A Web-Based AI System for Summarizing Islamic Historical Biographies Using NLP Techniques

A’lya Nadhirah binti Abdul Hadi1, a) and S Prabha Kumaresan1, 2, b)

1Faculty of Computing and Informatics, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Malaysia.

2Center of Natural Language Processing (NLP), CoE for Artificial Intelligence, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Malaysia.

*b) Corresponding author: prabha.kumaresan@mmu.edu.my*

*a)*[*1221303601@student.mmu.edu.my*](mailto:1221303601@student.mmu.edu.my)

**Abstract.** Text summarization has been in the industry since before the 20th century, yet a significant gap remains in methodologies tailored to the Islamic historical texts domain. These texts feature rich narratives, complex details, and unique vocabulary, which pose unique challenges for summarization. This research addresses this gap by conducting a comparative analysis of four pre-trained Natural Language Processing (NLP) models: two extractive models, Lexical Rank summarization (LexRank) and Heterogeneous Transformer (HETFORMER), and two abstractive models, Bidirectional and Auto-Regressive Transformer (BART) and Pre-training with Extracted Gap-sentences for Abstractive Summarization (PEGASUS). These models are evaluated based on their effectiveness in summarizing Islamic historical biographies while ensuring the preservation of essential entities and thematic coherence. The output of the highest performance model is selected to be the summary texts in a web-based system. By utilizing AI-driven knowledge dissemination, this project seeks to modernize the study of Islamic history, making historical narratives more accessible to scholars, educators, and the general public in today’s age. The results contribute to computational humanities research and AI-assisted historiography, facilitating advancements in the new exploration of the Islamic history domain in text summarization.

# INTRODUCTION

In today’s fast-paced world, people often need quick access to information. Long factual texts that require a considerable amount of time to read and search for important points may become difficult to process. This is where text summarization techniques extract and condense key information, thus significantly saving reading time. This may be evident in several domains like news reporting, medical research, and literary novels.

In educational settings, a campus-based chatbot system is developed to assist students with academic queries using intelligent neural network models [1]. These systems indicate how NLP can handle large-scale and domain-specific information efficiently. The same issue arises with Islamic historical biographies, such as "The First Series" by Yaqeen Institute, where extensive documents make it difficult for users to obtain a brief overview of an episode before deciding whether to watch it.

Despite the advancements in general text summarization methodologies, the major research gap remains in tools specifically designed for the Islamic historical domain. To address this gap, our research conducted a comparison of four different pre-trained NLP models and tested out their performance, and then we developed a specialized web-based system to provide a tailored platform for Islamic biographies summarization.

The proposed system delivers concise and thematic summaries of Islamic historical biographies, thus facilitating quicker access to essential information. By addressing this specific need within the domain, this research not only expands the new domain of text summarization but also makes an important contribution to enhancing educational engagement and knowledge dissemination related to the influential figures in Islamic history.

# RELATED WORKS

## Extractive Summarization Models

Extractive summarization is widely used to extract key sentences, which is useful for summarizing historical biographies where factual accuracy is crucial. This section reviews studies on extractive models and assesses their relevance to Islamic historical texts. One study explored automatic text summarization using Word2Vec embedding combined with k-means clustering, then applied a ranking algorithm to select the most relevant sentences. The model consistently outperformed other extractive methods like OPINOSIS and PyTextRank across various summary lengths [2]. Recent studies have used deep learning to improve the coherence of the text. For example, a study presented a Siamese-BERT model for semantic text-matching, achieving a high ROUGE-1 score on CNN/DailyMail using RoBERTa-base. Although this method improves semantic consistency, it may struggle in fluency and depth of context found in historical narratives [3].

Next, other works have focused on domain-specific optimization [2]. For example, the BERT model was fine-tuned to facilitate the summarization of unstructured call transcripts for higher ROUGE scores [4]. Similarly, TextRank-based clustering showed improved Recall performance in an unsupervised setting [5]. However, inputs like call transcripts have turn-based structures that differ from the chronological and thematic complexities of historical biographies.

Among graph-based ranking methods, TextRank and LexRank have been popular for extractive summarization. TextRank constructs a similarity graph based on sentence overlaps, which potentially summarizes without considering the relevance of topic-related terms. In contrast, LexRank employed TF-IDF weighted vectors and cosine similarity to compute sentence similarity, which helps capture semantically important content while normalizing for sentence length. [6].

A notable advancement in this field is HETFORMER. This pre-trained model utilized multi-granularity sparse attention mechanisms for long-text extractive summarization. In its evaluation on CNN/DailyMail and Multi-News datasets, HETFORMER achieved state-of-the-art performance while being more efficient in memory and parameters than existing models [7]. From the analysis above, LexRank and HETFORMER seem to be appropriate models for summarizing historical texts like Islamic biographies, where efficiency in handling long texts and accurate thematic extraction is important.

## Abstractive Summarization Models

Abstractive summarization generates new sentences rather than extracting them, making it useful for producing more fluent and coherent narratives. However, it may struggle with factual consistency, as abstractive models might generate nonexistent context, which can be challenging when summarizing historical texts.

A comparison was made between T5, BART, and PEGASUS on medical documents and found PEGASUS consistently outperformed other models across all ROUGE metrics [8]. Another approach combined an extractive pre-processing phase with transformer-based abstractive summarization, achieving state-of-the-art ROUGE scores across multiple datasets [9]. These studies indicate that pretraining with domain-specific data significantly improves abstractive summarization performance.

Another study evaluated BART’s encoder-decoder architecture and compared it with BERT, T5, and RoBERTa. They found that BART outperformed its peers in generating coherent summaries, reinforcing its potential as a strong candidate for refining structured summaries into high-quality narratives [10]. Next, a study focusing on intelligent abstractive summarization of scholarly publications introduced GRU-based and Transformer-based models (including ProphetNet, PEGASUS, and BART) for headline generation [11]. Their findings showed that fine-tuned Transformer models, particularly BART, achieved the best ROUGE scores while being computationally efficient.

While models like GPT-3 have been used for large-scale summarization tasks such as book summarization and received high ratings in human evaluations, they come with significant computational costs. These resource demands raise concerns about the practicality of using such models in real-time or large-scale applications, such as summarizing historical biographies across multiple transcripts [12]. Given these considerations, BART and PEGASUS stand out as a balanced solution, offering strong coherence, readability, and computational efficiency. However, for use in historical biography summarization, additional fact-constrained fine-tuning is needed to ensure historical accuracy in the generated outputs.

# Research methodology

## Dataset Description

The dataset from Yaqeen Institute's The Firsts series consists of transcribed lectures that provide in-depth narratives of early Islamic figures [13]. This dataset contains the story of the lives of the Prophet Muhammad’s (SAW) companions, their struggles, contributions, and spiritual journeys. The narratives are presented in an unstructured manner, incorporating historical references from the Quran and Hadith, personal reflections, and scholarly insights, making the dataset highly valuable for researchers and learners alike. One of the defining characteristics of this dataset is its lengthy texts, which contain informal expressions, repetitions, and filler words, making preprocessing essential for extracting meaningful content. For this research, a maximum of ten transcripts from "The Firsts" series were selected due to limited time constraints for manual editing, while ensuring thorough and manageable analysis.

## Data Pre-processing

First, we manually reviewed and edited each transcript to align the texts with the video content. We also corrected the grammatical errors and improved coherence to simplify the subsequent preprocessing steps. We eliminated repetitive words and sentences, Arabic text, and transliterations, while preserving the English translations. These refined transcripts served as the primary input for our summarization pipeline. Next, we removed filler words (e.g., ‘um’, ‘uh’, ‘you know’, etc.) and common repeated phrases (e.g., ‘Subhanallah’, ‘Alhamdulillah’, ‘said’, etc.) that added no meaning to the context. Following this, we performed sentence segmentation to help the models identify key sentences. Each sentence was then tokenized into individual words. Tokenization significantly improves the performance in NLP tasks, especially for texts with complex structures and limited resources [14]. The tokens are then processed to remove stop words and punctuation to minimize noise within the data. Finally, custom topic extraction is built through defining keywords relevant to each topic, ensuring that only the most relevant sentences are selected for summarization. The idea of utilizing keywords to guide content selection is inspired by keyword-based recommendation techniques, which have proven effective in capturing user preferences and personalizing recommendations even within large and diverse datasets [15]. The tokens are then processed to remove stop words and punctuation to minimize noise within the data. Finally, custom topic extraction is built through defining keywords relevant to each topic, ensuring that only the most relevant sentences are selected for summarization. The idea of utilizing keywords to guide content selection is inspired by keyword-based recommendation techniques, which have proven effective in capturing user preferences and personalizing recommendations even within large and diverse datasets [16]. By adopting a similar approach, keywords in this project serve as signals of topic relevance, helping to focus the summarization process on essential biographical themes. The full research methodology is shown in Figure 1.

## Summarization Approach

### Extractive Summarization: LexRank and HETFORMER

To generate extractive summaries based on unsupervised graph-based ranking, we implemented LexRank using the sumy library. Each input file was structured into thematic blocks, each introduced by a custom marker format ‘--- Topic Name ---'. A regular expression parser was developed to segment these blocks. For each identified topic section, we applied LexRank with a fixed output length of 7 sentences per block. The LexRank algorithm builds a similarity graph of sentences and ranks them using eigenvector centrality, allowing for efficient selection of the most representative sentences without requiring training data. The results were stored in an output directory with summaries grouped by topic.

To explore a supervised, context-aware approach to extractive summarization, we employed a BERT-based model (bert-base-uncased) to simulate the HetFormer framework. Each topic and its corresponding text block were processed into (topic, sentence) pairs to allow the model to score sentences for relevance. We tokenized the pairs using the Hugging Face transformers library and passed them through the BERT model to obtain class logits. The scores for the "positive" class were interpreted as relevance probabilities. We then selected up to 3 top-scoring to construct concise summaries. This method emphasizes semantic alignment between the topic and candidate sentences. Summaries were stored similarly to the LexRank output for comparison. All extractive summarization experiments were conducted using pre-trained models without additional fine-tuning.

### Abstractive Summarization: BART and PEGASUS

To explore abstractive summarization, we employed two transformer-based models: BART (facebook/bart-large-cnn) and PEGASUS (google/pegasus-cnn\_dailymail). Both models are pre-trained on large corpora and fine-tuned for summarization tasks, enabling them to generate coherent summaries that may go beyond directly copying from the source text.

**Manual correction**

Remove repetitive words and sentences, Arabic texts and transliterations, verify it tallies with video

**Preprocessing**

Remove filler words and redundant phrases, sentence segmentation, word tokenization, stop words, and punctuation removal

**Custom Topic Extraction**

Define relevant keywords for each topic

**Model Comparison**

LexRank and HETFORMER for extractive summarization. BART and PEGASUS for abstractive summarization.

**Model Evaluation**

Evaluate models using BERTScore

BERTScore > 0.8

**Website Development**

Using Django. Summaries with the best score is implemented into the website

Does not meet threshold

Meet threshold

Yes

No

**Testing and Refinement**

Test summarization output and website functionality

0

Error

**Testing and Refinement**

Test summarization output and website functionality

**FIGURE 1.** Research methodology flowchart

To make the summaries more focused and context-aware, we introduced a keyword conditioning mechanism. For each topic block, segmented using the same regex-based method as in other pipelines, we generated a short prompt that included the topic name and a set of curated keywords relevant to that topic (e.g., “tribe” for lineage content, “plague” and “martyr” for death-related content). This prompt was prepended to the source content before feeding it to the model. This technique provided soft guidance to the model, encouraging it to attend to the semantic core of each topic.

The input length for both models was truncated to 1024 tokens per block due to model constraints. No fine-tuning was applied. Summaries were generated using the Hugging Face Transformers pipeline and stored in output directories for evaluation.

### Evaluation Metrics

To quantitatively assess the quality of summaries generated by the four models - LexRank, HetFormer (BERT), BART, and PEGASUS - BERTScore, a semantic similarity metric designed for evaluating natural language generation tasks, is used. Unlike traditional string-matching metrics like ROUGE, BERTScore compares token embeddings generated by a pretrained BERT model, allowing it to capture deeper contextual and semantic alignment between the predicted summaries and reference (gold) summaries [16]. The gold summaries were generated with the assistance of ChatGPT and then manually refined to improve clarity, tone, and flow.

This approach ensured that the final summaries were both consistent and human-like, preserving fidelity to the original content while enhancing readability. Each gold summary corresponds to a topic block in the original text, allowing for precise alignment during evaluation. The evaluation script reads all model-generated summaries and their corresponding gold summaries, matching them by filename. BERTScore is computed for Precision (Equation (1)), Recall (Equation (2)), and F1-score (Equation (3)), and these values are averaged across all files for each method. This provides a reliable, meaning-focused measure of how well each model performs at capturing the key information in a human-like way. Should the summarization outcomes meet predetermined quality thresholds, which is 0.8, the project advances to the development of a web platform using Django. If the evaluation criteria are not satisfied, iterative refinement and reassessment of the summarization methods are undertaken.

### Website Development

The ultimate deliverables of this project include a summarized biographical content and a functional website designed to enhance user engagement and facilitate efficient access to Islamic historical biographies, using Django as the backend framework, along with Python, HTML, and CSS for logic and presentation. The website features a home page that displays a list of available episodes for users to explore (Figure 2). Each episode links to a dedicated episode page, which includes the YouTube link to watch the original video, the cleaned long-form transcript, a "Summarize" button to generate the summary, and a download button for saving the transcript as a PDF (Figure 3).

Once the summarization process is triggered, a pop-up summary page is displayed, presenting the generated summary using the model with the highest BERTScore. This summary page also includes a download button, allowing users to save the summarized version as a PDF for offline access (Figure 4).

A screenshot of a computer

AI-generated content may be incorrect.

**FIGURE 2.** Home page

A screenshot of a computer

AI-generated content may be incorrect.

**FIGURE 3.** Episode page

# Results and discussion

Among all models, BART achieved the highest performance across all three metrics, with a precision of 0.841, a recall of 0.836, and an F1 score of 0.839 as illustrated in Table 1. This indicates that BART produces summaries that are not only semantically similar to the gold references but also highly consistent in identifying key information. The results support the effectiveness of BART’s encoder-decoder architecture and its fine-tuning on summarization tasks.

PEGASUS followed closely with an F1 score of 0.829, demonstrating strong semantic retention, especially in recall (0.839), which aligns with its design for abstractive summarization through sentence masking. HetFormer and LexRank, while performing slightly lower, still provided reasonable outputs. HetFormer (a supervised extractive approach) marginally outperformed LexRank (an unsupervised extractive model), particularly in recall and F1.

The gap between extractive models (LexRank and HetFormer) and abstractive models (BART and PEGASUS) highlights the advantage of leveraging pretrained sequence-to-sequence architectures for generating coherent, human-like summaries. It also validates the use of BERTScore, which prioritizes semantic similarity over surface-level matching, as a suitable metric for evaluating such outputs.

Based on these results, BART was selected as the default summarization model in the deployed web application, offering the best trade-off between informativeness, fluency, and faithfulness to the original content.

A screenshot of a computer

AI-generated content may be incorrect.

**FIGURE 4.** Summary pop-up

|  |  |  |  |
| --- | --- | --- | --- |
| **TABLE 1.** Precision, Recall, and F1 measures of BERTScore, underlined and bolded are the top-performing models in each category | | | |
|  | **BERTScore** | | |
| **Summarization Models** | **Precision** | **Recall** | **F1** |
| LexRank | 0.822 | 0.822 | 0.822 |
| HETFORMER | 0.826 | 0.827 | 0.826 |
| BART | **0.841** | 0.836 | **0.839** |
| PEGASUS | 0.827 | **0.839** | 0.829 |

# CONCLUSION

In conclusion, this study explored the application of both extractive and abstractive summarization techniques for Islamic historical biographies, focusing on content derived from transcribed video lectures. We evaluated four pre-trained models: LexRank, HETFORMER (BERT-based), BART, and PEGASUS, while using BERTScore as the evaluation metric. Among these, BART achieved the highest performance in Precision and F1, making it the most effective choice for producing coherent, meaningful summaries in this domain. The findings highlight the potential of pre-trained abstractive summarization models in delivering concise and informative overviews of Islamic historical content, especially for users seeking quick access to essential information. This has practical value for learners, educators, and researchers, particularly in note-taking and reviewing lengthy biographical material. Additionally, by transforming spoken content into structured, readable text, the system can support accessibility for users with hearing impairments. The success of this approach demonstrates how NLP can be leveraged to enhance the usability of religious and historical resources. In future work, researchers could focus on fine-tuning summarization models on Islamic historical datasets to improve contextual understanding and factual consistency. Moreover, interactive features could be introduced in the web-based system to enhance the learning experience and encourage users' discussion, such as 'Notes' and 'Reflections' sections, where users can write their insights about an episode and engage with thoughts shared by others. In conclusion, this project marks the first step in the interdisciplinary study of how AI-driven summarization can modernize engagement with Islamic history, and through the website developed, it helps in content navigation efficiency and improves user experience.

# References

1. T.-J. Goh, L.-Y. Chong, S.-C. Chong, and P.-Y. Goh, “A Campus-based Chatbot System using Natural Language Processing and Neural Network,” JIWE **3**(1), 96–116 (2024).
2. M. Mohd, R. Jan, and M. Shah, “Text document summarization using word embedding,” Expert Systems with Applications **143**, 112958 (2020).
3. M. Zhong, P. Liu, Y. Chen, D. Wang, X. Qiu, and X. Huang, “Extractive Summarization as Text Matching,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (Association for Computational Linguistics, Online, 2020), pp. 6197–6208.
4. P.K. Biswas, and A. Iakubovich, “Extractive Summarization of Call Transcripts,” IEEE Access **10**, 119826–119840 (2022).
5. K. Patil, M. Badamikar, and S. Sonawane, “NLP based Text Summarization of Fintech RFPs,” in *2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, (IEEE, Erode, India, 2023), pp. 865–869.
6. M. Okumura, and H. Nanba, “Automated Text Summarization: A Survey,” Journal of Natural Language Processing **6**(6), 1–26 (1999).
7. Y. Liu, J. Zhang, Y. Wan, C. Xia, L. He, and P. Yu, “HETFORMER: Heterogeneous Transformer with Sparse Attention for Long-Text Extractive Summarization,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, (Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021), pp. 146–154.
8. E. Lalitha, K. Ramani, D. Shahida, E.V.S. Deepak, M.H. Bindu, and D. Shaikshavali, “Text Summarization of Medical Documents using Abstractive Techniques,” in *2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, (IEEE, Salem, India, 2023), pp. 939–943.
9. J. Pilault, R. Li, S. Subramanian, and C. Pal, “On Extractive and Abstractive Neural Document Summarization with Transformer Language Models,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Association for Computational Linguistics, Online, 2020), pp. 9308–9319.
10. A. Venkataramana, K. Srividya, and R. Cristin, “Abstractive Text Summarization Using BART,” in *2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon)*, (IEEE, Mysuru, India, 2022), pp. 1–6.
11. F. Zaman, M. Afzal, P.S. Teh, R. Sarwar, F. Kamiran, N.R.A. Aljohani, R. Nawaz, M.U. Hassan, and F. Sabah, “Intelligent Abstractive Summarization of Scholarly Publications with Transfer Learning,” Journal of Informatics and Web Engineering  **3**(3), 256–270 (2024).
12. A. Lukina, “Combining Runtime Monitoring and Machine Learning with Human Feedback,” AAAI **37**(13), 15448–15448 (2023).
13. J. Steinhauer, *The Firsts: The inside Story of the Women Reshaping Congress*, First edition (Algonquin Books of Chapel Hill, Chapel Hill, North Carolina, 2020).
14. A. Khan, K. Khan, W. Khan, S.N. Khan, and R. Haq, “Knowledge-based Word Tokenization System for Urdu,” Journal of Informatics and Web Engineering **3**(2), 86–97 (2024).
15. A.A. Zaveri, R. Mashood, S. Shehmir, M. Parveen, N. Sami, and M. Nazar, “AIRA: An Intelligent Recommendation Agent Application for Movies,” JIWE **2**(2), 72–89 (2023).
16. School of Computing and Communications, Lancaster University, Lancaster, United Kingdom, P. Saxena, M. El-Haj, and School of Computing and Communications, Lancaster University, Lancaster, United Kingdom, “Exploring Abstractive Text Summarisation for Podcasts: A Comparative Study of BART and T5 Models,” in *Proceedings of the Conference Recent Advances in Natural Language Processing - Large Language Models for Natural Language Processings*, (INCOMA Ltd., Shoumen, BULGARIA, 2023), pp. 1023–1033.