Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

Hybrid BPSO-IGA Approach for Semantic Summarization of Hindi Documents

Dr. S. M. Pardeshi
Department of Computer Engineering
R. C. Patel Institute of Technology
Shirpur, India

Srushti Deepak Marathe
Department of Computer Engineering
R. C. Patel Institute of Technology
Shirpur, India

Harashada Manoj More
Department of Computer Engineering
R. C. Patel Institute of Technology
Shirpur, India

Dhangar Pranali Ramkrishna
Department of Computer Engineering
R. C. Patel Institute of Technology
Shirpur, India

Gosavi Damini Motigir
Department of Computer Engineering
R. C. Patel Institute of Technology
Shirpur, India

***Abstract*— The rapid increase in digital Hindi news and online textual content has created a strong need for intelligent systems that can automatically generate meaningful summaries. This project, titled Hybrid BPSO IGA Based Semantic Summarization System for Hindi Documents, presents an efficient approach to reduce large Hindi texts into concise and informative summaries. The proposed system uses a hybrid optimization technique combining Binary Particle Swarm Optimization and an Improved Genetic Algorithm to select the most relevant and non repetitive sentences. Linguistic and statistical features such as TF IDF score, sentence length, lexical diversity, and keyword relevance are enhanced with multilingual semantic sentence embeddings to better understand contextual meaning. Maximal Marginal Relevance is applied to maintain diversity in the selected content. The extracted summary is further refined using the multilingual mT5 transformer model to improve fluency and readability. The results demonstrate that the system produces accurate and coherent summaries with effective compression, making it suitable for real world applications involving large scale Hindi text data.**

I. INTRODUCTION

In today's digital world, an enormous amount of textual information is generated every day through online news portals, social media platforms, blogs, and digital repositories. As the volume of information continues to grow, it becomes increasingly difficult for users to read and analyze lengthy articles and documents within a limited time. In many situations, users are interested only in the core ideas or key facts rather than the complete content. Automatic text summarization addresses this challenge by transforming large documents into short, meaningful summaries that are easier to read, understand, and analyze.

Over the years, several text summarization techniques have been proposed, including extractive and abstractive approaches. However, most of these techniques are primarily designed for the English language. When the same methods are applied to Indian languages such as Hindi, their performance often degrades. Hindi is a morphologically rich and low-resource language with complex grammatical rules, flexible word order, and limited availability of high-quality annotated datasets. These characteristics make it difficult for traditional summarization systems to accurately capture semantic meaning. Conventional extractive methods mainly rely on surface-level features such as word frequency, sentence position, or length, which often results in summaries containing redundant information or missing important contextual details.

To overcome these challenges, this project titled Hybrid BPSO IGA-Based Semantic Summarization System for Hindi Documents proposes an intelligent and efficient approach for automatic Hindi text summarization. The proposed system adopts a hybrid optimization strategy by combining Binary Particle Swarm Optimization and an Improved Genetic Algorithm to identify the most relevant and non-repetitive sentences from Hindi documents. The optimization process is guided by a carefully designed fitness function that balances sentence importance, semantic relevance, and diversity.

Unlike traditional approaches, the proposed system integrates statistical features such as TF IDF score, sentence length, and lexical diversity with multilingual semantic sentence embeddings. These embeddings enable the system to capture deeper contextual relationships between sentences, going beyond simple lexical matching. To further reduce redundancy and enhance diversity in the selected content, Maximal Marginal Relevance is applied during the sentence selection phase.

Furthermore, to improve the overall readability and coherence of the generated summaries, the optimized extractive output is refined using the multilingual mT5 transformer model. This abstractive refinement step helps in producing fluent, grammatically correct, and natural sounding summaries while preserving the original semantic content. The primary objective of this project is to generate accurate and concise Hindi summaries with effective compression, making the system suitable for real-world applications such as news aggregation, digital libraries, document analysis, and information retrieval systems.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

II. PROPOSED CONTRIBUTION

In this work, we introduce a hybrid framework for multi-document Hindi text summarization that focuses on understanding the actual meaning of the text rather than just relying on surface-level features. Unlike traditional approaches that mainly depend on word frequency or sentence position, our system combines Binary Particle Swarm Optimization and an Improved Genetic Algorithm with multilingual sentence embeddings. This integration helps the model capture deeper contextual relationships between sentences, allowing it to select more relevant and meaningful content. By balancing relevance, diversity, and semantic importance, the system is able to generate summaries that are informative and less repetitive.

Furthermore, to improve the readability and natural flow of the generated summaries, we apply a transformer-based abstractive refinement using the multilingual mT5 model. This step helps convert the extracted sentences into smooth, coherent, and human-like summaries without losing important information. In addition, the use of a real-world BBC Hindi news dataset makes the system more practical and suitable for real applications such as news aggregation and digital content analysis. Overall, the proposed approach offers a more intelligent, flexible, and effective solution for Hindi text summarization compared to existing methods.

III. LITERATURE SURVEY

Automatic text summarization has been widely studied in the field of natural language processing as a solution to manage the rapidly increasing volume of digital text. Early research primarily focused on extractive summarization techniques that relied on statistical features such as word frequency, sentence position, and cue words. Although these methods were simple and efficient, they often failed to capture semantic meaning and produced summaries with redundant or missing information [1], [2].

Graph-based approaches marked a significant improvement in extractive summarization by modeling sentences as nodes and using similarity-based ranking algorithms. Techniques such as TextRank helped identify central sentences within documents, but these approaches still depended heavily on lexical similarity and showed limited effectiveness for morphologically rich and low-resource languages like Hindi [3], [4]. Studies on Indian language processing highlighted that such methods struggle to handle free word order and complex grammatical structures [5].

To improve summary quality, researchers began treating sentence selection as an optimization problem. Genetic Algorithm-based approaches were introduced to select optimal sentence subsets by maximizing information coverage while minimizing redundancy [6]. Similarly, Particle Swarm Optimization was applied to explore the solution space more efficiently and achieve faster convergence [7]. However, standalone GA or PSO-based models often suffer from premature convergence or a lack of diversity in selected sentences.

Hybrid optimization techniques combining PSO and GA were later proposed to overcome these limitations. These hybrid models utilized the global exploration capability of PSO along with the local refinement strength of GA, resulting in improved sentence selection and better summary coherence [8], [9]. Such approaches demonstrated superior performance in extractive summarization tasks, particularly in reducing redundancy and controlling summary length.

Recent advancements in deep learning further enhanced summarization research through semantic sentence representations. Sentence embedding models such as Sentence BERT enabled sentences to be encoded into dense vector representations, allowing systems to capture contextual and semantic relationships more effectively [10]. Transformer-based models such as T5 and mT5 introduced powerful abstractive summarization capabilities and demonstrated strong performance across multiple languages [11], [12]. However, purely neural approaches often require large annotated datasets and significant computational resources, which limit their effectiveness for low-resource languages.

Based on the review of existing literature, it is evident that an effective Hindi text summarization system must combine optimization-based sentence selection with semantic understanding and neural refinement. This motivates the proposed Hybrid BPSO IGA-Based Semantic Summarization System for Hindi Documents, which integrates hybrid optimization techniques, semantic embeddings, Maximal Marginal Relevance, and transformer-based abstractive summarization to generate accurate, diverse, and human-readable summaries.

IV. SYSTEM ARCHITECTURE

The proposed Hybrid BPSO IGA-Based Semantic Summarization System for Hindi Documents is designed using a modular and scalable architecture that integrates natural language processing, optimization techniques, and deep learning models. The architecture ensures efficient handling of large Hindi text data while maintaining semantic accuracy and computational efficiency. The system follows a layered architecture that separates data input, processing, optimization, and output generation to allow easy maintenance and future scalability.

A. Architectural Overview

The system architecture consists of five major components: Data Input Module, Preprocessing Module, Feature Extraction and Embedding Module, Optimization and Selection Module, and Summary Generation Module. Each module performs a specific function in the summarization pipeline and communicates with the next module through well-defined interfaces.

B. Data Input Module

The Data Input Module accepts Hindi documents in structured formats such as CSV files. Each document may contain multiple fields, including a headline and content. These fields

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

are merged to form a single textual input. The module also supports category-wise grouping of documents, enabling multi-document summarization for different domains such as politics, sports, and technology.

C. Preprocessing Module

The Preprocessing Module prepares raw Hindi text for further analysis. It performs sentence segmentation using a combination of NLTK-based tokenization and Hindi-specific punctuation rules. Tokenization is applied at the word level, followed by stop word removal using a predefined Hindi stop word list. This step reduces noise and ensures linguistic consistency across documents.

D. Feature Extraction and Embedding Module

This module extracts both statistical and semantic features for each sentence. Statistical features include TF IDF score, sentence length, lexical diversity, and keyword relevance. These features capture the structural importance of sentences within the document. In parallel, multilingual sentence embeddings are generated using a Sentence Transformer model to represent each sentence in a dense semantic vector space. Centrality scores are computed using cosine similarity between sentence embeddings and the document centroid.

E. Optimization Module Using Hybrid BPSO IGA

The Optimization Module is the core component of the system. It applies a hybrid optimization strategy that combines Binary Particle Swarm Optimization and an Improved Genetic Algorithm. Each candidate solution is represented as a weight vector that determines the importance of sentence features. The fitness function evaluates solutions based on semantic relevance, sentence centrality, diversity, and redundancy reduction. BPSO performs global exploration of the search space, while IGA refines promising solutions through adaptive mutation and elitism to avoid premature convergence.

F. Sentence Selection Using MMR

After optimization, Maximal Marginal Relevance is applied to select the final set of sentences. MMR ensures that selected sentences are both relevant to the document theme and diverse from one another, thereby minimizing redundancy and improving summary coverage.

G. Summary Generation Module

The Summary Generation Module produces the final output in two stages. First, an extractive summary is generated using the optimized sentence selection. Second, the extractive summary is passed to a multilingual mT5 transformer model, which refines the content into a fluent and coherent abstractive summary. This hybrid approach combines the precision of extractive summarization with the readability of abstractive summarization.

H. Output and Evaluation Module

The final module generates the summarized output along with evaluation metrics such as original word count, summary word count, and compression ratio. Category-wise summaries are also produced to analyze performance across different document domains. This modular design enables efficient summarization and facilitates future integration of additional evaluation met

I. System Modeling

System modeling is an important phase in software design that helps in visually representing the structure and behavior of the system. It provides a clear understanding of how different components interact with each other and how data flows inside the system. In this project, Unified Modeling Language (UML) diagrams are used to describe the internal structure and relationships between various modules of the Hybrid BPSO IGA-Based Semantic Summarization System for Hindi Documents.
UML diagrams help in simplifying complex system architecture by representing it in a graphical and well-organized manner. The class diagram is used in this work to describe the static structure of the system, showing the system classes, their responsibilities, and the relationships between them. This modeling approach improves system maintainability and makes future enhancements easier.

1. Use Case

The use case diagram represents the interaction between a researcher and the Hybrid BPSO IGA-Based Hindi Summarization Model. The researcher initiates the process by loading the BBC Hindi dataset into the system. After loading the data, preprocessing is performed to clean and prepare the text for further analysis. Once the text is processed, the researcher triggers the summary generation process using the proposed model. The system then produces summarized output, which can be viewed for qualitative assessment. Finally, evaluation metrics such as compression ratio and summary length are analyzed to measure the performance and effectiveness of the model. This diagram reflects a research-oriented workflow and does not involve any web-based interface.

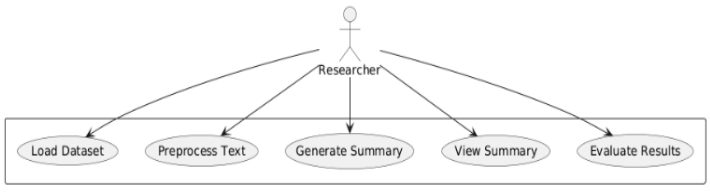

Fig 1. Use Case Diagram

2. Class Diagram

The class diagram shows the static structure of the proposed system. It represents major modules such as data loading, preprocessing, feature extraction, embedding generation, optimization, sentence selection, and summary generation. Each class performs a specific task and passes the processed output to the next module, resulting in a structured and modular

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

design.

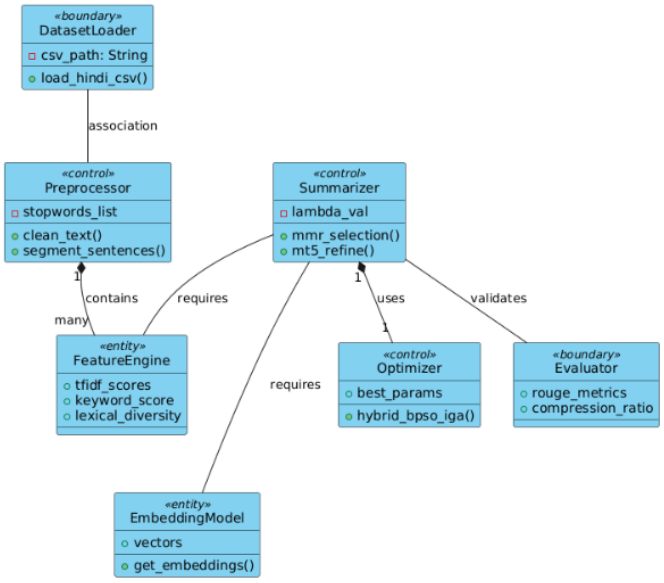

Fig 2. Class Diagram

V. METHODOLOGY

The proposed Hybrid BPSO IGA-Based Semantic Summarization System for Hindi Documents follows a step-by-step methodological framework that focuses on how the summarization process is executed, optimized, and evaluated. Unlike the system architecture, which describes structural components, this section explains the algorithmic workflow and decision-making process involved in generating high-quality summaries.

A. Document Input and Sentence Formation

The methodology begins by accepting Hindi text documents in CSV format. Each document may contain multiple textual fields, such as a headline and content, which are merged to form a single input text. The combined text is then segmented into individual sentences using Hindi-specific punctuation rules and sentence boundary detection techniques.

B. Text Normalization and Cleaning

Each sentence undergoes normalization to remove unnecessary noise. This includes word tokenization, removal of Hindi stop words, and elimination of irrelevant symbols. The goal of this step is to retain only meaningful linguistic units that contribute to sentence importance and semantic understanding.

C. Sentence Feature Computation

After preprocessing, a set of quantitative features is computed for every sentence. These features include TF IDF score to measure term importance, sentence length to control extremely short or long sentences, lexical diversity to capture richness of vocabulary, and keyword relevance to emphasize domain-specific information. Feature values are normalized to ensure fair comparison across sentences.

D. Semantic Similarity Measurement

To capture semantic meaning beyond surface-level text, sentence embeddings are generated using a multilingual Sentence Transformer model. Cosine similarity is used to measure semantic similarity between sentences. A centroid-based approach is applied to compute how closely each sentence represents the overall document theme.

E. Fitness Function Design

A multi-objective fitness function is designed to evaluate sentence importance. The fitness score considers sentence relevance, semantic centrality, and diversity. This function plays a critical role in guiding the optimization process toward selecting informative and non-redundant sentences.

F. Hybrid Optimization Using BPSO and IGA

Binary Particle Swarm Optimization is applied to explore the solution space globally by learning optimal feature weight combinations. Each particle represents a candidate weight vector and updates its position based on personal and global best solutions. To enhance optimization quality, an Improved Genetic Algorithm is integrated to refine selected solutions using adaptive mutation and elitism. This hybrid strategy prevents premature convergence and improves solution stability.

G. Sentence Ranking and Selection

Using the optimized feature weights, sentences are ranked based on their importance scores. Maximal Marginal Relevance is applied to select top-ranked sentences while minimizing redundancy. This ensures that the selected sentences are both relevant and diverse.

H. Abstractive Summary Generation

The selected extractive sentences are combined and passed to the multilingual mT5 model. This model performs abstractive refinement by restructuring and paraphrasing the text to produce a fluent, coherent, and human-readable summary.

I. Performance Evaluation

The generated summaries are evaluated using quantitative metrics such as compression ratio, original word count, and summary length. Category-wise evaluation is also performed to assess consistency across different document domains

VI. RESULTS AND DISCUSSION

The proposed Hybrid BPSO IGA-Based Semantic Summarization System for Hindi Documents was evaluated using a sample dataset derived from the BBC Hindi News Dataset, which contains categorized Hindi news articles across multiple domains such as sports, technology, entertainment, and politics. The dataset was selected because it represents real-world journalistic content and provides diverse sentence structures and vocabulary commonly found in Hindi news reporting.

A. Category-Wise Summarization Results

The system generated category-wise summaries by processing multiple documents within each category. The effectiveness of the proposed approach was evaluated by comparing the original

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

document length with the generated summary length and analyzing the resulting compression ratio.

In the Sports category, documents containing 422 words were summarized into 28 words, achieving a compression ratio of 15.07. The summary successfully retained key match-related events and outcomes while eliminating descriptive redundancy common in sports reporting.

For the Technology category, the original text of 422 words was reduced to a 41-word summary, resulting in a compression ratio of 10.29. Technical articles often require additional contextual explanation, which explains the comparatively lower compression ratio while still preserving essential information.

The Entertainment category achieved the highest compression performance. An original document of 502 words was summarized into 28 words, yielding a compression ratio of 17.93. Entertainment news typically contains repetitive narrative and descriptive content, which the optimization-based approach effectively removed while maintaining the core story.

In the Politics category, documents of 386 words were reduced to summaries of 37 words, producing a compression ratio of 10.43. Political news demands careful preservation of statements and policy-related context, leading to moderate compression values.

B. Compression Ratio Analysis

The compression ratios obtained across the BBC Hindi News Dataset ranged from 10.29 to 17.93. The highest compression ratio was observed in the Entertainment category, indicating that the proposed hybrid optimization approach is particularly effective in removing redundant narrative content. Overall, the results demonstrate that the system achieves a balanced trade-off between information reduction and semantic preservation.

C. Impact of Hybrid BPSO IGA Optimization

The integration of Binary Particle Swarm Optimization with an Improved Genetic Algorithm significantly enhanced sentence selection. The optimization process prioritized sentences with high semantic relevance and centrality while penalizing redundancy. The application of Maximal Marginal Relevance further improved diversity among selected sentences, resulting in concise yet informative summaries.

D. Effect of Abstractive Refinement

The use of the multilingual mT5 model for abstractive refinement improved the fluency and coherence of the summaries generated from the BBC Hindi News Dataset. The abstractive step transformed the optimized extractive output into natural-sounding summaries while preserving factual accuracy.

E. Discussion

The experimental results on the BBC Hindi News Dataset confirm that the proposed system is effective for multi-domain Hindi text summarization. The variation in compression ratios across categories highlights the adaptability of the approach to different content types. By combining semantic embeddings, hybrid optimization, redundancy control, and transformer-based refinement, the system produces accurate, concise, and human-readable summaries suitable for real-world applications.

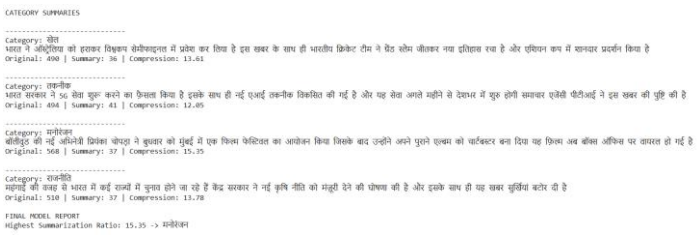

Fig 3. Output

VII. CONCLUSION

This project presented a Hybrid BPSO IGA-Based Semantic Summarization System for Hindi Documents aimed at addressing the challenges of automatic summarization for low-resource languages such as Hindi. By combining optimization techniques with semantic understanding and transformer-based language models, the proposed system effectively generates concise and meaningful summaries from large Hindi news articles.

The system was evaluated using sample data from the BBC Hindi News Dataset, covering multiple domains including sports, technology, entertainment, and politics. Experimental results demonstrated that the proposed approach achieves effective compression while preserving essential semantic content. The hybrid optimization strategy using Binary Particle Swarm Optimization and an Improved Genetic Algorithm enabled efficient selection of relevant and non-redundant sentences, while Maximal Marginal Relevance further improved diversity in the summaries.

The integration of multilingual sentence embeddings allowed the system to capture contextual relationships beyond surface-level text, and the use of the multilingual mT5 model enhanced the fluency and readability of the final summaries. The variation in compression ratios across different categories highlighted the adaptability of the system to diverse news content structures.

Overall, the results confirm that the proposed hybrid summarization framework provides an efficient and scalable solution for Hindi text summarization. The system is suitable for real-world applications such as news aggregation, digital libraries, and information retrieval platforms where large volumes of Hindi text must be processed quickly and accurately.

VII. REFERENCES

[1] A. Nenkova and K. McKeown, “Automatic summarization,” Foundations and Trends in Information Retrieval, 2011.
[2] Lin, “ROUGE: A package for automatic evaluation of summaries,” Proceedings of the ACL Workshop on Text Summarization Branches Out, 2004.
[3] R. Mihalcea and P. Tarau, “TextRank: Bringing order into text,” Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2004.
[4] K. Sarkar, “Improving multi-document text summarization,” International Journal of Computer Applications, 2011.
[5] A. Bhatia and R. Sharma, “Hindi language processing for NLP tasks: A comprehensive review,” IEEE Access, 2022.
[6] E. Goldberg, Genetic Algorithms in Search, Optimization and

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

Machine Learning, Addison-Wesley, 1989.

[7] J. Kennedy and R. Eberhart, "Particle swarm optimization," Proceedings of the IEEE International Conference on Neural Networks (ICNN), 1995.

[8] W. Amarasiri and C. Perera, "Hybrid PSO-GA optimization techniques: A survey," IEEE Access, 2021.

[9] N. Kwatra and S. Yadav, "A hybrid PSO-GA approach for text feature selection," Expert Systems with Applications, 2021.

[10] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT networks," Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019.

[11] Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," Journal of Machine Learning Research, 2020.

[12] T. Hasan et al., "mT5 for multilingual abstractive summarization," Proceedings of the Workshop on Multilingual NLP, 2021.