, India

J. K. Mandal et al. (eds.), *Recent Advances in Artificial Intelligence and Smart Applications*, Lecture Notes in Networks and Systems 1380,
https://doi.org/10.1007/978-981-96-5822-0_13

1 Introduction

Indian mythology, with its rich and intricate narratives, presents unique challenges for natural language processing due to the complexity of its stories and the vast number of characters involved. One of the two great epics, the **Mahabharata**,[1] presents a particularly difficult scenario because of its intricate and interwoven character interactions. This intricacy makes it an excellent work for investigating character-centric summarization, in which the roles, actions, and interactions of certain characters are highlighted. Transformer-based models like BART, PEGASUS, and T5 have recently improved the efficiency and performance of text summarization methods. However, the research work on character-centric text summarization is not so common. This research suggests fine-tuned versions of these models to address the problem of character-specific summaries in the **Mahabharata**. Using the in-depth knowledge of linguistic structures, these models produce summaries that are coherent and appropriate for the context. So the research of fine-tuning these models on more specific task such as character-centric summarization has a rising need, even though the traditional summarization focuses on entire texts.

In the narratives like the **Mahabharata** of mythology, the character interactions are pivotal and a dedicated approach is needed to identify each character's significance. This effort is driven by the goal of improving the quality and relevance of summaries of mythical texts by highlighting the roles and actions of particular characters. When applied to character-centric narratives, traditional summary techniques frequently ignore these important details, producing summaries that are less informative. In order to provide summaries that are more representative of each character's contributions to the **Mahabharata** and offer deeper, more contextual insights into the story, this work intends to optimize T5, BART, and PEGASUS especially for character-centric summarization. In this research, we faced a number of challenges like—due to the narrative's intricate structure, it is challenging to isolate and summarize character-specific actions. Next, it requires a careful balance to guarantee that important activities of all the characters from the most significant ones like *Krishna* and *Arjuna* to the minor ones are accurately depicted. The lack of adequate resources to meet the current demand results in significant manual effort and time required to create the dataset. Finally, there are technological challenges in fine-tuning the models such as BART, PEGASUS, and T5 to produce character-centric summaries while preserving narrative coherence.

This work aims to achieve three primary objectives: (1) fine-tune the BART, PEGASUS, and T5 models for character-centric summarization by capturing the roles and actions of individual characters in the text; (2) develop a task-specific dataset ($DS_{character}$) of character-centric sentences; and (3) keyphrase extraction and visualization that will show the key themes and actions associated with each character to give readers a better understanding. This paper contributes in a number of significant ways. For character-centric summary, it first presents improved versions of BART, PEGASUS, and T5 that are especially suited for intricate mythological works

[1] https://www.wisdomlib.org/hinduism/book/the-mahabharata-mohan.

like the **Mahabharata**. Secondly, it provides a character-specific content-focused custom dataset ($DS_{character}$), which is an essential training and assessment resource. Thirdly, it provides a new framework for keyphrase extraction and visualization, which makes it easy to investigate and comprehend the roles and connections between each character in the narrative.

The remainder of the paper is structured as follows: Related work is covered in Sect. 2, with an emphasis on current methods for keyphrase extraction, text summarization, and the applications of the analysis of mythological literature. The $DS_{character}$ dataset was prepared in detail in Sect. 3, which also explains the extraction of character-based phrases for fine-tuning. The system framework is presented in Sect. 4, including the fine-tune and implementation of BART, PEGASUS, and T5 for character-centric summarization and keyphrase extraction. In Sect. 5, the outcomes are investigated, and each model's performance on the character-specific dataset is assessed. A discussion and observation section is provided in Sect. 6, which examines the significant findings and how they advance our knowledge of mythological writings. The work is finally concluded and new research options are suggested in Sect. 7.

2 Related Work

2.1 Text Summarization

With the introduction of deep learning models like BART, PEGASUS, and T5, text summarization has seen substantial progress to generate concise summaries. T5 [1] reframes various NLP tasks as text-to-text tasks, showing strong performance in producing coherent summaries by converting tasks into a text generation framework. BART [2] utilizes a denoising autoencoder structure to reconstruct disrupted input text, creating fluent summaries by effectively capturing text structure and context.

PEGASUS [3] focuses specifically on abstractive summarization by implementing a gap-sentence generation approach during pre-training, making it highly suitable for generating detailed, contextually rich summaries. Additional advancements include BERTSUM [4], which enhances extractive summarization with BERT embeddings, and the Hierarchical Transformer model [5], which manages long-form document summarization. SummaRuNNer [6] applies attention mechanisms to improve the generation of abstractive summaries.

Further studies focus on optimizing encoder–decoder structures for enhanced coherence [7] and applying contrastive learning to better extract key sentences [8]. However, the summarization of mythological texts remains relatively unexplored. Using these summarization models for mythology can simplify narratives, facilitate comparative studies, support digital archiving projects, and produce educational resources that make these texts more accessible and engaging.

2.2 *Keyphrase Extraction*

Keyphrase extraction focuses on identifying important phrases within text, which assists with indexing, summarization, and retrieval. In mythological texts, where narratives are complex and symbolic, keyphrase extraction helps identify essential elements such as characters and key events.

KeyBERT [9] is an unsupervised model that employs BERT embeddings to extract keyphrases without requiring labeled data, proving effective in classification and summarization contexts. EmbedRank [10] uses embeddings combined with graph-based ranking to identify keyphrases, ranking them by semantic similarity, which is particularly useful for unsupervised tasks.

By accounting for word position and co-occurrence, PositionRank [11] improves keyphrase extraction. CopyRNN [12] uses a supervised approach that generates keyphrases not necessarily present in the text, making it effective for research papers and complex fields like mythology.

For both summarization and keyphrase extraction tasks, multi-head attention networks [13] utilize attention to focus on critical parts of the text. YAKE! [14] performs effectively across various types of texts based on word frequency and position. Neural networks with attention were applied by [15] to enhance the context-dependent keyphrase extraction.

Due to the lack of research work done on character-centric summary and keyphrase extraction from texts like **Mahabharata** our work will have a significant contribution.

3 Preparation of Dataset

The dataset preparation is a critical step to enhance the performance of character-centric summarization and keyphrase extraction. In this study, we used the $\mathbf{DS}_{character}$ dataset, specifically curated for summarizing and extracting essential information related to characters in Indian mythology. The preparation process included several important steps:

The $\mathbf{DS}_{character}$ dataset was developed based on the PouranicTopic dataset by Paul et al. [16]. This dataset provides an extensive collection of mythology-related texts, particularly emphasizing descriptions and narratives of various characters. To optimize the dataset's quality and applicability, we applied the following pre-processing steps: the text was tokenized into sentences and words using a specific tokenization tool, such as SpaCy.[2] Named entities, especially character names, were recognized and tagged using a Named Entity Recognition (NER) model. Unnecessary content, such as metadata or irrelevant information, was removed to keep only the character-centric text.

[2] https://spacy.io/.

After tagging character names through NER, we extracted sentences associated with each character. This process involved segmenting the text into individual sentences and filtering them to retain only those mentioning the recognized characters. For each character, we compiled a list of relevant sentences. Additionally, each character is paired with an image to facilitate visualization, helping users to visually associate each summary with the corresponding character. To support fine-tuning, concise summaries were manually crafted for each character's set of sentences, providing an accurate, context-rich representation of each character's role in the narrative, which in turn serves as valuable input for model training. Quality checks were performed through both manual review of summaries to confirm accuracy and contextual fidelity and automated validation to maintain uniformity and thoroughness across the dataset. An example of this dataset preparation approach is provided below.

Original: *Abhimanyu was born to Subhadra, the sister of Vasudeva, and Arjuna, making him the grandson of the illustrious Pandu. Known in his previous life as the mighty Varchas, son of Soma, he was reborn as Abhimanyu, a warrior of extraordinary deeds and accomplishments. Abhimanyu fathered a son, Parikshit, who was beloved by Vasudeva himself. Among all the warriors, Abhimanyu stood as the perpetuator of his family line, earning renown for his strength and valor.*

Summary: *Abhimanyu, son of Subhadra and Arjuna, was a renowned warrior known for his extraordinary deeds and valor. A reincarnation of Varchas, he continued his legacy through his son Parikshit, earning admiration as the perpetuator of his family line.*

This dataset preparation process was essential to tailor the data for character-specific summarization and keyphrase extraction, forming a robust basis for the tasks that followed.

4 System Framework

The system framework developed for this study offers an integrated approach to character-centric summarization, keyphrase extraction, and visualization. It comprises three main components: character-centric summarization, keyphrase extraction, and visualization, each essential for efficient data processing and presentation. This structure is illustrated in Fig. 1.

4.1 Character-Centric Summarization

The initial component of the framework focuses on character-centric summarization, aimed at generating concise summaries that revolve around individual characters within the dataset. The process includes the following steps:

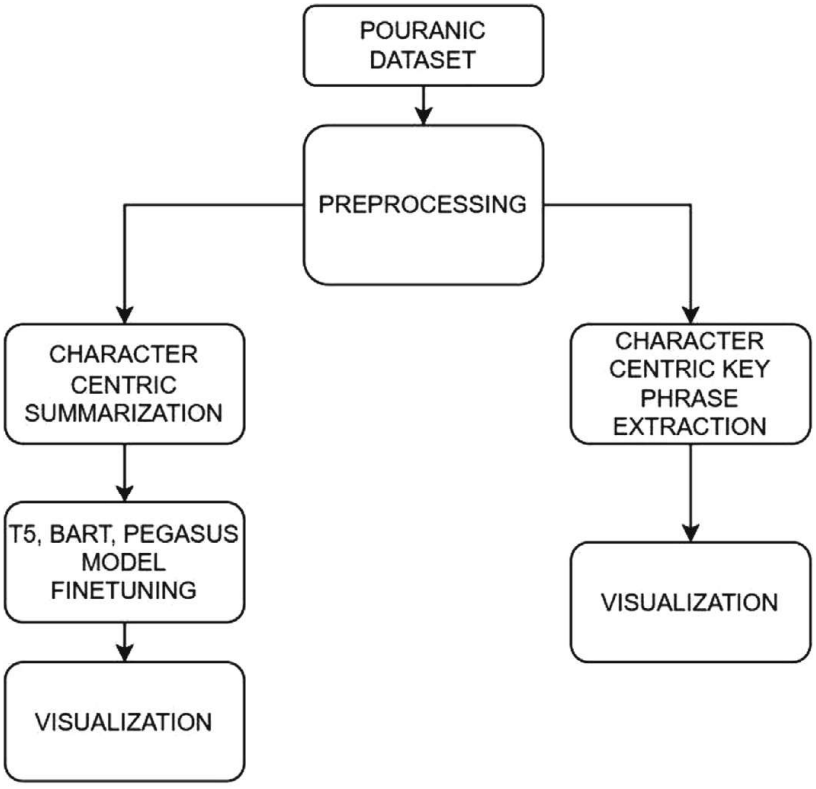

Fig. 1 System framework

- **Model Fine-Tuning:** Models like BART, PEGASUS, and T5, which are pre-trained, undergo fine-tuning on sentences that are specifically related to individual characters. This fine-tuning process enables the models to generate summaries that highlight each character's narrative and attributes. For T5, fine-tuning involves training on character-specific sentences and corresponding summaries, optimizing the model to minimize the loss between generated and reference summaries, with evaluation using ROUGE scores. BART is fine-tuned using a sequence-to-sequence methodology, where input sentences are paired with target summaries, focusing on reconstructing accurate summaries. Evaluation is done using ROUGE metrics and manual checks for contextual consistency. Similarly, PEGASUS is fine-tuned to enhance its summarization capabilities for character-specific content, focusing on generating precise summaries and optimizing its accuracy in the process. Performance is evaluated through both ROUGE scores and qualitative assessments.
- **Summary Generation:** The fine-tuned models generate summaries that capture the essence of each character's actions, attributes, and their role in the narrative of Indian mythology.

4.2 Keyphrase Extraction Using KeyBERT

The second component of the framework is Keyphrase Extraction, which utilizes KeyBERT, a model that extracts significant phrases from text. The steps are as follows:

- **Keyphrase Identification:** KeyBERT, utilizing BERT embeddings to understand contextual meanings, identifies the most relevant key phrases within the character-centric sentences. This model helps in pinpointing the essential phrases associated with each character.
- **Keyphrase Extraction:** KeyBERT extracts key phrases that highlight critical aspects of the character's story and role, focusing on terms and phrases that are essential for understanding their importance in the narrative.

4.3 Visualization

The final component, Visualization, showcases the summarized data and extracted key phrases in an interactive and engaging format. The process includes

- **Summary Display:** Character-centric summaries are presented on a web interface. To improve the recognition, each character's name in the summary is substituted with their corresponding image, helping users better connect the summary to the appropriate character.
- **Keyphrase Visualization:** The extracted key phrases are visualized using methods like word clouds, helping users understand the prominence and relationships of key phrases tied to each character.

5 Result Analysis

5.1 Character-Centric Summarization

The effectiveness of the BART, PEGASUS, and T5 models in producing character-specific summaries was assessed using ROUGE and BLEU scores. The quality of the summaries produced by these models was assessed using the ROUGE metrics. The results of this evaluation are presented in Table 1.

Table 1 illustrates a comparison of the BART, PEGASUS, and T5 models across four evaluation metrics: ROUGE-1, ROUGE-2, ROUGE-L, and BLEU. In the case of ROUGE-1, BART outperforms the other models with a score of 0.76. This shows that T5 is relatively less effective in capturing relevant unigrams. For ROUGE-2, PEGASUS leads with a score of 0.74. T5 follows closely with a score of 0.70,

Table 1 ROUGE scores for character-centric summary

Model	Rouge-1	Rouge-2	Rouge-L	BLEU
T5	0.65	0.70	0.68	0.66
BART	0.76	0.66	0.72	0.77
Pegasus	0.71	0.74	0.75	0.70

while BART scores 0.66, indicating its relative weakness in word matching. In terms of ROUGE-L, PEGASUS scores 0.75. BART follows at 0.72, and T5 scores 0.68, further highlighting PEGASUS's strength in capturing of relevant keyphrases. For the BLEU metric, BART performs with a score of 0.77. PEGASUS scores 0.70, while T5 scores 0.66, which suggests that BART is effective at matching n-grams.

BART shows excellent precision in terms of n-gram matching, particularly with the BLEU score. T5 tends to fall behind slightly across all the evaluation metrics.

5.2 Keyphrase Extraction Evaluation

To extract meaningful key phrases the evaluation was carried out using precision, recall, and F1-score metrics presented in Table 2.

The performance of the KeyBERT model in extracting keyphrases was evaluated through precision, recall, and F1-score metrics. A precision of 0.73 indicates that 73% of the key phrases identified were relevant, showcasing the model's capacity to select meaningful phrases while minimizing irrelevant ones.

The recall of 0.78 shows its capacity to identify a significant proportion of important terms. The F1-score of 0.75 shows the model's strong overall performance. These results collectively suggest that KeyBERT excelled at extracting key phrases, offering valuable insights into the character-centric analysis.

5.3 Visualization

The visualization component of the system aims to improve the interaction and presentation of character-focused summaries and key phrases through a user-friendly web interface. This component includes several important features:

Table 2 Performance metrics of keyphrase extraction model

Precision	Recall	F1-score
0.73	0.78	0.75

Summary Display: The character-specific summaries are prominently featured on the web interface. To demonstrate the visualization of summaries and key phrases, we present images of the web interface. These images highlight the primary features and functions of the visualization system.

In Fig. 2, selecting a character will display the corresponding text lines associated with that character under the **Original Text** section.

The summary produced by the system is shown in Fig. 3 under the **Original Summary** section, where the names of characters in the summary are substituted with their respective images in the following section.

As shown in Fig. 4, the keyphrases associated with the character *Abhimanyu*, including *King Pariksit*, *Abhimanyu*, *Subhadra*, *son of Arjuna*, and *great hero Abhimanyu*, are represented as a word cloud.

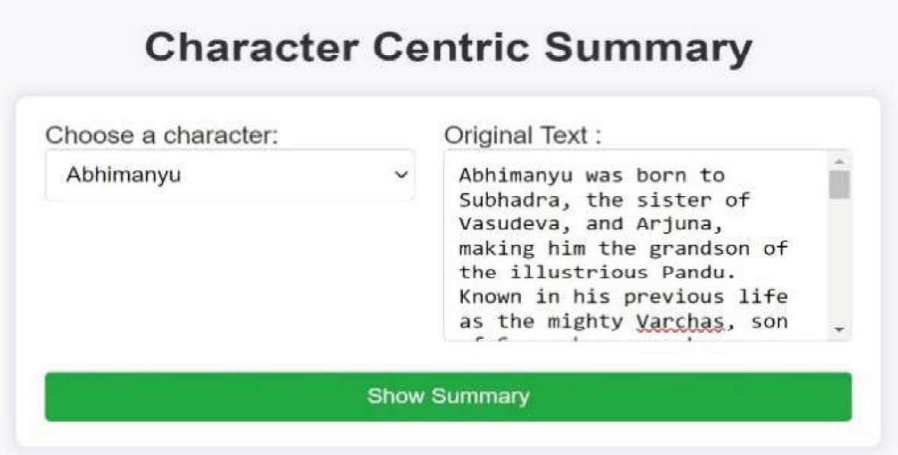

Fig. 2 Character-centric summary interface

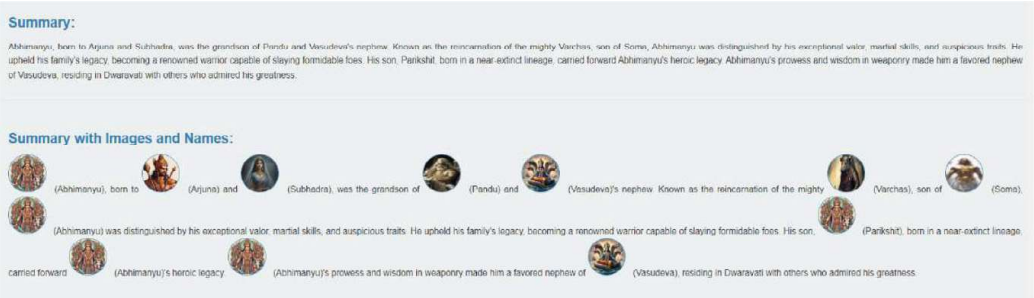

Fig. 3 Character-centric summary output

Fig. 4 Character-centric keyphrases

6 Discussion and Observations

The data highlights the performance of BART, PEGASUS, and T5 across both the ROGUE and BLEU metrics, revealing distinct patterns in their efficiency and effectiveness. T5 shows moderate effectiveness, maintaining relatively balanced scores across all ROGUE metrics and a BLEU score of 0.66, indicating a consistent but not outstanding performance in both precision and recall tasks.

BERT, with a strong BLEU score of 0.77, further confirms its high efficiency in language translation and summarization, though its dip in ROGUE-2 suggests challenges in capturing complex relationships. PEGASUS, while slightly behind BERT in BLEU with a score of 0.70, exhibits superior effectiveness across the ROGUE metrics, particularly excelling in ROGUE-2 and ROGUE-L. This suggests that PEGASUS strikes a strong balance between precision and the ability to capture broader textual relationships, making it the most effective model overall across both performance metrics.

KeyBERT demonstrated varying performance in keyphrase extraction across different summaries. In particular, summary achieved the highest precision (0.73) and recall (0.78), reflecting effective extraction of key phrases. However, there was noticeable variability in performance across other summaries, highlighting areas where the method could be improved to achieve more consistent results.

7 Conclusion and Future Scope

This study analyzes the effectiveness of BART, PEGASUS, and T5 models, trained on the PouranicTopic Dataset, for both summarization and keyphrase extraction tasks. PEGASUS excelled in ROUGE metrics, particularly ROUGE-L, indicating its strength in capturing coherent summaries. BART achieved the highest BLEU score, showing precision but lower bigram performance. T5 performed well in ROUGE-2 but could improve in overall coherence. KeyBERT was effective in extracting key

phrases, though its consistency varied across summaries. Future research could focus on combining model strengths, expanding the dataset for better generalization, and exploring hybrid approaches for improved summarization and keyphrase extraction.

References

1. Raffel, C., Shinn, E., Roberts, A., Lee, K., & Liu, P. J. (2020). *Exploring the limits of transfer learning with a unified text-to-text transformer*. arXiv preprint arXiv:1910.10683.
2. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., & Gouws, S. (2020). *BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension*. arXiv preprint arXiv:1910.13461.
3. Zhang, J., Zhao, Y., & Salehi, M. (2020). *PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization*. arXiv preprint arXiv:1912.08777.
4. Lin, C. Y., & Liu, X. (2019). *BERTSUM: BERT-based extractive summarization*. arXiv preprint arXiv:1904.02340.
5. Liu, Y., & Lapata, M. (2019). *Hierarchical transformer for extractive summarization*. arXiv preprint arXiv:1907.12428.
6. Gao, Y., He, J., & Li, Z. (2020). *SummaRuNNer: A recurrent neural network based sequence model for extractive summarization*. arXiv preprint arXiv:2004.08056.
7. Zhou, W., Zhang, J., & Xu, Y. (2021). *Abstractive document summarization with pre-trained language models*. arXiv preprint arXiv:2102.10920.
8. Gong, Y., & Liu, X. (2021). *Extractive summarization with contrastive learning*. arXiv preprint arXiv:2105.00119.
9. Grootendorst, M. (2020). *KeyBERT: Simple and effective keyphrase extraction with BERT*. arXiv preprint arXiv:2010.11972.
10. Bennani-Smires, K., Musat, C., Hossmann, A., Baeriswyl, M., & Jaggi, M. (2018). *Simple unsupervised keyphrase extraction using sentence embeddings*. arXiv preprint arXiv:1801.04470.
11. Florescu, C., & Caragea, C. (2017). PositionRank: An unsupervised approach to keyphrase extraction from scholarly documents. In *Proceedings of the 55th annual meeting of the association for computational linguistics*.
12. Meng, R., Huang, X., Yuan, J., & Wang, D. (2017). *CopyRNN: A sequence-to-sequence model with copy mechanism for keyphrase generation*. arXiv preprint arXiv:1704.06377.
13. Vaswani, A., Shazeer, N., Parmar, N., & Uszkoreit, J. (2017). *Attention is all you need. Advances in neural information processing systems*.
14. Campos, R., Mangaravite, V., Pasquali, A., Jorge, A. M., Nunes, C., & Jatowt, A. (2020). YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences, 509*, 257–289.
15. Sun, Q., Wang, H., Li, P., Ren, Z., Chen, J., & Ma, J. (2019). *Attention-based neural keyphrase extraction*. arXiv preprint arXiv:1906.09861.
16. Paul, A., Seal, S., & Das, D. (2024). Transformer-based PouranicTopic classification in Indian mythology. *Sadhana, 50*. https://doi.org/10.1007/s12046-024-02555-3