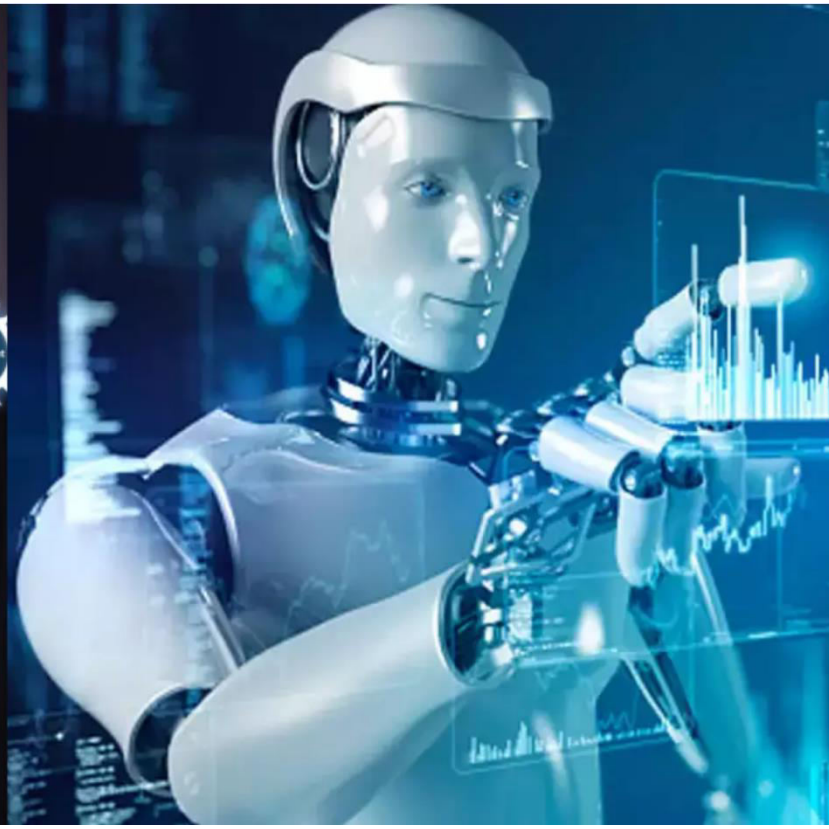
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 15, Issue 3, March 2026

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503103|

Intelligent Research Paper Summarizer Using ML and NLP

Mr. M Hari Krishna[1], I Sai Teertha Sri[2], V Tejaswini[3], G Bhanu Prakash[4], P Eashwar[5]

Assistant Professor, Department of CSE (Data Science), ACE Engineering College, Hyderabad, Telangana, India[1]

IV B. Tech Students, Department of CSE (Data Science), ACE Engineering College, Hyderabad, Telangana, India[2,3,4,5]

ABSTRACT: With the increase in the number of academic publications, it has become hard to understand research papers due to the complexity of the language and the use of terminology specific to the field. The current summarization systems provide only one type of summary, which may not be suitable for different levels of users. The proposed project will be an Intelligent Research Paper Summarizer, which will be able to generate different summaries for different levels of users, such as school, undergraduate, and experts. The system will take a research paper as a PDF file, process it, and split it into logical sections for processing. The readability of each section will be determined by the machine learning-based readability classifier. Then, the readability of each section will be analysed by the pre-trained Transformer model, and the summary will be generated for each section based on the audience type.

KEYWORDS: Automatic Text Summarization, Readability Classification, Transformer-based NLP, Machine Learning for Text Analysis, Multi-level Research Paper Summarization.

I. INTRODUCTION

Understanding research papers has become increasingly challenging due to the rapid growth of academic publications and the use of complex technical language. Many research papers contain dense structures and domain-specific terminology, which makes them difficult to understand for readers with different levels of expertise such as school students, undergraduate learners, and non-specialist readers. Most existing summarization tools generate only a single generic summary that does not consider the reader's level of understanding. As a result, these summaries may be too complex for beginners while lacking sufficient detail for advanced readers. This creates a gap in accessibility and makes it harder for many people to effectively understand academic literature.

To address this issue, this paper proposes an Intelligent Research Paper Summarizer that uses Machine Learning and Natural Language Processing (NLP) techniques to generate summaries tailored to different levels of readers. The system accepts research papers in PDF format, extracts and cleans the textual content and divides the document into meaningful sections for analysis. Various linguistic and statistical features are then extracted and used by machine learning models to predict the readability level of each section. Based on this analysis, a Transformer-based model generates separate summaries for school-level, undergraduate-level, and expert-level readers. This approach makes research papers more accessible and easier to understand while still preserving the key ideas of the original document.

II. LITERATURE SURVEY

Research paper summarization has gained significant attention due to the rapid growth of academic publications. Early approaches mainly focused on extractive summarization methods, where important sentences are selected from the original document. Luhn [1] introduced one of the earliest techniques using word frequency and keyword importance to identify significant sentences. Later, Edmundson [2] improved this approach by introducing features such as cue words, title words, and sentence position to determine sentence importance. These methods formed the foundation for many traditional extractive summarization systems.

With the advancement of machine learning and Natural Language Processing (NLP), more advanced techniques were developed for text summarization. Nallapati et al. [3] proposed neural network-based models that use deep learning to understand the semantic meaning of sentences and produce better summaries. More recently, transformer-based models such as BERT and GPT have significantly improved summarization performance by capturing contextual relationships

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503103|

between words and sentences. Liu and Lapata [4] demonstrated the effectiveness of transformer architectures for summarization, while Cohan et al. [5] focused on summarizing long scientific research papers by extracting key sections such as abstracts and conclusions. Inspired by these approaches, the proposed system aims to generate concise research paper summaries using NLP preprocessing and machine learning techniques.

III. METHODOLOGY

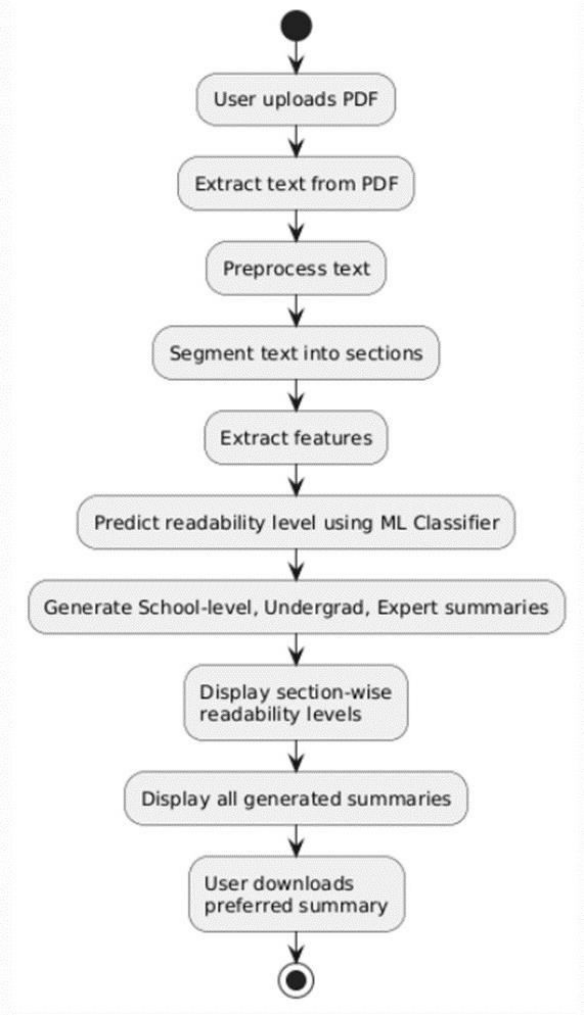


Fig 3.1 Intelligent Research Paper Summarizer Methodology

The methodology of this project focuses on developing an automated system that summarizes research papers using Natural Language Processing (NLP) and Machine Learning techniques. The system accepts research papers in PDF or text format as input. The text is first extracted from the document using text extraction methods. After extraction, preprocessing is performed to remove stop words, punctuation, and unnecessary characters to clean the data. NLP techniques such as tokenization, sentence segmentation, and word normalization are then applied to understand the structure of the text. Important sentences are identified based on relevance and significance.

The model is developed using Python with NLP and machine learning libraries. Research papers are collected for training and testing the summarization system. Sentence scoring and keyword importance techniques are used to identify meaningful sentences. The system architecture includes modules such as input, text extraction, preprocessing, NLP processing, summarization, and output. Finally, the model generates a concise summary that captures the key

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503103|

ideas of the research paper and presents it to the user. The final summarized version of the research paper is displayed to the user. This architecture ensures that the system processes research documents efficiently and generates clear summaries that highlight the key ideas of the paper.

IV. SYSTEM ARCHITECTURE

The architecture of the Intelligent Research Paper Summarizer is designed as a modular pipeline that integrates document processing, machine learning prediction, and transformer-based summarization. The system consists of two primary components: an offline model training pipeline and an online document processing pipeline. This architecture ensures efficient processing while allowing the system to incorporate improved models in future upgrades. In the offline training phase, readability datasets are processed to extract linguistic features used for model training. Multiple machine learning algorithms are trained and evaluated to determine the most effective classifier for predicting text complexity. After evaluation, the best-performing model along with its pre-processing scaler is saved and integrated into the deployed application. This separation between training and deployment improves scalability and allows new models to be incorporated without modifying the entire system.

The operational pipeline begins when a user uploads a research paper through the Streamlit interface. The system extracts textual content from the uploaded PDF and performs pre-processing to remove irrelevant formatting artifacts. The cleaned document is then segmented into sections or manageable chunks to preserve the structural organization of the research paper. For each section, linguistic features are computed and passed to the trained readability classifier. The classifier predicts the complexity level of the content, enabling the system to display section-wise readability insights that help users understand which parts of the paper are easier or more difficult to read. Simultaneously, each section is processed by the FLAN-T5 summarization engine, which generates summaries for three reader groups: School-level, Undergraduate-level, and Expert-level. Prompt-based instructions guide the summarization model to adjust the level of explanation, vocabulary, and technical depth according to the target audience.

Finally, the generated summaries are aggregated across sections to produce complete document-level summaries. These summaries, along with readability analysis results, are presented through the Streamlit interface where users can view and download summaries according to their preferred reading level. This modular architecture ensures efficient document processing while providing adaptive summaries for diverse audiences.

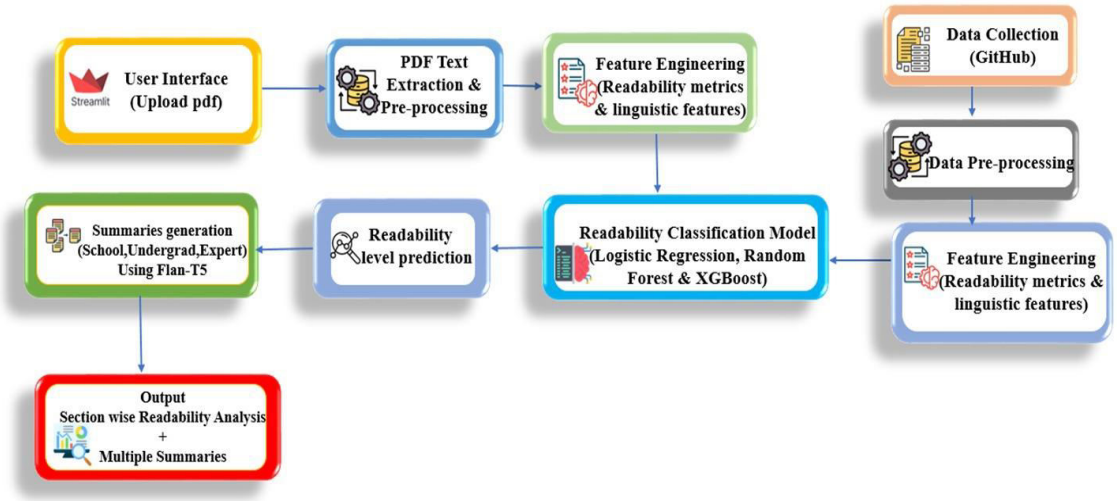


Fig 4.1 System Architecture

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503103|

V. EVALUATION METRICS

To evaluate the readability classifier, the OneStopEnglish Paragraphs Dataset was used, which contains English passages rewritten at three readability levels: Elementary, Intermediate, and Advanced. In this study, these were mapped to School, Undergraduate, and Expert categories. The dataset originally contained 3570 paragraphs, and after removing 11 duplicates, the final dataset consisted of 3559 samples. Each entry includes the paragraph text, a numeric label, and the readability level. Linguistic and statistical features were extracted from the paragraphs and used to train machine learning models.

Three classifiers Logistic Regression, Random Forest, and XGBoost were trained on the extracted features and evaluated on a held-out test set. Performance was measured using accuracy, precision, recall, F1-score, and macro F1-score, where macro F1 ensures balanced evaluation across all classes. In addition to test performance, training accuracy was also examined to assess model learning and generalization. The training accuracy, test accuracy, and macro F1-score for each model are reported in Table 5.1. Although Random Forest and XGBoost achieved perfect training accuracy, their high test accuracy indicates strong generalization, suggesting only minor overfitting, which is common in tree-based ensemble models.

Model	Train Accuracy	Test Accuracy	Macro F1 Score
Logistic Regression	94.01%	94.39%	0.92
Random Forest	100%	96.21%	0.94
XGBoost	100%	96.36%	0.95

Table 5.1 Evaluation Metrics

Among the models evaluated, the best test accuracy and Macro F1 score were obtained by the XGBoost model, which suggests that the model has better capabilities to deal with complex relationships between the linguistic features. Therefore, the XGBoost model has been selected as the final readability classification model for the proposed system.

VI. OUTPUT SCREENS

As depicted in Figure 6.1 below, the major interface of the Intelligent Research Paper Summarizer developed using the Streamlit library is presented, whereby the user can upload a research paper in PDF format using the drag-and-drop box or the browse option. After the research paper is uploaded, the system will then perform the processing of the uploaded research paper by extracting the text and carrying out the text readability analysis. Figure 6.2 below depicts the section-wise input text complexity analysis, whereby the trained readability classification model analyses each section of the uploaded research paper and assigns the complexity level as low, medium, or high based on the linguistic features. Figure 6.3 below depicts the summaries generated using the transformer-based summary model for three different audience levels: School, Undergraduate, and Expert. The user can now view the summaries for the same text at three different complexity levels. Figure 6.4 below depicts the download functionality of the generated summaries in the form of a .txt file, whereby the user can download each summary for easy accessibility.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503103|

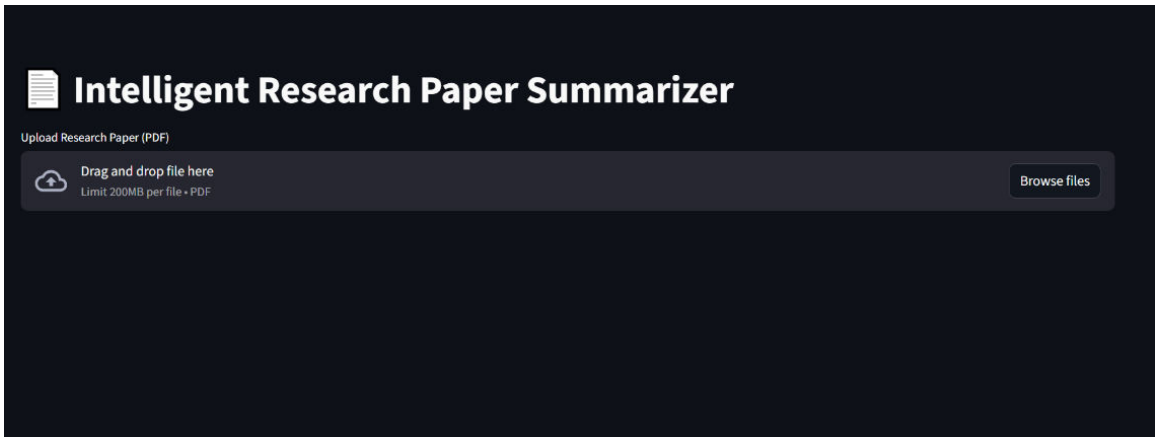


Fig 6.1 Upload Screen

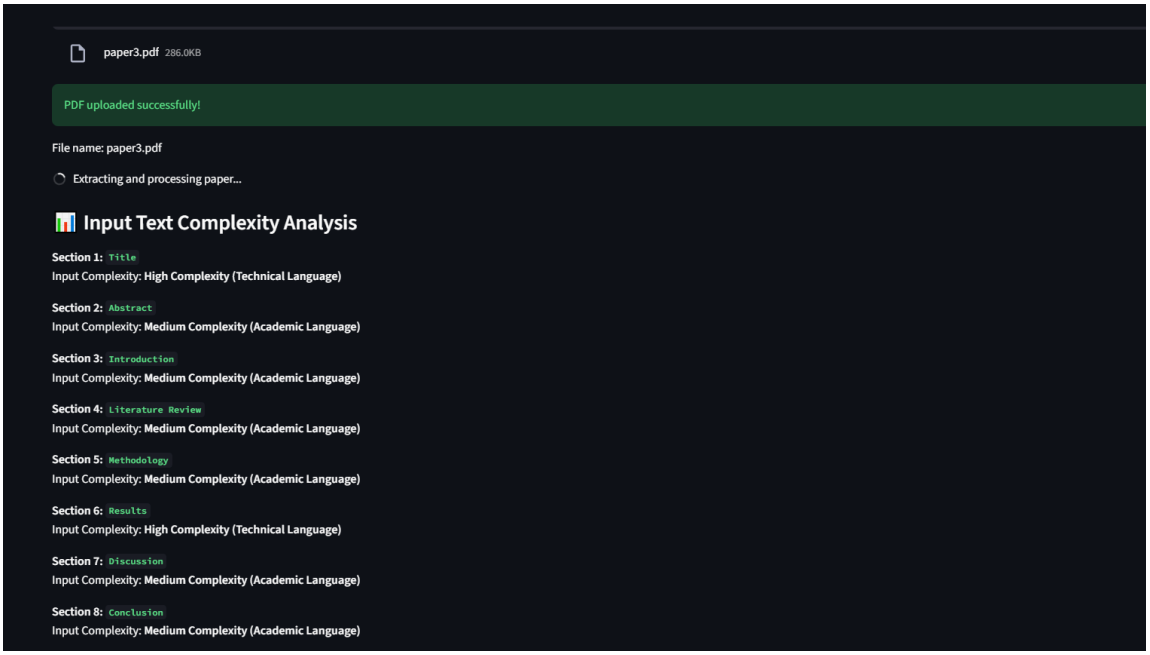


Fig 6.2 Section Wise Readability Analysis

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503103|

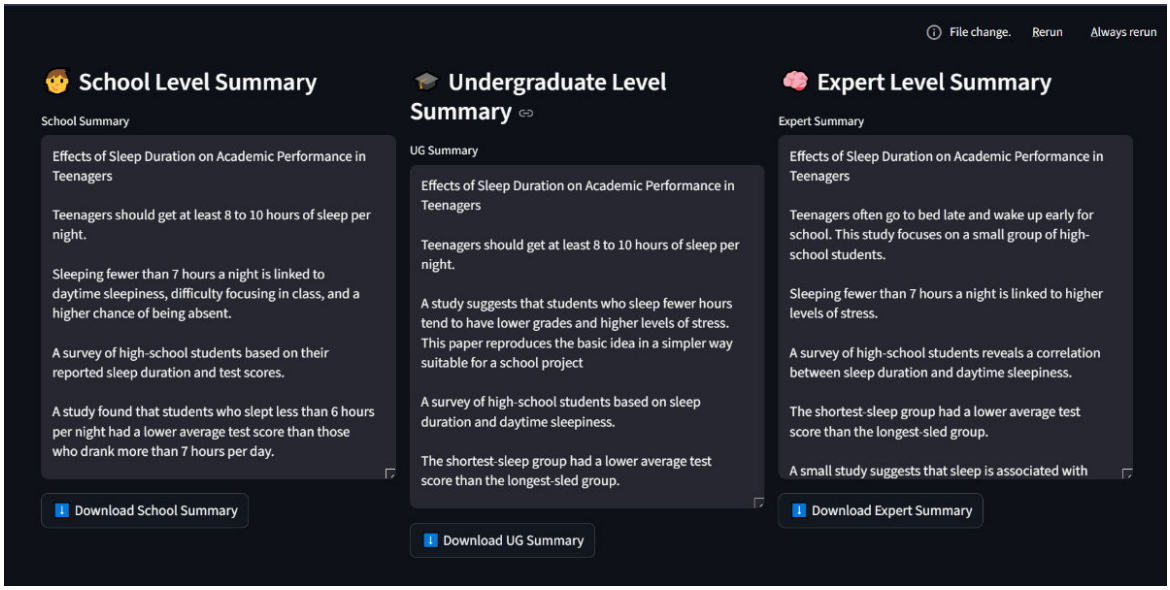
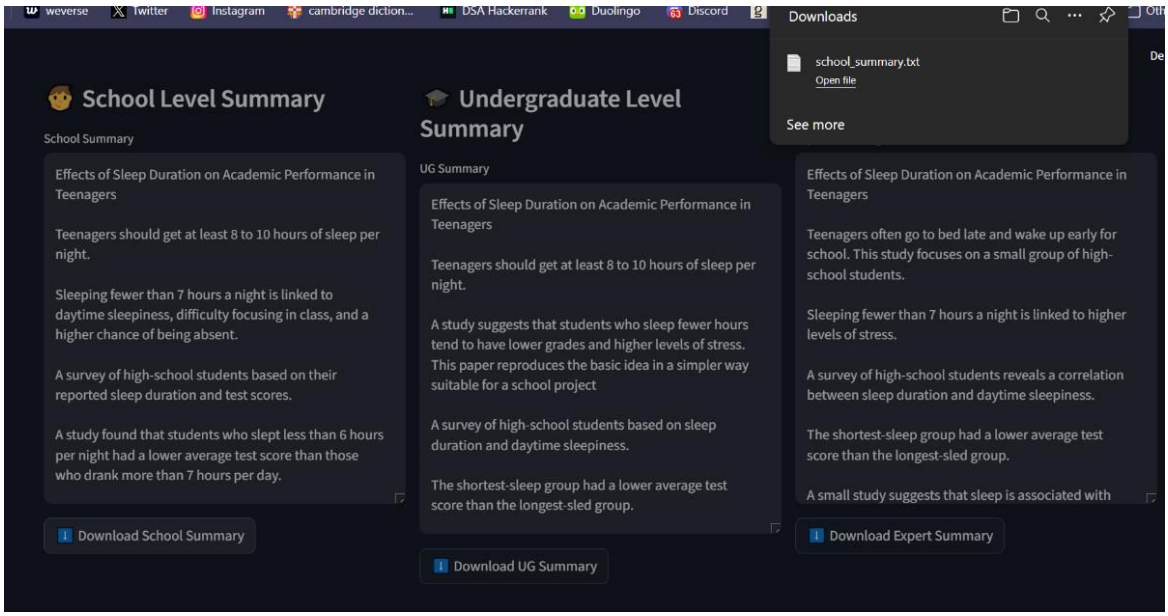


Fig 6.3 Generated Multiple Summaries



6.4 Download Confirmation

VII. CONCLUSION

In this project, an intelligent system for summarizing research papers was developed using Machine Learning and Natural Language Processing techniques. Research papers usually contain large amounts of detailed information, which makes it difficult for students and researchers to quickly understand the main ideas. The main aim of this project was to create a system that can automatically analyse a research document and generate a short and meaningful summary. The system works by processing the text, removing unnecessary words, and identifying the most important sentences in the document. With the help of NLP techniques and machine learning models, the system can understand the structure of the text and produce summaries that still keep the main meaning of the original paper. This project shows how artificial

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503103|

intelligence can be used to make academic reading easier and faster. By automatically summarizing research papers, the system helps users save time and quickly understand the key concepts of a document. Overall, the project demonstrates the usefulness of ML and NLP technologies in managing and simplifying large amounts of textual information.

VIII. FUTURE SCOPE

Although the system works well for generating summaries, there are many ways it can be improved in the future. The model can be enhanced by using more advanced deep learning algorithms to produce even more accurate summaries. Another improvement would be adding support for multiple languages so that research papers written in different languages can also be summarized. The system can also be converted into a web or mobile application where users can upload research papers and instantly get summaries. Additional features such as keyword extraction, topic identification, and visual representation of important points can make the tool more useful for researchers.

In the future, the system may also be integrated with digital libraries or academic platforms to automatically summarize newly published research papers. These improvements would make the tool more powerful and helpful for students, researchers, and professionals.

REFERENCES

[1] See, A., Liu, P. J., & Manning, C. D. (2017). Get To The Point: Summarization with Pointer Generator Networks. Proceedings of ACL.
Link: https://arxiv.org/abs/1704.04368
[2] Liu, J., Guo, S., & Jin, Y. (2019). Fine-Grained Readability Prediction. Proceedings of the 57th Annual Meeting of the ACL.
Link: https://aclanthology.org/P19-3010
[3] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., et al. (2020). BART: Denoising Sequence to-Sequence Pre-training for Natural Language Generation. ACL.
Link: https://arxiv.org/abs/1910.13461
[4] Raffel, C., Shazeer, N., Roberts, A., et al. (2020). Exploring the Limits of Transfer Learning with T5. JMLR. Link: https://arxiv.org/abs/1910.10683
[5] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL.
Link: https://arxiv.org/abs/1811.00196
[6] Cohan, A., Feldman, S., Beltagy, I., et al. (2018). A Discourse-Aware Attention Model for Scientific Document Summarization. NAACL.
Link: https://arxiv.org/abs/1804.09068
[7] Vajjala, S., & Meurers, D. (2016). Readability Assessment for Text Simplification. Workshop on Computational Linguistics for Linguistic Complexity.
Link: https://aclanthology.org/W16-4111

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details