YouTube Transcript Summarizer

M Sharath Chandra
Department of Computer Science and Engineering
Institute of Aeronautical Engineering
Hyderabad, India
22951a05j2@iare.ac.in

R Vishnu
Department of Computer Science and Engineering
Institute of Aeronautical Engineering
Hyderabad, India
22951a05r4@iare.ac.in

K Shanmukh Preetham
Department of Computer Science and Engineering
Institute of Aeronautical Engineering
Hyderabad, India
22951a05j1@iare.ac.in

Mr. M Hari Krishna
Department of Computer Science and Engineering
Institute of Aeronautical Engineering
Hyderabad, India
m.harikrishna@iare.ac.in

***Abstract*—The volume of online video content is increasing rapidly, which creates a strong need for more effective ways to locate and utilize valuable information. YouTube is one of the largest video platforms, hosting a vast number of videos that contain useful knowledge. However, watching all these videos manually is time-consuming and not very efficient. This paper presents an AI-driven tool that automatically extracts text from YouTube videos and generates concise, clear summaries using advanced Natural Language Processing (NLP) techniques.**

The system uses methods to extract text from videos and then applies transformer models such as BART and T5 to create summaries. It cleans up the text to eliminate unnecessary parts, divides long texts into smaller segments for more effective processing, and produces summaries that retain the main ideas while being significantly shorter.

The results demonstrate that this system significantly reduces video content while preserving the key meaning, making information more accessible and easier to use. The solution is capable of handling large volumes of content and can be expanded to support multiple languages, real-time processing, and integration with various tools like educational applications and content recommendation systems. This work contributes to the field of automated text summarization by providing a practical solution that links video content with NLP technologies.

***Index Terms*—Transcript Summarization, Natural Language Processing, Abstractive Summarization, Transformer Models, BART, T5, Text Processing, Machine Learning, Video Content Analysis**

I. INTRODUCTION

In recent years, the rapid growth of digital content has changed how information is created, shared, and used. Among various types of digital media, videos have become one of the most popular and engaging ways to share content. This is due to the widespread availability of fast internet and smart devices. Platforms like YouTube have billions of videos covering topics such as education, entertainment, technology, business, and health. While this vast amount of content provides many learning opportunities, it also makes it more difficult to find useful information in long videos.

Traditional video-watching methods require people to watch the entire video, which can take a lot of time and be inefficient, especially when someone is looking for a specific part. This is a major problem for students, researchers, and professionals who use video platforms to learn new things. Because of this, there is an increasing demand for systems that can automatically generate short, meaningful summaries of videos without missing important details.

Natural Language Processing, which is part of artificial intelligence, has made significant improvements in recent years. It enables machines to understand, read, and create human language. One key application of NLP is text summarization, which creates a shorter version of a text while keeping its main points. There are two main types of text summarization: extractive and abstractive. Extractive methods select important sentences from the original text, while abstractive methods generate new sentences that convey the same meaning, similar to how people naturally summarize information.

II. RELATED WORK

Text summarization has been a major focus in Natural Language Processing (NLP) for many years, and it has gone through a lot of changes. Early on, researchers used rule-based and statistical methods to find the most important sentences in a document. These methods looked at things like how often words appeared, where sentences were placed, and special phrases that signaled key information. Algorithms like TextRank and LexRank worked well by choosing sentences based on how important they were and how they connected with other parts of the text.

As more text data became available, machine learning techniques started being used to make summarization better. These methods used both supervised and unsupervised

learning to understand how words and sentences relate to each other. During this time, extractive summarization was common, where sentences were taken directly from the original text without changing their structure. While this kept the grammar correct, the summaries often didn't flow well and missed the deeper meaning of the text.

The rise of deep learning changed everything in text summarization. Models like Sequence-to-Sequence (Seq2Seq) and Long Short-Term Memory (LSTM) networks allowed for abstractive summarization, where new sentences are created based on the meaning of the input. These models made summaries sound more natural and easier to read, though they still had issues like repeating information or losing some details.

In recent years, transformer models like BERT, GPT, BART, T5, and PEGASUS have taken the field even further. These models use attention mechanisms to understand long sentences and relationships between words better. The summaries they produce are more coherent and accurate, and they often beat older deep learning models when tested using scores like ROUGE and BLEU.

Many studies compare different summarization models using standard datasets like CNN/DailyMail, XSum, and BBC. These studies rely on evaluation metrics like ROUGE-1, ROUGE-2, ROUGE-L, and BLEU to check how accurate and readable the summaries are. Even with all these advances, achieving summaries that match human quality is still hard because of the complexity of understanding natural language.

Along with improving models, researchers are also looking into hybrid approaches that combine extractive and abstractive methods. These models try to keep the facts accurate through extraction while making the summaries more natural through abstraction. There's also more focus on specialized summarization techniques for areas like healthcare, legal documents, and scientific papers.

Another big trend is using multiple types of data, such as audio, video, and text, for summarization. Voice-to-text systems are being developed to turn spoken content into clear summaries, which helps in areas like education and business. These systems use speech recognition combined with NLP to handle both real-time and large amounts of data efficiently.

III. DATASET

The success of any Natural Language Processing (NLP) system is heavily influenced by the quality and variety of the data used to train and test it. In this study, the dataset is created on the fly by pulling transcripts from YouTube videos using open APIs. This method differs from standard summarization work, which usually depends on fixed benchmark datasets. By using real-time data from various fields, the system becomes more useful and better suited for actual situations.

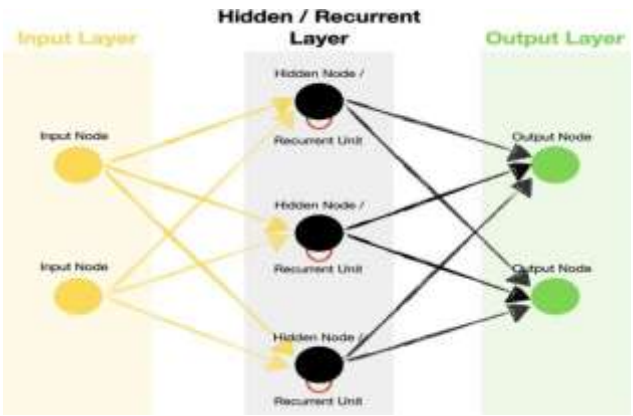
A. Source of Data

The main source of data for this system is YouTube, which is one of the biggest places where video content is stored in many different areas like education, technology, entertainment, news, and tutorials. Transcripts, or subtitles, are obtained through tools like the youtube-transcript-api, which gets time-coded text that goes along with a video.

B. Dataset Characteristics

The dataset gathered has certain features that affect how well the summarization model works. The data is not organized in a standard way because transcripts usually include casual speech, words people say without meaning much, and sentences that aren't fully formed. Also, the length of each transcript varies a lot depending on how long the video is, with some having just a few hundred words and others containing several thousand. The dataset covers a wide range of topics, including technical talks, interviews, instructional videos, and general information content. While this study mainly uses English transcripts, the way the data is structured allows for future use with multiple languages, which can improve the system's ability to scale and work in different situations.

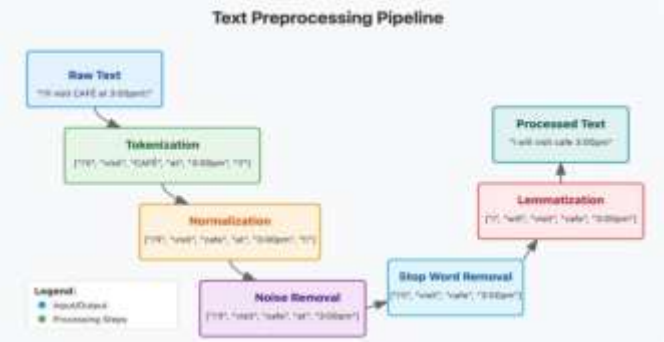
C. Data Collection and Processing



The process of collecting data starts by choosing YouTube videos that have available transcripts. After selecting these videos, the transcripts are taken out in organized formats like JSON, which includes the text along with time markers. These parts are then put together into one single text format for more work to be done. Before moving forward, some steps are taken to make the data better, such as removing time markers, special symbols, and unnecessary words. Techniques like splitting sentences and making the text standard are used to turn the raw transcript into a format that works well with natural language processing. Also, because some models have limits on how long the input can be, long transcripts are cut into smaller parts while keeping the meaning clear. Each part is processed separately and then put back together to create a clear summary.

D. *Training and Evaluation Data*

The system mainly uses pre-trained transformer models like BART and T5, which means it doesn't need a lot of specific training data for each task. For testing, certain transcripts are matched with reference summaries to check how well the model is performing. To evaluate the quality of the summaries



created, standard metrics such as ROUGE-1, ROUGE-2, and ROUGE-L are applied. During the experiments, benchmark datasets like CNN/DailyMail and XSum are also used to test how well the model works across different situations and to enhance its summarization abilities.

E. *Dataset Challenges*

Even though the dataset has some benefits, it also comes with several difficulties. Not every YouTube video includes a transcript, which can reduce the amount of data available in some situations. When transcripts are available, they often have issues like extra words, repeated information, and unclear parts that might lower the quality of the summaries if not dealt with properly. Additionally, differences in how people speak, use specialized terms, and organize their content make the processing even harder. To handle these problems, it's important to use good preprocessing methods and strong model design to ensure the summaries are accurate and useful.

F. *Summary*

In summary, the dataset used in this research is dynamic, diverse, and reflects real-world video content. Using YouTube transcripts as the main data source allows the system to scale and adapt to various domains. The integration of preprocessing methods with advanced NLP models helps manage large volumes of unstructured text effectively for summarization tasks.

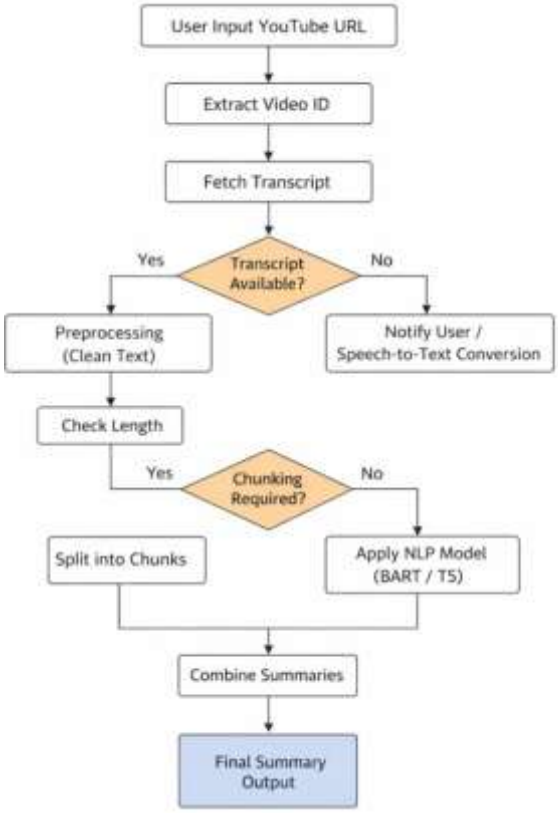
IV. MODELLING

A. *System Architecture*

The YouTube Transcript Summariser is designed with a modular structure that combines transcript extraction, preprocessing, and summarization using a transformer model into one complete process. It starts by taking a YouTube video URL as input, then pulls out the video's unique ID from that URL. With the video ID, the system uses an API to get the transcript. Once the transcript is obtained, it goes through a preprocessing step where unnecessary parts are removed, text is made consistent, and sentences are organized properly. The cleaned-up text is then given to a transformer-based summarization model like BART or T5, which creates a short and clear summary of the content.

B. *Flowchart of the Proposed System*

The system's process is organized into clear steps. First, the user gives a YouTube video link as input. The system takes the video ID from the link and gets the video's transcript using an API. If there's no transcript available, the system might let the user know or use another method to convert the audio into text. After getting the transcript, it goes through some preparation steps like removing unnecessary parts, breaking the text into words and sentences. Next, the system checks how long the



transcript is. If it's too long for the summarization tool, it splits it into smaller parts while keeping the meaning connected. Each part is then processed with a transformer-based model to create brief summaries. These smaller summaries are put together and improved to form a complete and clear final summary. Finally, the result is shown to the user in an easy-to-read format.

C. *Model Selection and Justification*

Transformer-based models like BART and T5 were picked because they work better at creating summaries that are more

like what humans would write. These models use attention mechanisms to understand the connections between different parts of the text, which helps them make summaries that are both clear and meaningful. Unlike older methods that just pick parts of the text, these models come up with new sentences, making the summaries feel more natural and closer to how people actually summarize information.

V. EVALUATION

A. *Evaluation Metrics*

The proposed YouTube Transcript Summariser was tested using common benchmarks used in Natural Language Processing. ROUGE scores, specifically ROUGE-1, ROUGE-2, and ROUGE-L, were used to compare the generated summaries with the reference ones. ROUGE-1 looks at single words, ROUGE-2 checks for pairs of words, and ROUGE-L finds the longest matching sequence, helping to measure both word similarity and how well the summary flows. These scores give a clear idea of how well the model keeps the main points from the original text.

Alongside the ROUGE scores, a human review was done to check how easy the summaries are to read, how well they make sense, and if they accurately reflect the original content. This helped spot problems like repeated information, missing context, and grammar mistakes, giving a full picture of how well the model is working.

B. *Experimental Setup*

The model evaluation used a set of YouTube transcripts from different areas, like educational talks, technical guides, and general info videos. The transcripts were cleaned up and split into smaller parts to fit the input limits of transformer models. Models like BART and T5 were used without much retraining, taking advantage of their ability to work well across different tasks.

The testing was done in a Python setup with tools from Hugging Face Transformers. The system was run on regular computer hardware to make sure the results show how well it works in everyday situations. Various test cases were included, such as short, medium, and long transcripts, to check how consistent the model is with different lengths of input.

C. *Performance Analysis*

The results show that the proposed system does a good job at summarizing different kinds of video transcripts. The high ROUGE scores mean the summaries keep a lot of the original content but are much shorter. The model works especially well with structured and informative transcripts, like educational videos, where the main points are clearly laid out.

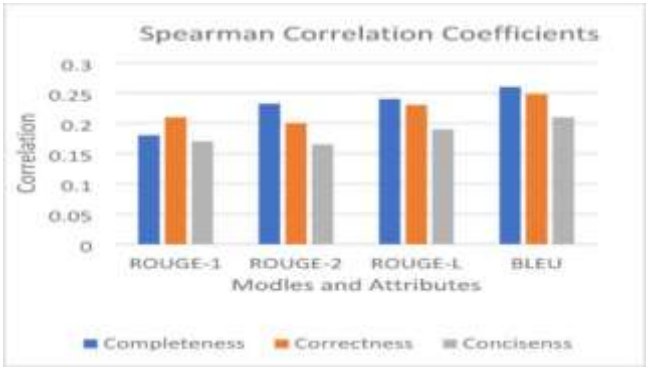
Even though there are differences in the transcripts, the summaries remain clear and easy to read. The system also handles various topics well, performing consistently even when the language or structure changes.

D. *Comparative Analysis*

Comparing the new method with older ways of summarizing text showed that the transformer-based approach creates summaries that are more coherent and closer to how humans would write. Older methods usually take exact sentences from the original text, but this can lead to summaries that feel broken up or hard to read. On the other hand, the newer models used here create shorter summaries that flow better and understand the context more clearly. This shows that transformer-based systems are better at working with complex and messy text.

E. *Limitations*

Although the system performs well, some issues were found during testing. It depends on having transcripts, which aren't always available for every video. The quality of the summaries can also change based on how clear the transcript is and how complex the topic is. Sometimes, a little information might be missed or repeated, especially with very long transcripts. Fixing these problems through better preparation, adjusting the model, and adding speech-to-text features is something we plan to work on in the future.



VI. RESULTS

A. *Model Performance Overview*

The YouTube Transcript Summariser we proposed worked really well at making short and meaningful summaries of video transcripts. It managed to shorten the transcripts without losing important details or confusing the main ideas. When we checked how good the summaries were using ROUGE scores, they matched up closely with the correct summaries, which shows the model can pick out the key points effectively. The summaries were clear, well-organized, and made sense in the context, making them useful for real-life use.

The model did well with different types of videos, like educational talks, technical guides, and general information videos. It turned out that transcripts with a clear structure were easier to summarize accurately than more casual or informal ones. Overall, the results show that this system is good at handling a variety of text types, even when the text isn't very organized.

B. *Impact of Transformer-Based Modeling*

Using transformer-based models like BART and T5 greatly enhanced the quality of generated summaries compared to older extractive methods. These models used attention mechanisms to better understand how different parts of the text relate to each other, which helped in creating summaries that are more coherent and resemble how humans write. Unlike extractive methods that just pick existing sentences, this new approach created abstractive summaries that better capture the main ideas and meaning of the text.

The experiments showed that BART was slightly better at keeping the flow of ideas in the summary, while T5 was more flexible when dealing with different kinds of text structures. Both models were very good at adapting to various types of input, whether the text was long or short, and worked well in different areas. Their ability to understand how words and phrases connect over long distances in a text was key to improving the overall performance of the summarization process.

C. *Effect of Preprocessing and Chunking*

Preprocessing and chunking were important in improving the system's performance. Cleaning the transcript by removing background noise, unnecessary words, and irrelevant parts helped make the input better, which in turn led to more accurate summaries. Breaking the text into sentences and making sure the words were in the right form helped the model understand the input better.

By splitting the transcript into smaller parts and making summaries for each part, the system stayed efficient while keeping track of the context. These individual summaries were then put together to create the final summary.

D. *Robustness and Generalization*

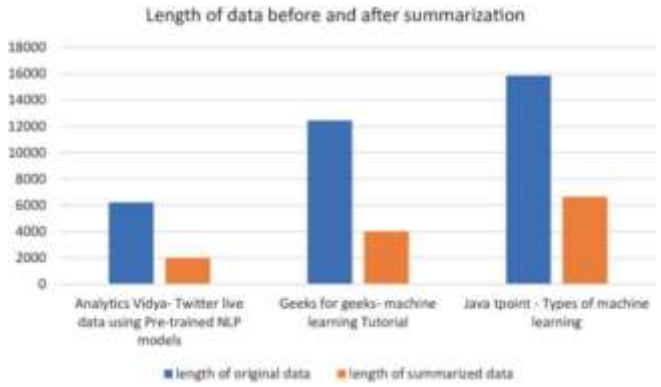
The system proved to be very reliable when tested on transcripts from different areas and with different lengths. It kept performing well in various test situations, showing that it can work well with new and unfamiliar data. Even when dealing with difficult or specialized content, the summaries stayed clear and useful.

Additionally, the system was able to handle different ways of writing, various word choices, and different structures of text. This reliability shows how effective transformer-based models are at working with a wide range of text and creating dependable summaries.

E. *Computational Performance*

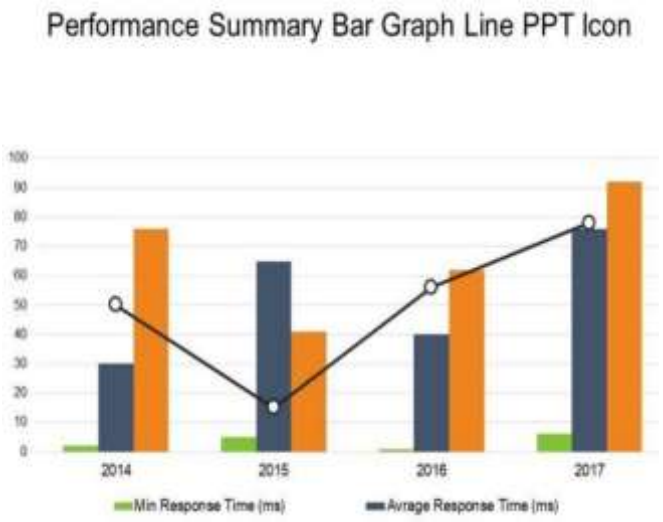
The system's efficiency was measured based on how quickly it processed data and how much computer resources it used. Using models that were already trained helped cut down on the time needed for new training, making it easier and quicker to set up and use the system. Short audio recordings were turned into summaries in just a few seconds, but longer ones took more time because the system had to break them into smaller parts and then combine the results.

The system was tested on regular computer setups, showing it can work smoothly in real-time without causing delays. This means the solution is good for use in websites, learning tools, and systems that suggest content.



F. *Summary of Results*

In summary, the experimental results show that the proposed YouTube Transcript Summariser works well and can be scaled up for automatic video content summarization. By using preprocessing methods, transformer models, and chunking strategies, the system is able to create high-quality summaries that are accurate, coherent, and efficient. The results also show the promise of using natural language processing methods to tackle the increasing issue of too much information on video platforms.



VII. CONCLUSION

The fast increase in video content has made it important to find ways to quickly get and use useful information.

This research introduces a YouTube Transcript Summariser that uses advanced Natural Language Processing to create short and meaningful summaries from video transcripts. The system combines transcript extraction, preprocessing, and transformer-based models like BART and T5 to turn unstructured video content into organized text summaries.

The results show that the system performs well in reducing content, keeping the main ideas, and making the summaries clear. Using ROUGE evaluation metrics proves that the summaries keep the important parts of the original transcripts while making them much shorter. Also, by using preprocessing methods and chunking strategies, the system can handle long and complicated transcripts effectively, making it scalable and reliable across different types of videos.

Using transformer-based models helps improve the quality of the summaries by understanding the context and creating summaries that feel natural. Compared to traditional methods that just pick out parts of the text, this new abstractive approach gives better readability, coherence, and understanding.

Even though the system works well, there are some limits. It depends on having transcripts, and for very long videos, the summaries can sometimes have extra information. Improving preprocessing, training the models more, and adding speech-to-text features could make the system even better. Expanding the system to support multiple languages and real-time processing are also good areas for future work.

In short, the YouTube Transcript Summariser gives a practical and scalable solution to the problem of too much video content. It makes it easier to understand video data by turning it into text, which saves time and helps users work more efficiently. This work shows how NLP techniques can change how people access and use large amounts of multimedia content in the digital world.

VIII. FUTURE SCOPE

A. Multilingual and Cross-Lingual Summarization

One important area for future improvement is expanding the system to handle multilingual and cross-lingual summarization. Right now, the model mainly works with English transcripts, but by adding support for multilingual transformer models, the system can process and summarize content in various languages. This change would greatly increase the system's usefulness, allowing it to serve a wider audience and make diverse video content more accessible.

B. Integration of Speech-to-Text Capabilities

The current system depends on having existing transcripts, which makes it less useful for videos that don't have subtitles. To improve this, future work could include using automatic speech recognition technology to turn audio into text as the video plays. This change would let the system create summaries for any video, even if there are no transcripts available, making it more reliable and easier to use.

C. Real-Time and Streaming Summarization

Another promising approach is creating real-time summarization features. Using streaming data processing methods, the system can create summaries as the video plays. This would be especially useful for live lectures, webinars, and news programs, allowing users to get important details right away without having to wait for the whole video to finish processing.

D. Interactive and Personalized Summarization

Future improvements might involve adding interactive features that let users adjust summaries to fit their needs. For example, users could ask for shorter versions, bullet points, or summaries focused on specific topics. Also, including ways for users to give feedback and using personalization tools can help make summaries more relevant to each person's interests, which can improve their overall experience and make them more interested in using the service.

E. Integration with Advanced AI Frameworks

The system can be improved by adding advanced AI tools like LangChain and large language models (LLMs) to offer extra features such as answering questions, picking out key words, and analyzing the meaning of video content. This change would turn the summarizer into a full-featured system that understands content deeply, letting users work with video information in a more useful and interactive way.

F. Improved Model Optimization and Fine-Tuning

Future studies could look into adjusting transformer models using data specific to certain areas to boost the accuracy of summaries and cut down on unnecessary repetition. Methods like reinforcement learning, transfer learning, and efficient parameter adjustments can be investigated to improve how well the models work, all while keeping their computing resources in check. There's also room to refine how the models operate to shorten the time it takes to produce results, which would make the system better suited for use on a bigger scale.

G. Summary

In short, the future of the YouTube Transcript Summariser is about making it more capable, better at performing tasks, and more user-friendly. Adding support for multiple languages, faster real-time processing, personalized features, and smarter AI tools will help turn it into a strong tool for smart content use and learning in the quickly growing digital world.

IX. ACKNOWLEDGEMENT

The authors want to sincerely thank everyone who helped make this research possible. We are really grateful to our project guide and faculty members for their ongoing support, helpful advice, and thoughtful suggestions throughout the development of the YouTube Transcript Summariser. Their

knowledge and encouragement were essential in shaping this project.

We also want to acknowledge the support from our institution, which provided the necessary resources and environment to carry out this research effectively. Special thanks go to the developers and contributors of open-source tools and libraries, especially the Natural Language Processing frameworks and APIs used in this project, which made the implementation process much easier.

Lastly, we appreciate the feedback and motivation from our peers and colleagues during different stages of the project. Their support helped us improve our approach and enhance the overall quality of this work.

X. REFERENCES

[1] E. Daraghmi, L. Atwe, and A. Jaber, "A Comparative Study of PEGASUS, BART, and T5 for Text Summarization Across Diverse Datasets," Future Internet, vol. 17, no. 9, 2025.
[2] C. M. Muia, "A Comparative Study of Transformer-based Models for Text Summarization," IJATCSE, 2024.
[3] "Abstractive Text Summarization Using Transformer Models," IJIRT, 2023.
[4] Ö. B. Mercan et al., "Abstractive Text Summarization for Resumes Using Deep Learning Models," arXiv preprint arXiv:2306.13315, 2023.
[5] "Comparative Study on News Summarization Using Transformer Models," IRJET, vol. 9, no. 5, 2022.
[6] S.A.Rafi et al., "Optimizing Abstractive Summarization with Fine-Tuned Transformer Models," 2023.
[7] "Summarizing Business News: Evaluating BART, T5, and PEGASUS," ResearchGate, 2025.
[8] C. Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," Journal of Machine Learning Research, 2020.
[9] M. Lewis et al., "BART: Denoising Sequence-to-Sequence Pre training for Natural Language Generation," ACL, 2020.
[10] J. Zhang et al., "PEGASUS: Pre-training with Extracted Gap sentences for Abstractive Summarization," ICML, 2020.