

International Exposure Program (IEP) Report

PDEU(Pandit Deendayal Energy University) and AIT(Asian Institute Of Technology)

**Natural Language Processing application**

**By**

Rurakumar Alpeshbhai Madhu

Vraj Chetankumar Patel

Soham Rawal

Zeeshan Sanaullakhan Ghori

**Nationality:** Indian

**Pursuing Degree:** Bachelor of Engineering in

Computer Science Engineering,

Information Communication Technology Engineering,

At Pandit Deendayal Energy University (PDEU),

Gandhinagar, Gujarat, India.

**Under the guidance of:** Dr. Chaklam Silpasuwanchai

And other mentors: Amanda Raj Shreshtha,

Sakib Bin Alam, Ayush Koirala, Abhinav Lugun.

**Examination Committee**

Prof. Nitin Kumar Tripathi

Mr. Ranadheer Reddy

**Submitted At**

Asian Institute of Technology

School of Engineering and Technology, Thailand

July-August 2023

**ACKNOWLEDGMENTS**

We are immensely indebted to **Pandit Deendayal Energy University (PDEU)** for organizing the International Exposure Program at **Asian Institute of Technology (AIT), Thailand.** We would like to thank the Office **of international relations** and **Mr. Maulik Shah** for having good collaborative relations with the Asian Institute of Technology and being all time interactive to provide us the best outcomes from the IEP. It gave us an opportunity to get exposed to the latest technology and helped us to learn the research perspectives in our domain. It gave us great pleasure to have completed the entire one month as a student of AIT, gaining knowledge in the field of wherein the world is progressing rapidly. We thank **Dr. Nitin Tripathi** for managing the entire IEP and organizing the domain specific programs for students of all branches. Our sincere thanks to **Mr. Ranadheer Reddy**, project coordinator- Special Degree Programs at AIT, for mentoring us in the entire program for a month and being all time available to us for any guidance needed whether it be in the academic specific aspect or anything related to non-academic purpose. Furthermore, thanking **Arthur Lance Gonzales,** Program Officer at AIT, for giving us the best possible hospitality over the entire program duration. Most importantly we thank our professor at AIT, **Dr. Chaklam Silpasuwanchai**, Assistant Professor Computer Science & Data Science and AI Asian Institute of Technology, and our mentors **Amanda, Ayush, Abhinav, Sakib** under whose guidance we were trained for the overall program and gained the knowledge in the field of artificial intellegence with research perspective. Lastly, we thank student coordinators **May Hnin, Sai** and **Jirawat D.** for being all time available to us, providing us with our needs, and being best guides to us.

**ABSTRACT**

This paper presents a project that implements paraphrasing and summarizing features using the T5 and FactorSum models. The T5 model is a large language model that is trained on a massive dataset of text and code. It can be used to paraphrase the text, which means to restate it in different words while preserving the original meaning. The FactorSum model is a summarization model that is also trained on a massive dataset of text. It can be used to summarize text, which means to reduce it to its most important points. The project was implemented using the Python programming language. The T5 model was implemented using the Hugging Face Transformers library. The FactorSum model was implemented using the FactorSum library. The project was evaluated using a set of test data. The results showed that the project was able to paraphrase and summarize text with high accuracy. Our project is a valuable tool for anyone who needs to paraphrase or summarize text.

The implementation of both models was carried out using python, one of the most popular languages in the field of data science and machine learning, The T5 model was integrated in the project using Hugging Face Transformer Library, a widely used library that offers numerous pre-trained models for nlp tasks. The FactorSum model was implemented using FactorSum Library, a specialized library for text summarization tasks

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **SR NO.** | **CONTENTS** | **PAGE NO.** |
| **1** | **ACKNOWLEDGMENT** | **2** |
| **2** | **ABSTRACT** | **3** |
| **3** | **LIST OF FIGURES** | **5** |
| **4** | **LIST OF ABBREVIATIONS** | **6** |
| **5** | **CHAPTER 1: INTRODUCTION**  1.1: Background of the Study  1.2: Statement of the Problem  1.3: Research Questions  1.4: Objectives of the Study | **7**  **7**  **8**  **8**  **8** |
| **6** | **CHAPTER 2: DESCRIPTION**  2.1: Frontend  2.2: Backend  2.3: Factor-sum Model  2.4: T5 Model  2.5: Future Work | **9**  **10**  **14**  **16**  **17**  **20** |
| **7** | **REFERENCES** | **21** |
| **8** | **APPENDIX** | **22** |

**LIST OF FIGURES**

**2.1:** Overview of the entire pipeline of the project

**2.2:** Homepage of the website

**2.3:** Pdf viewing in website

**2.4:** Latex symbol writing in website

**2.5:** Paraphrasing in website

**2.6:** Transformer architecture. Left part is the encoder, the right part is the decoder. T5's architecture is very similar to this one.

**LIST OF ABBREVIATIONS**

* **AI -** Artificial Intelligence
* **ML -** Machine Learning
* **BART -** Bidirectional and Auto-Regressive Transformers
* **HTTP -** Hypertext Transfer Protocol
* **IEP -** International Exposure Program
* **LSTM -** Long Short-Term Memory, a type of recurrent neural network
* **NLP** - Natural Language Processing
* **RNN** - Recurrent Neural Network
* **SEO -** Search Engine Optimization
* **API -** Application Programming Interfaces

**INTRODUCTION**

This project leverages the power of a large language model for paraphrasing and summarization tasks, we have used the T5 and FactorSum model for tasks, in the realms of academic professional writing these tasks are crucial and time consuming. Our aim is to provide a tool that helps with Paraphrasing, Summarization & proofreading. These models are very efficient and accurate in their domain.

However, to use these tools effectively, one must understand their working and limitations. These models are powerful but they are bounded by parameters like computational resources at disposal and quality of Dataset.

The focus of this project extends beyond mere implementation, we also delve into theoretical and architectural complexity of these models to select best for our use case and their implementations.

We aim to equip users with advanced paraphrasing tool tailored for scientific and research use unlike any!

**1.1 Background of the Study:**

The use of paraphrasing and summarization tools has become increasingly common in recent years. These tools can be used to save time and effort, and to improve the accuracy and clarity of text. However, there is still a lack of understanding about how these tools work, and how they can be used effectively.

The digital era has brought about significant changes in the way we handle information. These tools are software programs that have been designed to rewrite or condense text without altering its original meaning. They are particularly useful in several area such as Research , Content Creation and education , where large amount of information have to processed and presented in a concise and clear manner

**1.2 Statement of the Problem:**

Despite the growing popularity of paraphrasing and summarization tools, there is still a lack of understanding about how these tools work, and how they can be used effectively. This lack of understanding can lead to problems such as:

* The use of tools that are not appropriate for the task at hand.
* The use of tools that produce inaccurate or misleading results.
* The misuse of tools, which can lead to plagiarism or other ethical violations.

**1.3 Research Questions:**

This study will address the following research questions:

1. What are the different types of paraphrasing and summarization tools?
2. What are the theoretical underpinnings of paraphrasing and summarization?
3. How can paraphrasing and summarization tools be used effectively?
4. What are the limitations of paraphrasing and summarization tools?

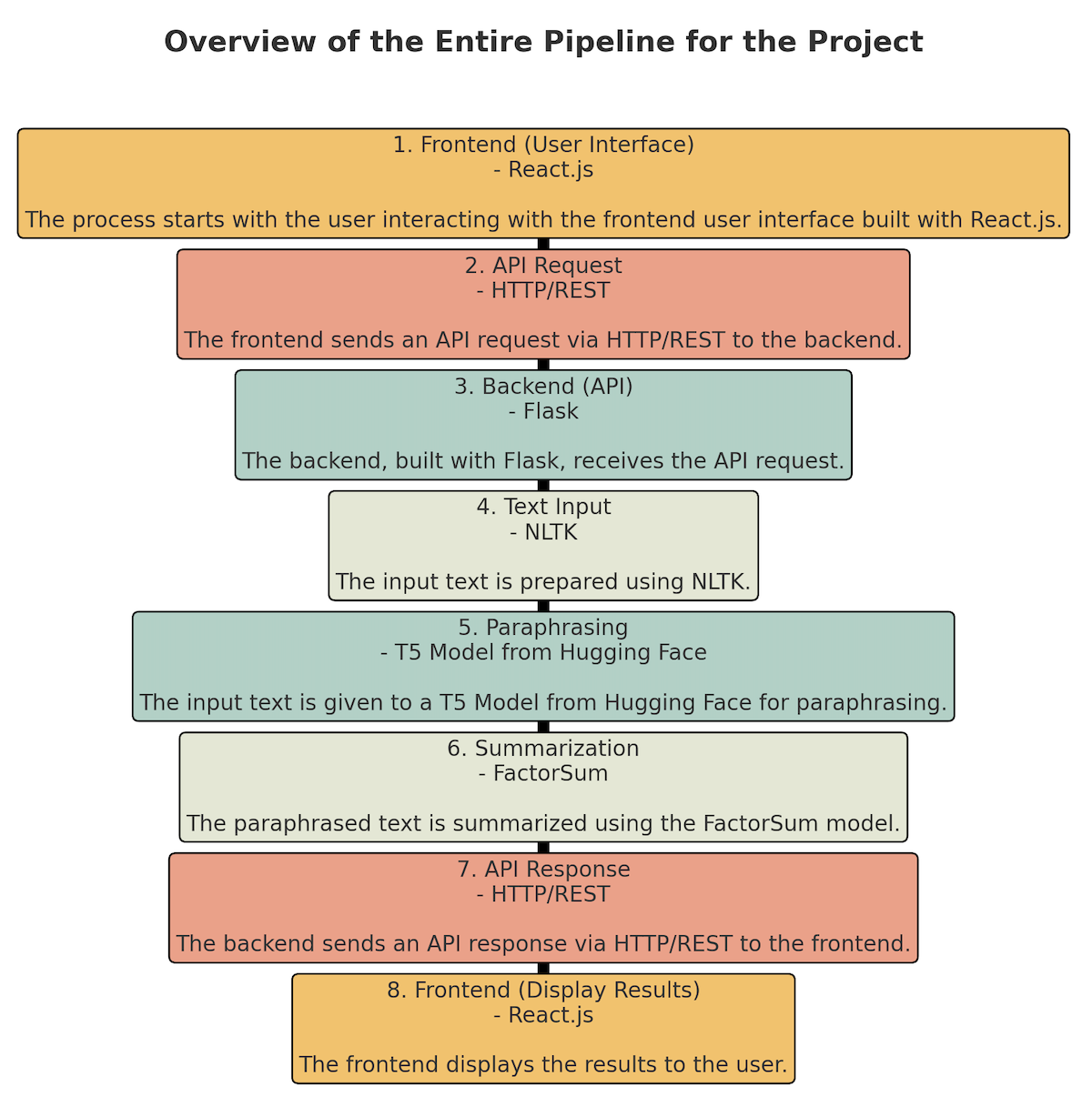
**1.4 Objectives of the Study:**

The objectives of this study are to:

* Provide a comprehensive overview of the background of paraphrasing and summarization tools.
* Discuss the theoretical underpinnings of paraphrasing and summarization.
* Explore the different approaches that can be used to implement paraphrasing and summarization tools.
* Identify the limitations of paraphrasing and summarization tools.
* Make recommendations for future research in this area.

**CHAPTER-2: DESCRIPTION**

The complete project was tackled by dividing it into segments, with each individual of the group focusing on the development of the specific part they were assigned. The project was primarily divided into three main sections: Front-End, Back-End, and Research. The main focus of the research was to discover a suitable pretrained model, along with a compatible dataset, that could be fine-tuned according to the specific requirements of our project. Subsequently, each of these sections details a distinct aspect of the project, including its implementation and the process involved.

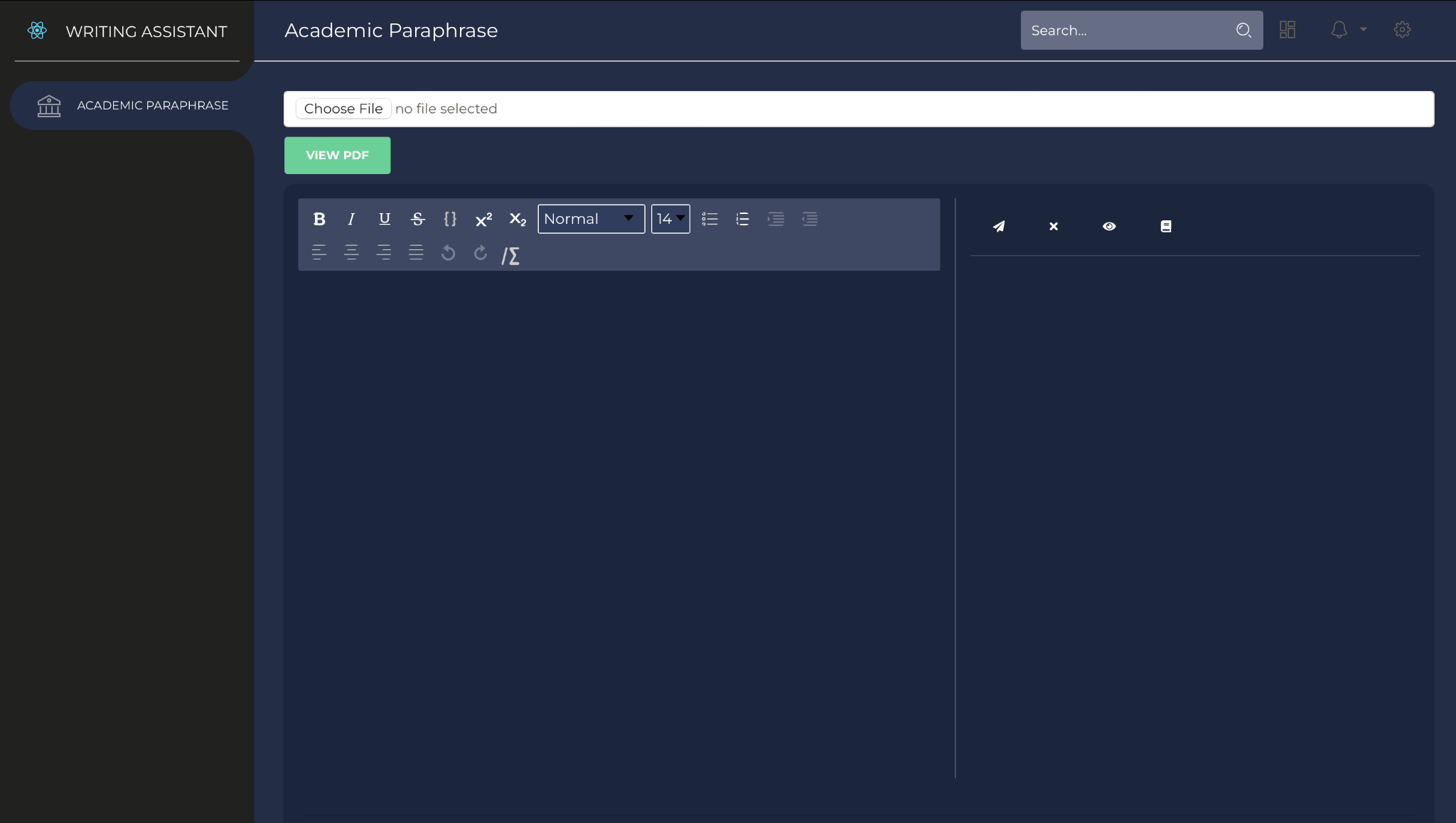
****

**Figure 2.1:** pipeline

**2.1 FRONTEND:**

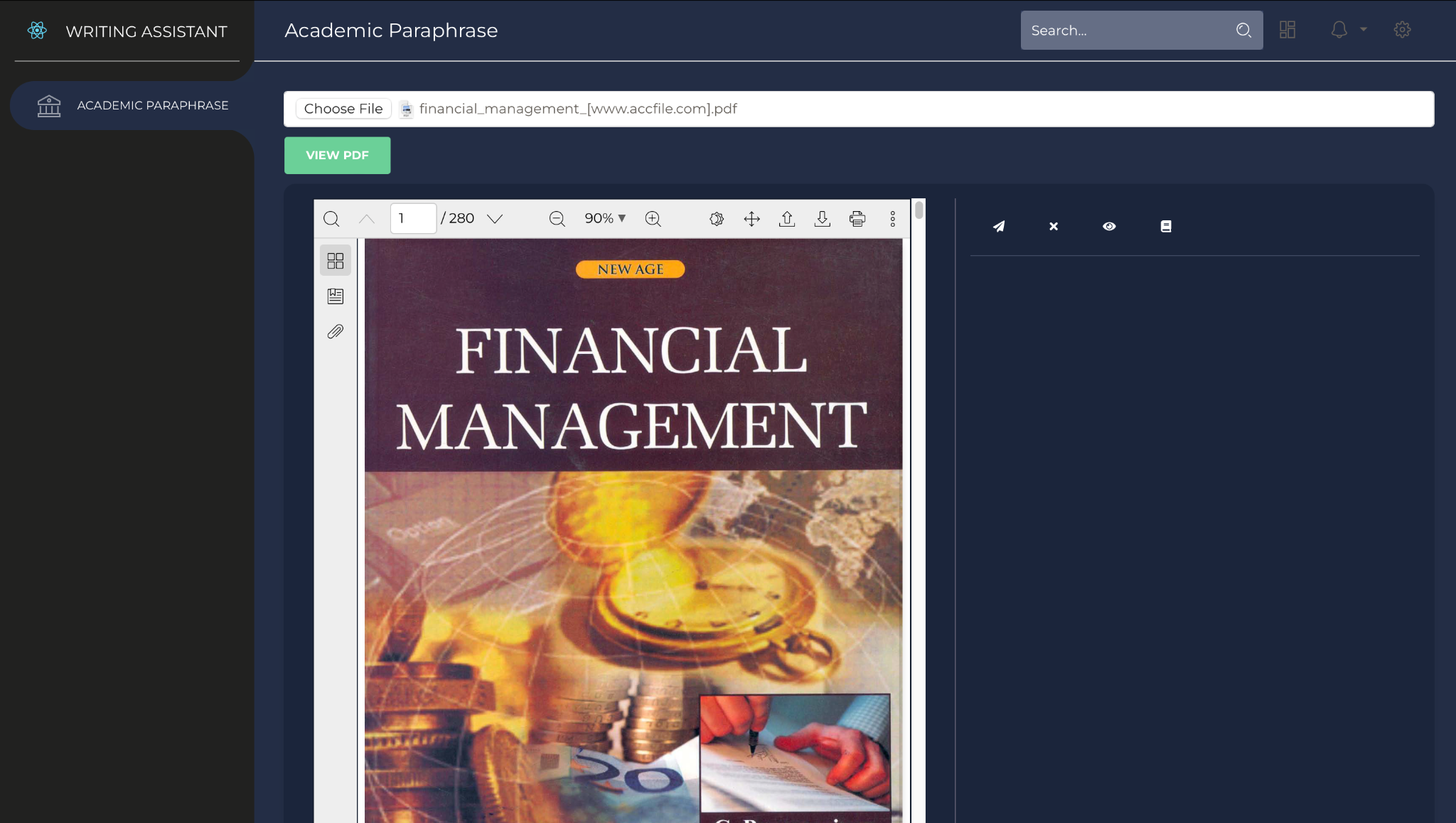
**2.1.1 Basic Features:**

* It contains some basic features like bold-italic-underline, alignment of words, adding bullets to sentences, indent-outdent and font-size. These features are used for basic changes in any text when a user wants to decorate their writing with the same features as any doc platform.
* There are some additional features like strikethrough, monospace, superscript, subscript and changing the block type of any sentence. It is a very user-friendly website that thinks about the needs of any user and it gives many different options.

****

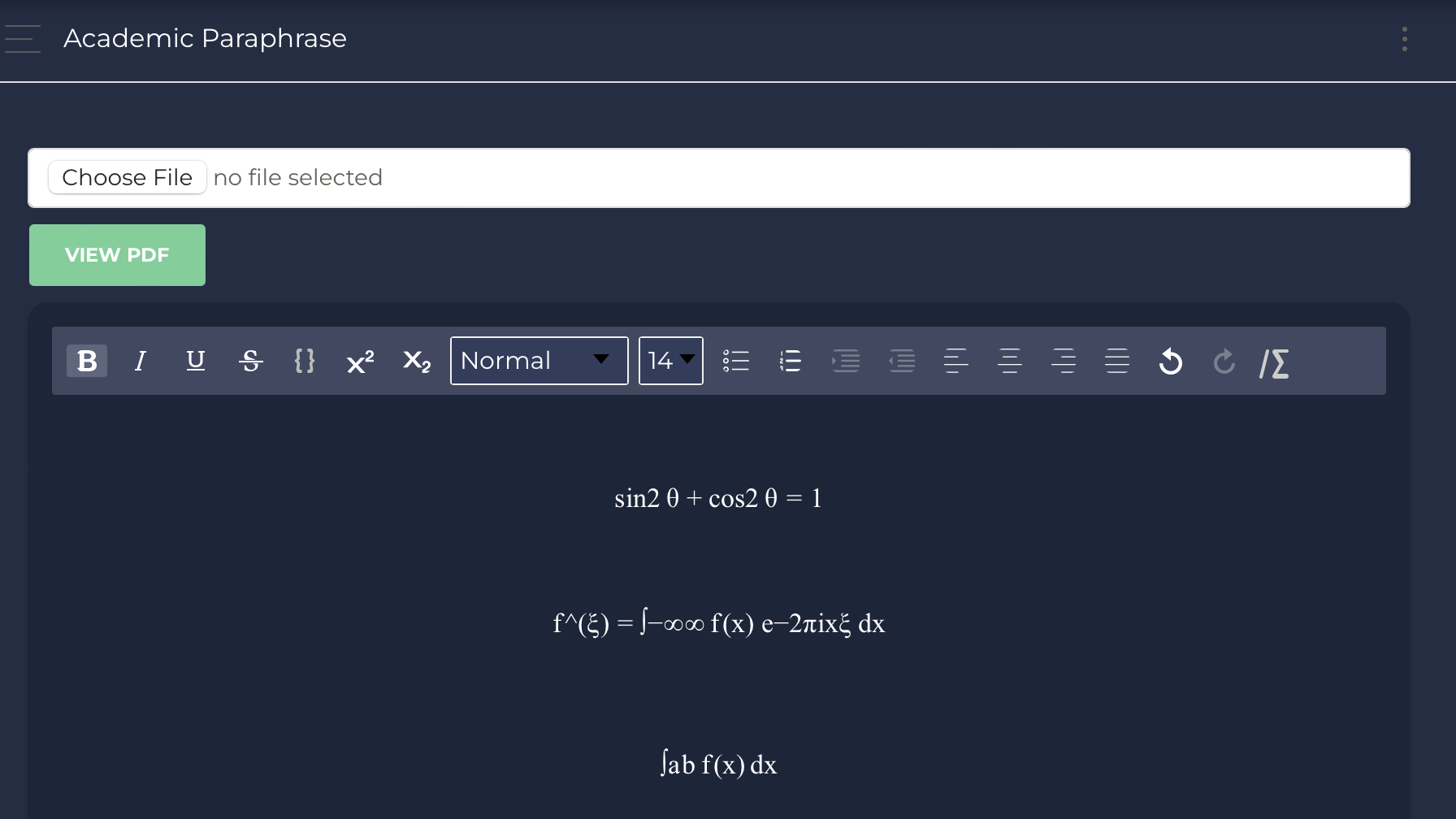
**Figure 2.2:** Homepage

* Furthermore, it has some special feature like uploading and reviewing a pdf. This feature has some important advantages like Enhanced User Experience, Time-Saving, Preserve Formatting, Maintain Source Context, Encourage Engagement, Expand User Base, Academic and Professional Use, Potential SEO Benefits, Analytical Insights, Competitive Advantage, etc.

****

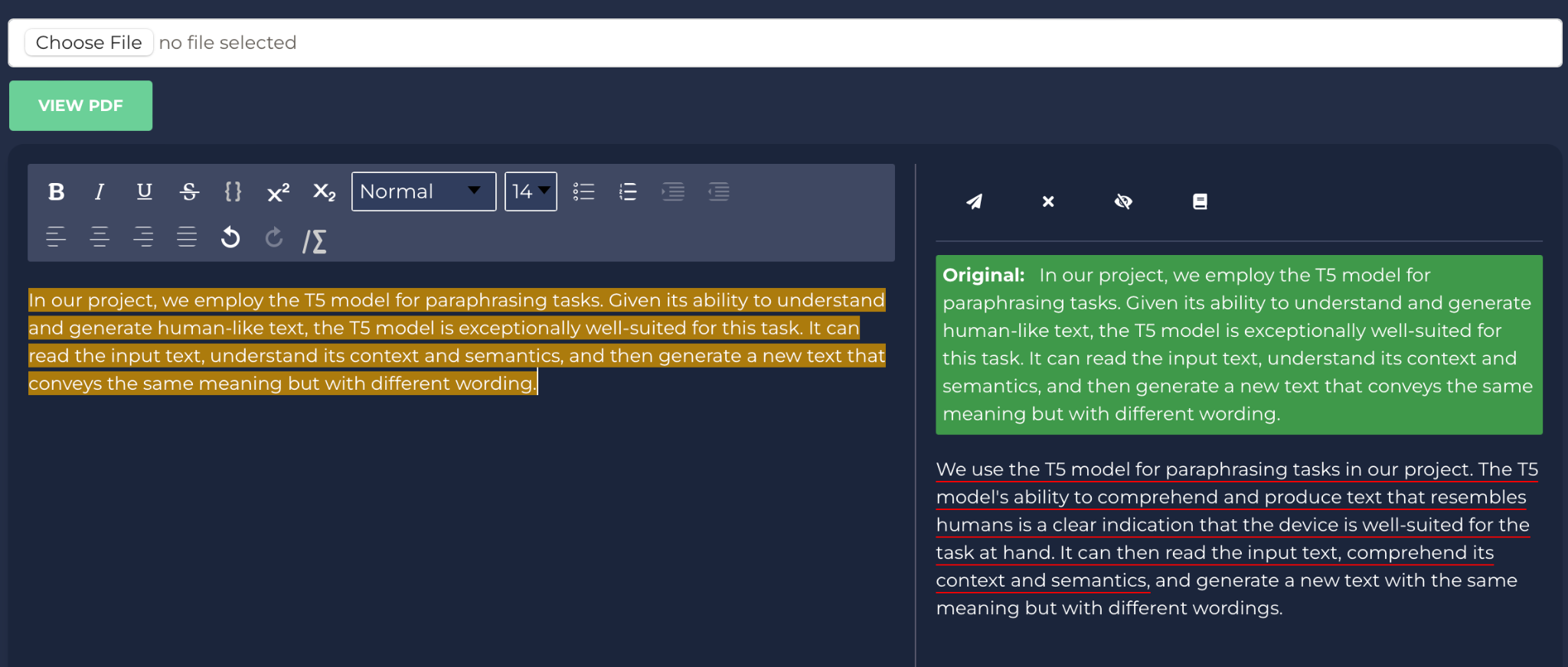
**Figure 2.3:** Pdf viewer

* Another special ability of this tool is writing latex symbols. And some known advantages of this ability are Mathematical Expression Support, Specialized Notation, Improved Academic and Technical Use, Time Efficiency, Avoiding Symbol Ambiguity, Consistency and Standardization, Cross-Platform Compatibility, Research and Technical Writing, Professional Appeal, Academic Integrity etc.

****

**Figure 2.4:** Latex Symbol

* It also underlines the changed part in any sentence, and it can also show the original sentence for comparison. This is a very necessary feature because when a user is writing a very big report or thesis they have many things to take care of so in this type of situation it is very important that the paraphrased part should be underlined for better understanding.



**Figure 2.5:** Paraphrasing

* It also contains a summarizing feature which can summarize many sentences and paragraphs by selecting them. This is an additional feature which we tried to add and we got success in that. Some advantages of this feature are Efficient Information Processing, Clarity and Conciseness, Complementary to Paraphrasing, Versatility, Language Learning, Accessible Content, User Engagement, etc.
* This paraphraser provides you a platform for writing any report, research paper, thesis, and articles at once at only one place without changing different platforms by just selecting and resolving the mistakes in it.

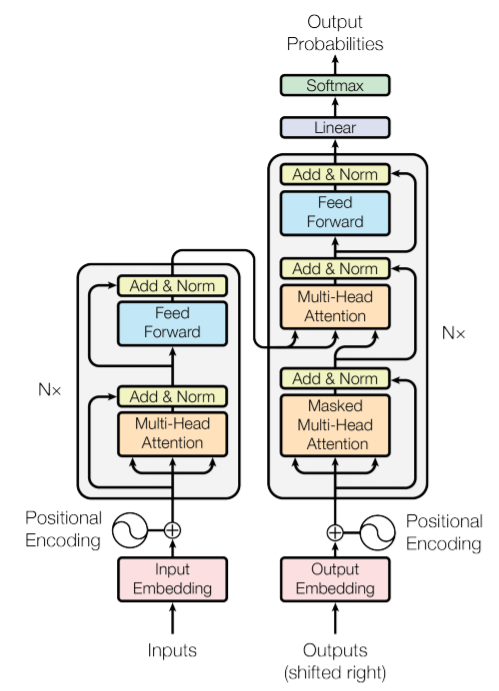
This tool was made to help the writers with a better platform for writing anything like reports, papers, novels, etc.

**2.1.2 Challenges and learnings:**

* For the frontend template ReactJS was used and as it was very new to us. SCRIMBA and some YouTube videos are used for learning purposes. It is a very good platform for newcomers where you can practice your codes within the video with the tutor.
* There was a major challenge of adding a pdf viewer in this and for that the learning was done by some basic codes from YouTube and completed with the help of mentors.
* And the main thing after making a frontend template is to connect that with the backend. For that some basic knowledge of axios was gained from YouTube which was used to send requests to backend and receive requests from backend.

**2.2 BACKEND:**

* Python was used to create the application's backend, and the Flask web framework was used. Python was a great option for creating the server-side logic because of its adaptability and simplicity. Being a micro-framework, Flask offered a simple and effective method for managing HTTP requests and responses. It made it possible for us to create seamless RESTful APIs for interacting with the front end and other external services.
* The development team used Flask blueprints to keep the codebase organized and modular. They were able to divide the programmed into more manageable, reusable components thanks to Flask blueprints. The application's features or functions were represented by individual blueprints, which made the codebase easier to maintain and grow. This architectural design pattern promoted developer cooperation and enhanced code clarity, allowing them to independently work on particular areas of the application.
* For paraphrasing text within the application, the team integrated the T5 model from Hugging Face's transformer library. T5, or "Text-to-Text Transfer Transformer," is a powerful language model capable of various natural language processing tasks, including paraphrasing. Leveraging T5, the application could generate alternate versions of input text, providing users with diverse and contextually similar phrasings for their content. This feature enhanced user experience, making the application more useful for content creation and curation.
* The backend offered summarizing functionality using the Factor Sum approach in addition to paraphrase. A sophisticated summary method called the Factor Sum model is used to extract key information from lengthy passages of text and condense it into succinct summaries. The backend of the programme might use this model to automatically generate summaries of books, documents, and other lengthy content. Users that needed quick access to the most important parts of lengthy content found this to be quite helpful, saving them time and effort and enabling them to process information more effectively.



**Figure 2.6:** Transformer architecture

**2.3 FACTOR-SUM MODEL:**

Introduced by Barzilay and Lee in 2004, the factor sum model is a statistical model used for the summarization of text. The principal belief behind this model is that an ideal summary should be a weighted summary of the most significant sentences from the source text. A factor analysis model determines the weight of each sentence by evaluating the semantic similarities between sentences and the summary's overall consistency.

The factor sum model has demonstrated its efficiency for summarizing diverse text types, encompassing news articles, academic research papers, and legal documents. The implementation of this model is quite straightforward and it requires a relatively minimal amount of data for training.

The process of the factor sum model can be broken down into a few steps. Initially, each sentence in the original text is represented as a vector of features. These features can stem from the words in the sentence, the words' part-of-speech tags, or the sentence's syntactic structure. Following this, a factor analysis model is trained on this set of sentence vectors. This model identifies a set of latent factors that encapsulate the semantic similarity among sentences. The final step involves generating a summary using the factor analysis model, where the most significant sentences from the original text are chosen and weighted according to the latent factors' values.

The factor sum model is an effective and simple method for text summarization. The model is not complicated to implement and requires a relatively small dataset for training. The model's effectiveness spans various text genres. However, it has its limitations. It can be sensitive to the selection of features, the training of the factor analysis model can be computationally expensive, and the model may fail to fully comprehend the intricate relationships among sentences. Nonetheless, the factor sum model stands as a straightforward and effective approach to text summarization with the demonstrated capability to handle a variety of text genres. Its potential limitations in capturing sentence relationships' full complexity and sensitivity to feature selection are the areas to keep an eye on.

**2.4 T5 MODEL:**

**Introduction:**

T5 stands for “Text to Text Transfer Transformer”, which is a transformer architecture-based model. We used the T5-base model for this project which has over 220 million parameters and works well with our application for paraphrasing. We used the ParaSCI dataset for fine tuning the model, which consists of around 350 thousand paraphrase pairs from the scientific academic papers.

The Transformer model was first introduced in research paper ‘Attention is All you need’ by Vaswani, which revolutionized the field of machine learning and AI. The paper proposed a new architecture that moved away from conventional RNN {recurrent neural network} & LSTM {long short-term memory neural network}.

The Transformer architecture was introduced with the concept of ‘attention mechanisms’ demonstrating that focusing on different parts of input sequence is crucial for better performance in tasks such as translation & summarization. This shift had drastically improved the performance of NLP models.

**Transformative Influence of the Transformer Model:**

The advent of the Transformer model, marked by the research paper 'Attention is All You Need', brought about a significant revolution in the sphere of machine learning and artificial intelligence. The architecture of this model introduced an innovative approach that veered away from conventional models like Recurrent Neural Networks (RNNs) and Long Short-Term Memory networks (LSTMs). A pioneering concept, known as 'attention mechanisms', was introduced with the Transformer model, emphasizing the importance of paying detailed attention to diverse segments of an input sequence when executing tasks like translation and summarization.

This paradigm shift, initiated by the Transformer model, profoundly enhanced the efficacy of Natural Language Processing (NLP) models. The innovative attention mechanism provided the model with the ability to focus selectively on certain parts of the input sequence that were more pertinent for each step in generating the output sequence. This method greatly optimized the sequence processing mechanism, thereby boosting performance in an array of NLP tasks. Undoubtedly, the Transformer model's transformative impact is profound as it not only changed our approach to sequence-related tasks but also established a new benchmark for NLP models.

**The Power of T5 Model in NLP:**

Exploring the T5 model, it expands upon the principles introduced by the Transformer model. Instead of dealing with different NLP tasks individually, the T5 model views all tasks as a text-to-text problem. Whether it's translation, summarization, question answering, or any other NLP task, the T5 model manages them uniformly: by transforming text input into a text output. This unified approach endows the T5 model with remarkable flexibility and adaptability to a variety of tasks.

**Application of T5 Model in Our Project:**

In our project, we employ the T5 model for paraphrasing tasks. The ability of the T5 model to interpret and generate text in a way that mirrors human communication is a clear testament to its suitability for the task. It can analyze the input text, understand its context and semantics, and subsequently generate new text that retains the same meaning but uses different verbiage.

**Seq2seq Architecture and Its Role:**

We tap into the T5 model's sequence-to-sequence (seq2seq) architecture for paraphrasing tasks. The seq2seq model processes an input sequence into a fixed-dimensional vector, which it then decodes into an output sequence. This complex process enables the T5 model to grasp semantic connections between words and phrases in the input text and generate an output text that maintains these connections but employs different wording. As such, the T5 model excels in tasks like paraphrasing, where deep comprehension of text context and semantics is essential.

**Paraphrasing with the T5 Model:**

In the context of paraphrasing, the T5 model takes the input text (source sequence), turns it into a semantic vector, and then converts this vector into a rephrased text (target sequence) that retains the original meaning. This seq2seq property significantly boosts the efficacy of the T5 model for this task.

**Denoising AutoEncoders and Their Functionality:**

A denoising autoencoder, a specialized type of autoencoder, employs a neural network model to reconstruct input data. Unlike regular autoencoders that attempt to replicate the input data, a denoising autoencoder uses a slightly distorted version of the input data to learn to rebuild the original, uncorrupted data. The model purposefully introduces noise to the input data, creating a distorted version. During the reconstruction phase, it tries to reverse this corruption. This forces the autoencoder to discern more robust and critical data patterns instead of simply copying the input to the output. In essence, the denoising autoencoder must understand the data structure to restore the original input from its noisy version.

**Conclusion:**

The T5 model is a testament to the significant strides being made in the domain of Natural Language Processing (NLP), courtesy of its integrated text-to-text methodology and its use of the Transformer's sequence-to-sequence architecture. This innovative model has been integral to our project, allowing us to tap into its potent capabilities to develop a paraphrasing tool that delivers both efficiency and effectiveness.

The utility of the T5 model within our work reflects its potential within the broader field of machine learning. As we continue to witness advancements in this domain, it is clear that models like T5 are destined to play a pivotal role in steering the course of future NLP applications. These models, with their robust capabilities, are shaping up to become foundational elements of next-generation language processing tools.

**Furthermore:**

This transformation isn't just confined to theoretical models or controlled environments. Practical applications, such as the one we're developing in our project, serve as a tangible demonstration of these models' power and versatility. In this case, the T5 model enables us to create an effective paraphrasing tool, one that could have wide-ranging applications in areas such as content creation, academic research, and more.

As we forge ahead in the field of machine learning, the significance of models like T5 only continues to grow. Their utility in a variety of applications underscores their potential, marking them as key contributors to the ongoing evolution of NLP applications. As we unlock further capabilities and refine these models, we can expect them to reshape our understanding and interaction with language processing in unprecedented ways.

**2.5 FUTURE WORK:**

Because of the project's time constraints, we could only train the model to a certain extent. Our objective is to build upon the initial framework and gradually enhance the model's efficiency and accuracy by training it with diverse training examples. Additionally, we plan to introduce a question answering feature to assist researchers in reviewing academic papers and eventually streamline their own paper writing process. We are also considering utilizing other language models such as BART and comparing their efficiencies to determine the most suitable model for the project.

**REFERENCES**

* Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998-6008.
* Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 22, 1629-1640.
* Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735-1780.
* Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.
* M. Fonseca, Y. Ziser, and S. B. Cohen, "Factoring Content and Budget Decisions in Abstractive Summarization of Long Documents," arXiv preprint arXiv:2205.12486v2, 2022.
* DataChef Team. (2020). Paraphrasing and style transfer with GPT-2. *arXiv preprint arXiv:2004.01069*
* Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). T5: Text-to-text transfer transformer. arXiv preprint arXiv:2006.16163.

**APPENDIX**

**Lectures:**

1. **Orientation:**

The introductory session provided us with valuable insights into the cultural norms and practices of Thailand, including essential dos and don’ts, public manners, planned itineraries, outings, accommodations, food options, and even an introduction to the Thai language. Additionally, the lecture shed light on the layout of the campus, enhancing our understanding of the facility. We extend our sincere gratitude to Mr. **Ranadheer Reddy** for his guidance during this enlightening orientation.



1. **Fintech evolution & Solution**

This engaging lecture integrated the fields of technology and finance, offering students a comprehensive overview of the fintech revolution. It shed light on how fintech has paved the way for microtransactions, thereby connecting families and dismantling the complex, orthodox practices of traditional banking. The lecturer, Ms. **Aishwarya Kapoor**, skillfully served as a bridge between the intricate concepts of fintech and the students, making the subject both accessible and enlightening.



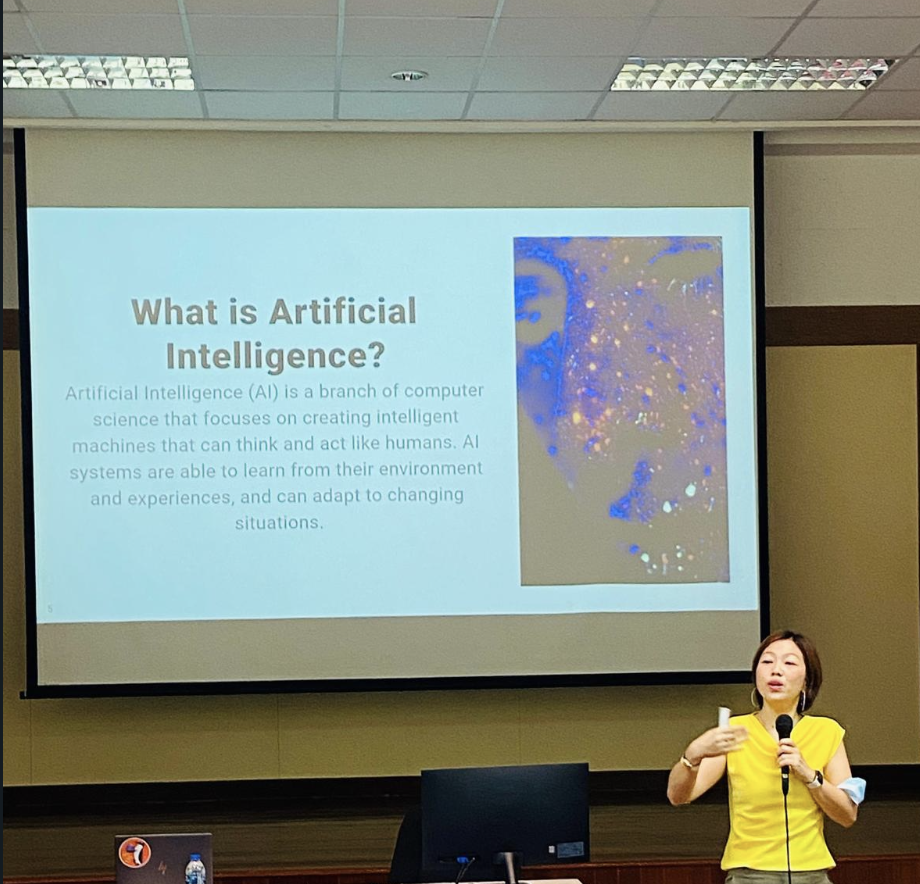
1. **AI & ChatGPT**

The lecture commenced with a demystification of AI terminologies, aiming to remove any abstractness and make the subject more accessible to all attendees. **Dr. Chutiporn Anutariya's** approach was inclusive, tailoring the content to be comprehensible to individuals from various backgrounds, including commerce and science, not just computer science.

A chronological exploration of AI's history was presented, starting with IBM's famous AI that defeated a world champion in a board game. The narrative then transitioned to the early 2000s, highlighting Sony's pet dog AI, followed by the more recent developments like Alexa and Google Assistant.

Dr. Anutariya skillfully navigated the AI timeline, elucidating key milestones and innovations that have shaped the field. The lecture's design allowed for an engaging and well-rounded overview, emphasizing the transformative impact of AI technologies.

The session culminated with a focus on ChatGPT, exemplifying the contemporary advancements in generative AI. By connecting historical context with current trends, the lecture inspired attendees to appreciate the multifaceted nature of AI and envision its future potentials.

****

1. **Career development:**

There’s a fine quote by Mahatma Gandhi that “A man is but a product of his thoughts.” And the lecture begins with a question that ‘what is the purpose of life?’

And answer of that is that the work you are doing should be satisfying and meaningful to you.

There is a very important thing for people who are looking for more money and think that for a good life you have to earn more money that if you are not sure about that thoughts you should ask 3 questions to yourself and that are:

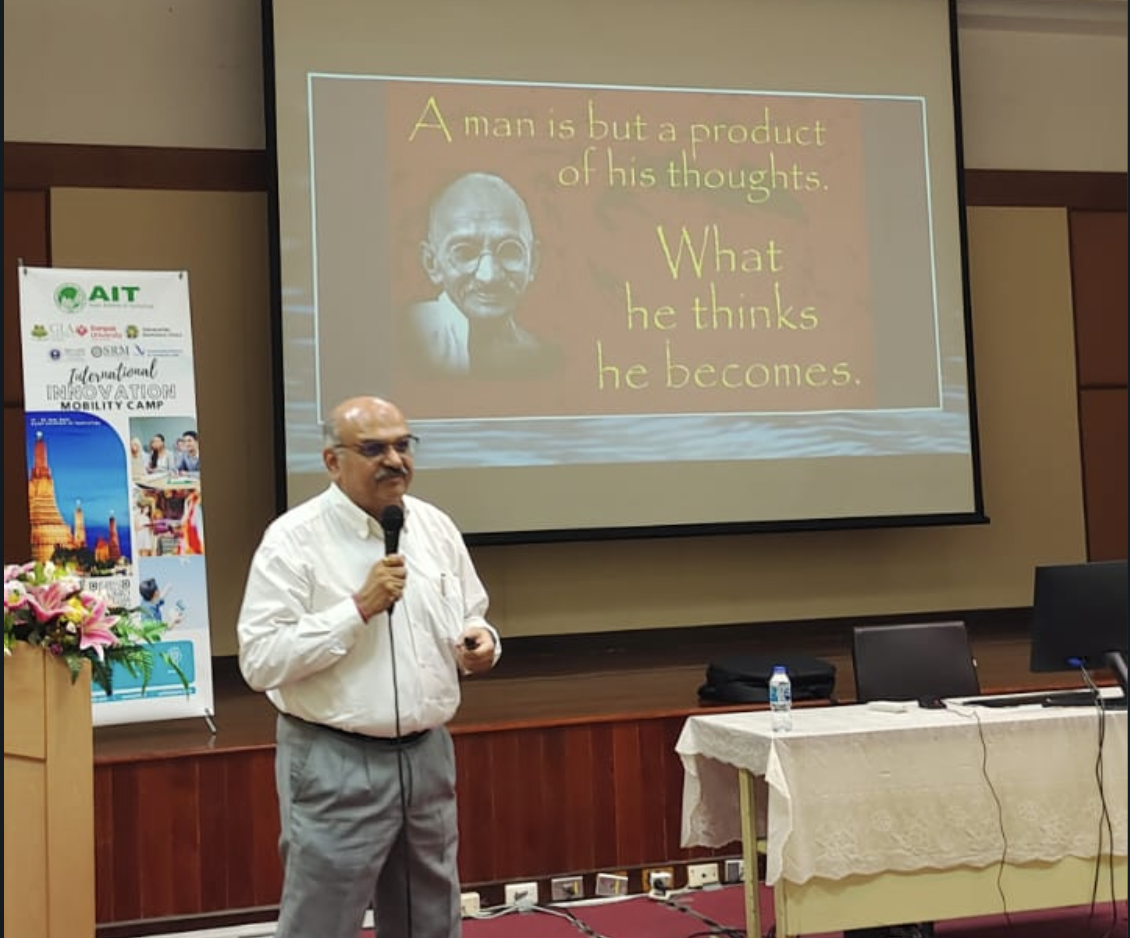
1. What is good life?
2. What is true value of money?
3. What should I not do these?

There are 3 types of goods which affects the life of any person:

1. Bandwagon goods: goods that other people have them.
2. Snob goods: goods that other people don’t have them.
3. Veblen goods: goods which are expensive.

And the lecture concluded with concept of IKIGAI:



****

**Visits:**

1. **BART Lab:**





1. **National Science Museum:**





