WriteWizard - Collaborative Document Editing Tool: Real-Time Multi-functional Platform

(Automatic IEEE formatter for collaborative documents)

TMP-24-25J-146

Project Proposal Report

Shandeep. J – IT21375682

B.Sc. (Hons) in Information Technology Specializing in Software Engineering

Faculty of Computing

Sri Lanka Institute of Information Technology
Sri Lanka

August 2024

WriteWizard - Collaborative Document Editing Tool: Real-Time Multi-functional Platform

(Automatic IEEE formatter for collaborative documents)

TMP-24-25J-146

Project Proposal Report

B.Sc. (Hons) in Information Technology Specializing in Software Engineering

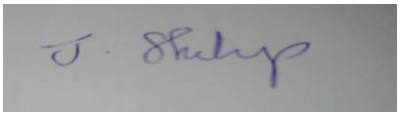
Faculty of Computing

Sri Lanka Institute of Information Technology
Sri Lanka

August 2024

Declaration

I declare that this is my own work, and this proposal does not incorporate without acknowledgementany material previously submitted for a degree or diploma in any other university or Institute of higher learning and to the best of our knowledge and belief it does not contain any material previously publish or written by another person expect where the acknowledgement is made in the text.

Name	Student ID	Signature
Shandeep. C	IT21375682	

The above candidate is carrying out research for the undergraduate Dissertation under my supervision.

……..

Signature of the Supervisor:

(Dr. Lakmini Abeywardhana) Date:22/08/2024

……..

Signature of the Co-supervisor:

(Ms. Karthiga Rajendran) Date:22/08/2024

Abstract

In the realm of academic writing, precise adherence to formatting standards such as IEEE is essential for the clarity and trustworthiness of scholarly work in the field of academic writing depends on strict adherence to formatting standards like IEEE. Nevertheless, the manual formatting procedure is frequently labor-intensive, which increases the risk of mistakes and inefficiencies. The powerful collaborative editing tool presented in this research automatically formats documents into IEEE format and offers context-aware formatting recommendations in real time.

In order to ensure that publications follow the necessary structure, citation style, and layout, the tool uses an optimized Language Model (LLM) to assess and apply the complex requirements of IEEE formatting with ease. Furthermore, the application provides ongoing, customized formatting advice to users as they write and modify their work, adjusting to their own writing preferences and style. This dynamic feedback loop does more than just make writing easier.

Researchers and students can concentrate more on content production than the time-consuming components of formatting by adopting this intelligent document formatting solution. By lowering formatting-related difficulties, increasing document preparation speed, and guaranteeing that submissions fulfill publication standards with little effort, this tool seeks to assist the academic community.

Table of Contents

Contents

List of Figures

List of Tables

List of Abbreviations

NLP	Natural Language Processing
NER	Named Entity Recognition
IEEE	Institute of Electrical and Electronics Engineers
LLM	Large Language Model
FLAN-T5	Fine-Tuned Language Model (FLAN) based on the T5 architecture
NER	Named Entity Recognition
ML	Machine Learning

1.0 Introduction

1.1 Background and Literature

Academic writing is a cornerstone of scholarly communication, demanding strict adherence to specific formatting guidelines such as IEEE standards. While these guidelines ensure consistency, clarity, and professionalism in research papers, manually formatting documents can be a tedious and error-prone process. Historically, authors have had to meticulously follow intricate formatting requirements [1], which can divert valuable time and attention away from the primary task of content creation. As a result, there is a growing demand for tools that can automate the formatting process while maintaining accuracy and adherence to these standards.

In recent years, several automated formatting solutions have been developed to assist authors in aligning their documents with specific style guidelines. However, many of these tools have limitations, often falling short in handling the complex requirements of academic formats or lacking real-time feedback mechanisms. Existing tools typically rely on static templates that do not adapt to the unique needs of individual users or evolve with changes in document content.

This project aims to bridge this gap by introducing an advanced Intelligent Document Formatter that provides real-time, context-aware formatting recommendations and automates the conversion of manuscripts into IEEE format. Leveraging the capabilities of the FLAN-T5 model, the application is designed to understand the intricacies of layout, citation styles, and document structure while offering continuous, personalized formatting assistance. Key features include the use of NLP techniques for content parsing and understanding, dynamic formatting conversion, real-time personalized suggestions, and automated compliance checks to ensure IEEE standards are met. By integrating these advanced functionalities, this tool seeks to enhance the accuracy and efficiency of academic writing, supporting researchers and students in producing high-quality, publication-ready documents with ease.

.

1.2 Research Gap

Significant research gaps remain despite advances in AI-driven document formatting, especially when it comes to automatically converting academic articles to IEEE format utilizing Large Language Models (LLMs). A primary obstacle is the scarcity of accurate, optimized models created especially for IEEE formatting. The complex formatting guidelines mandated by IEEE standards, such as precise citation styles, figure placements, and section organization, may not always be effectively applied by the current models since they are frequently generalized.

The scarcity of comprehensive datasets formatted for academic use is another significant gap. The majority of datasets used to train LLMs lack the specific instances required for exact formatting tasks, instead concentrating on generic text processing. The model's capacity to acquire and apply intricate formatting rules to a variety of document types is impacted by this constraint.

Moreover, although promising, real-time formatting adjustments and suggestions have difficulties accommodating individual user preferences while strictly adhering to IEEE standards. The proper balance between standardization and personalization is still poorly understood, thus more study is required to create models that can adapt to user needs dynamically without sacrificing formatting accuracy.

Additionally, ethical questions surface, namely in relation to the dependability and openness of formatting tools powered by AI. Users need to have faith that these technologies will not only adhere to the rules but also deliver repeatable, consistent outcomes across a range of documents. In order to increase the use of LLM-based formatting[6] tools in academic contexts and ultimately improve the effectiveness and precision of scholarly publishing, it is imperative that these research gaps be filled.

For this project, data collection will focus on generating real-world examples from various academic documents to train and validate the model, with particular attention to adapting these tools for the specific needs of the Sri Lankan academic context.

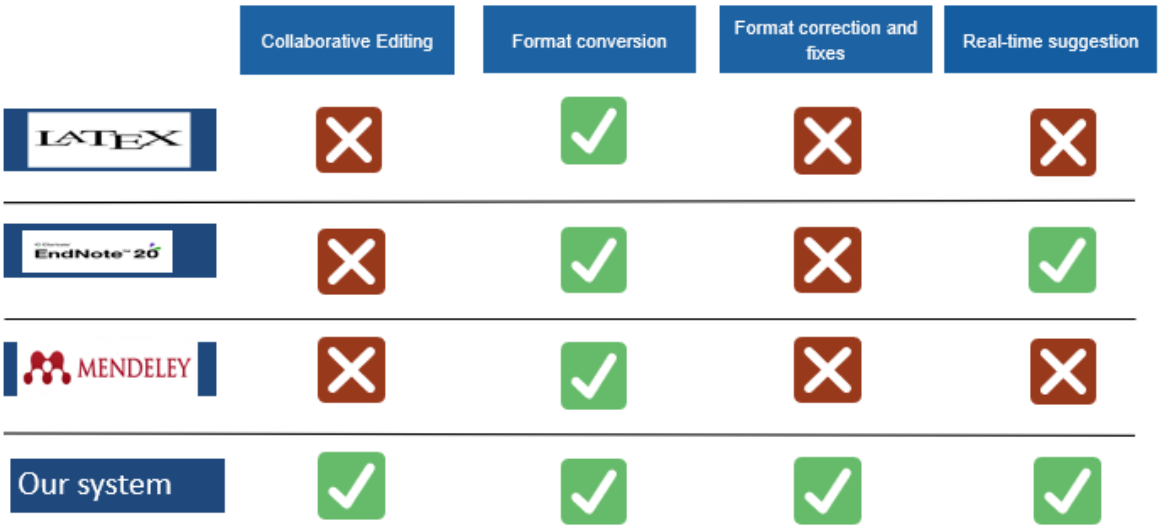


Figure 1: Competitive Analysis

It was assumed that filling these gaps will further develop more varied and reliable, which will promote the concept of highly effective real-time collaboration, increased focus on content, beyond-measure flexibility, and formatting accuracy.

1.3 Research Problem

As the demand for scholarly publications adhering to IEEE guidelines increases, the need for precise and consistent document formatting has become crucial. Researchers, academics, and students often face significant challenges when manually formatting their documents to meet IEEE standards. This process is not only time-consuming but also prone to errors, requiring a deep understanding of complex formatting rules.

Although AI-powered tools have been developed to assist with document formatting, existing solutions are inadequate for fully automating the conversion of documents into IEEE format. These tools frequently struggle with handling complex formatting elements such as equations, figures, tables, and proper citation styles. Moreover, they often fail to maintain consistency across different sections of a document and lack the ability to provide real-time formatting suggestions that adapt to the user's evolving document structure while ensuring strict adherence to IEEE standards.

This research seeks to address the critical need for an intelligent document formatting tool that can accurately convert academic papers into IEEE format. The project aims to develop a system that not only automates the formatting process but also offers real-time, personalized formatting suggestions tailored to the user's specific needs. By Large Language Models (LLMs) and advanced machine learning techniques, this tool will significantly reduce the manual effort required by researchers and students, ensuring their work meets the stringent

2.0 Objectives

2.1 Main Objective

The project aims to automate formatting documents to IEEE standards and providing real-time personalized suggestions. It also performs automated compliance checks to ensure and fix IEEE formatting issues.

2.2 Specific Objectives

The following are the sub-objectives for this project:

- **Content Parsing and Understanding:** Implement NLP techniques to extract and comprehend document content, ensuring accurate application of IEEE formatting standards.
- **Formatting Conversion:** Utilize the FLAN-T5 model to convert document content into IEEE format seamlessly.
- **Personalized Formatting Suggestions:** Offer real-time formatting recommendations based on user edits, with options to request suggestions at any point during document creation.
- **Automatic Formatting Issue Fixes:** Detect and automatically correct formatting issues to maintain IEEE compliance throughout the document.

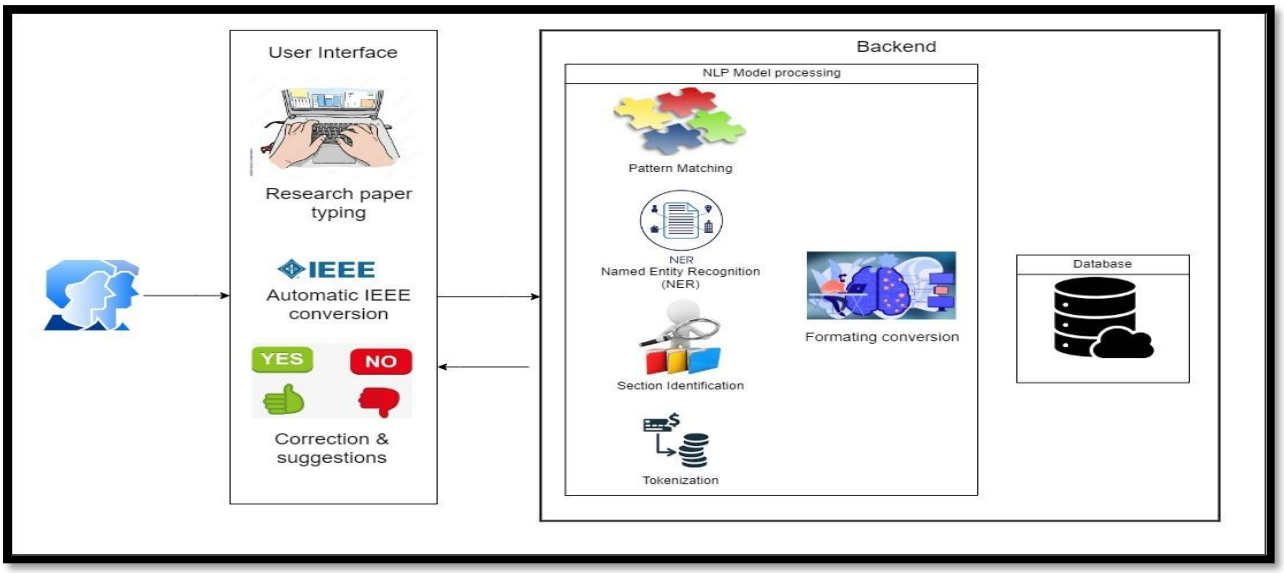


Figure 2: System Diagram

3.0 Methodology

requirements of IEEE publications without extensive manual intervention.

3.1 Requirement Gathering

Requirement gathering was through performing an extensive analysis of past research conducted throughout recent years, identification and analysis of the existing systems, as well as reading through a variety of online resources. Also, some real-world scenarios were used to figure out this research problem and requirement gathering.

3.1.1 Past Research Analysis

In the field of automated document formatting, particularly for IEEE standards, numerous research papers and studies have been conducted, focusing on various aspects such as natural language processing (NLP), machine learning, and document parsing techniques[2]. However, there is a relatively smaller body of research that directly addresses the challenges of fully automating the conversion of academic documents into IEEE format, especially with real-time personalized formatting suggestions.

The past research analysis primarily focused on identifying the methodologies and technologies used in existing tools and platforms for document formatting. Key areas of interest included the application of Large Language Models (LLMs) like FLAN-T5, document parsing libraries (e.g., python-docx, PyMuPDF), and techniques for ensuring compliance with complex formatting standards such as those set by IEEE. This analysis also shed light on the limitations and challenges faced by previous researchers, such as the difficulty in handling intricate formatting elements, maintaining consistency across different document sections, and providing adaptive, real-time formatting recommendations.

Understanding these challenges and the tools used in previous research has been instrumental in shaping the development of a more robust and intelligent document formatting system. The insights gained from past research have informed the approach of this project, particularly in addressing the gaps in current solutions and enhancing the accuracy of IEEE format conversion and compliance.

3.2 Feasibility Study

3.2.1 Technical Feasibility

3.2.1.1 Knowledge on Technologies

To develop the proposed automated document formatting solution with real-time personalized suggestions and automatic formatting issue fixes, the following technologies and tools are essential:

- **Natural Language Processing (NLP) Techniques:** NLP is critical for extracting and understanding the content of documents. Techniques such as Named Entity Recognition (NER), syntax parsing, and tokenization will be used to accurately interpret the structure and content of the documents to ensure they conform to IEEE standards.
- **Large Language Models (LLMs) like FLAN-T5:** LLMs such as FLAN-T5 are used for converting document content [3] into the IEEE format. The model will be fine-tuned on a diverse dataset to accurately apply IEEE formatting rules and make personalized suggestions based on user edits.
- **Document Parsing Libraries (e.g., python-docx, PyMuPDF):** These libraries are essential for reading, manipulating, and writing documents in various formats. They will be used to parse the content, detect formatting elements[6], and apply the necessary formatting transformations to conform to IEEE standards.
- **Machine Learning for Formatting Suggestions:** Machine learning algorithms will be implemented to analyze user preferences and document characteristics in real-time. These algorithms will generate personalized formatting suggestions that adapt to the evolving structure of the document.
- **Automated Compliance Checking:** The system will incorporate algorithms to automatically detect and correct formatting issues, ensuring compliance with IEEE standards. Techniques such as rule-based systems, pattern matching, and statistical methods will be employed to verify and adjust the formatting[5].
- **Real-time Processing and Integration:** To offer real-time suggestions and corrections, the solution will leverage technologies like real-time communication protocols to provide instant feedback and updates as the user edits the document.

3.2.1.2 Knowledge on Tools

To develop the proposed solution, it is essential to have a comprehensive understanding of the following tools:

1. **Python-docx**: Utilized for parsing and manipulating Word documents programmatically, which is crucial for extracting and applying formatting styles.
2. **FLAN-T5 Model**: A fine-tuned language model [4] used for generating personalized formatting suggestions and automating the conversion process into IEEE format.
3. **NLP Techniques**: Employed for content parsing, understanding, and ensuring the accuracy of document formatting suggestions.
4. **Version Control Tools**: Essential for managing code and collaborating effectively on different aspects of the project.

3.2.1.3 Data collection Knowledge

Effective data collection and preprocessing are vital for the project:

1. **Data Collection**:
   - **Document Samples**: Collecting academic documents, including IEEE papers, for model training and testing.
2. **Data Preprocessing**:
   - **Content Parsing**: Structuring document data for accurate formatting.
   - **Data Cleaning**: Removing inconsistencies and errors.
   - **Feature Extraction**: Identifying key formatting elements for precise application.

3.2.2 Schedule Feasibility

This project will be developed independently and aligned with the proposed timeline. The project schedule is carefully planned to ensure timely completion, with milestones and deadlines clearly outlined. This timeline includes phases for research, development, testing, and implementation, ensuring that the system is delivered within the set timeframe. Detailed schedule management plans are provided to track progress and adjust as needed to meet project goals.

3.2.3 Economic Feasibility

The proposed solution is designed to be cost-effective, aiming to minimize expenses while delivering high functionality. Compared to existing solutions, this approach is budget-friendly, though costs may vary with economic conditions.

3.3 System Analysis

3.3.1 Software Solution Approach

Below is the approach to the solution development for automating document formatting to IEEE standards with real-time personalized suggestions and automatic formatting issue fixes. Figure 5 illustrates the workflow of the solution.

1. **Data Collection:**
    - Collect sample academic documents in various formats (e.g., IEEE, Harvard) from different sources.
    - Analyze existing academic documents to identify common formatting patterns and anomalies.
2. **Document Parsing:**
    - Extract text and formatting information from the uploaded documents.
    - Identify key document elements such as headings, subheadings, citations, and figures.
3. **Preprocessing:**
    - Normalize the document content to a standard format (e.g., consistent font size, style).
    - Detect and extract structural elements such as sections, tables, figures, and citation formats.
    - Align document elements with the target format (e.g., IEEE).

4. **Feature Extraction:**
    - Extract meaningful features related to document formatting: a. **Document Structure:** Headings, subheadings, sections, and paragraph styles. b. **Formatting Elements:** Citation styles, bibliography format, figure captions, and table formats.
5. **Formatting Conversion:**
    - Map extracted features to the formatting rules of the target style[7] (e.g., IEEE).
    - Use predefined formatting templates to guide the conversion process.
    - Integrate the FLAN-T5 model to transform document content into IEEE format.
6. **Personalized Formatting Recommendations:**
    - Provide real-time formatting recommendations based on user edits and preferences.
    - Continuously update suggestions based on document changes and user interactions.
7. **Automatic Formatting Issue Fixes:**
    - Detect formatting issues using NLP techniques.
    - Automatically apply corrections to ensure IEEE compliance.
8. **Model Training:**
    - Train the formatting conversion model using labeled datasets that include documents in various formats.
    - Employ machine learning techniques to learn the mapping between source and target formats.
9. **Model Evaluation:**
    - Evaluate the performance of the formatting conversion model using metrics such as accuracy, consistency, and user feedback.
    - Perform iterative testing and refinement to improve the model's effectiveness in handling diverse document formats.

3.3.2 Tools & Technology

- **Programming Languages:**
  - Python
- **Frameworks:**
  - Keras
  - TensorFlow
  - Flask or Django (for web backend, if needed)
  - FLAN-T5 (for NLP and formatting tasks)
- **Document Parsing Libraries:**
  - python-docx
  - PyMuPDF
- **Web Development Frameworks:**
  - **Front End Development:** React JS
  - **UI Components Design:** MUI (Material-UI)
- **UI Design Implementation:**
  - Figma
- **Diagramming Tool:**
  - Draw.io
- **Integrated Development Environment (IDE):**
  - VS Code
- **Version Control System (VCS):**
  - Git (GitHub or GitLab)
- **Collaboration Tools:**
  - Microsoft Teams

    - WhatsApp
- **Testing Tools:**
    - **Unit Testing for Python:** pytest
    - **API Testing:** Postman
    - **JavaScript Unit Testing:** Jest
- **Deployment Tools:**
    - Docker (for containerization)
    - Kubernetes (for orchestration)
    - Jenkins (for automation)
    - Ansible (for configuration management)
- **Code Quality Assurance:**
    - SonarQube
- **Project Management:**
    - MS Planner
    - Jira

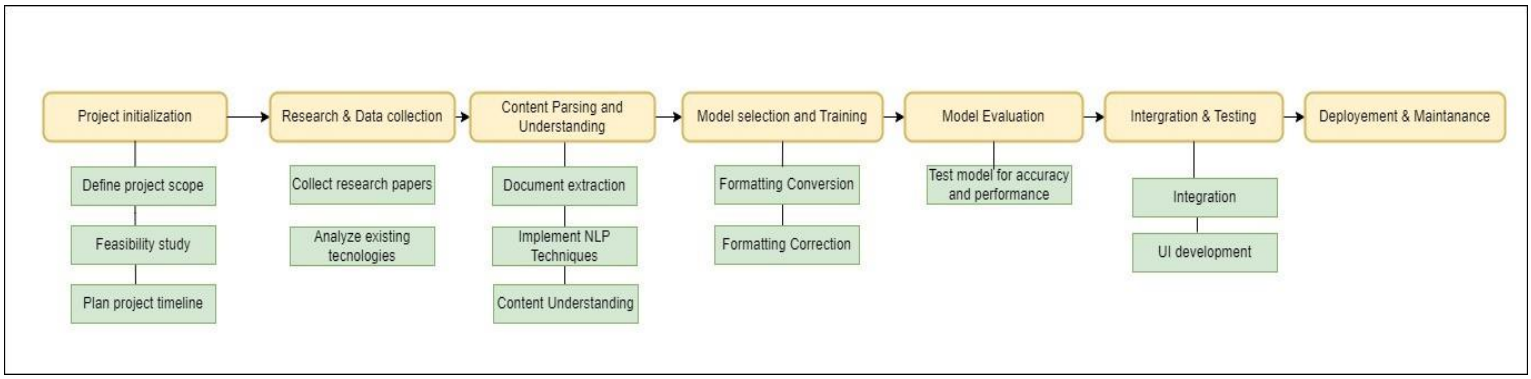


Figure 3: Work Breakdown Structure

3.4 Project Requirements

3.4.1 Functional Requirement

The functional requirements for the proposed model are as follows:

- Accurate Documents processing
- Convert document content into IEEE format using a Trainned model
- Provide real-time formatting recommendations based on user edits
- Identify formatting issues that deviate from IEEE standards and apply fixes

3.4.2 Non-Functional Requirements

The non-functional needs that were prioritized throughout the creation of the suggested model are listed below

- Performance
- Real-Time Processing
- Usability
- Data Protection

3.5 Testing

Testing is a critical component of the project's success. The initial testing phase will involve conducting unit and system testing by applying the solution to a range of sample documents, including those from peers and other academic sources. This stage will ensure that the tool correctly formats documents according to IEEE standards and provides accurate real-time suggestions.

Subsequently, the solution will undergo Acceptance Testing, including both Alpha and Beta testing phases. Alpha testing will be performed in a controlled environment, while Beta testing will involve real-world users such as students and researchers to validate the tool's effectiveness in various scenarios. Testing will primarily be manual, supported by automated tests for key functionalities, including the accuracy of formatting adjustments and the responsiveness of real-time suggestions

3.6 Timeline

The proposed timeline for the project is as follows

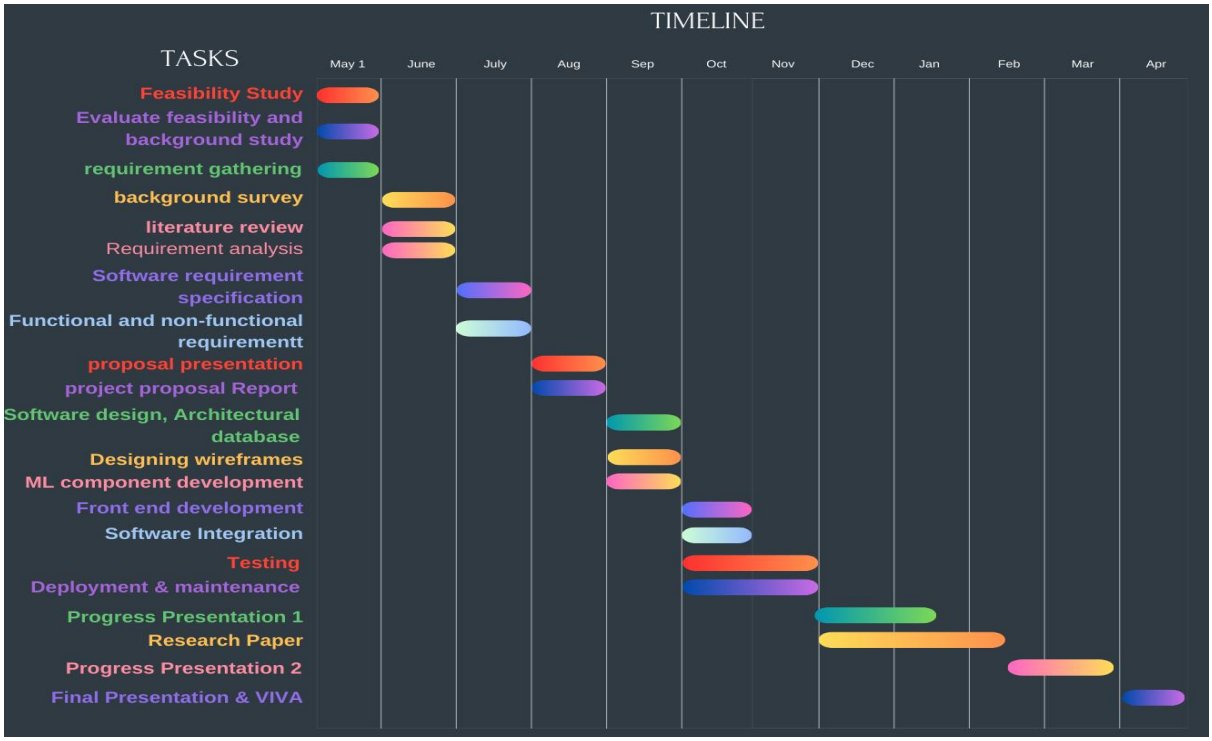


Figure 4: Gantt chart

3.7 Communication Management Plan

The Communication Management Plan ensures that all relevant stakeholders, including supervisors and any external collaborators, receive timely and accurate information throughout the project's lifecycle. The success of the project hinges on the effective planning and execution of communication strategies.

This plan will detail the communication methods for each stakeholder group, outlining the audience, content, format, frequency, and expected outcomes for each communication item. It will also define the roles and responsibilities of the stakeholders, assigning tasks and establishing a communication strategy based on their influence, interests, and expectations in the project. Regular updates, progress reports, and feedback loops will be implemented to keep all parties informed and engaged, ensuring alignment with project goals and timelines.

3.7.1 Communication Objectives

Proactive communication is vital for the success of this project. The communication approach will focus on the following objectives:

- **Adequate:** in the right format and right content
- **Specific:** for the targeted audience.
- **Sufficient:** providing all the necessary information.
- **Concise:** brief, avoiding repetition and non-important information.
- **Timely:** addressing points at the right time.

3.7.2 Communication Media:

The communication media that will be used for the project are:

- Email
- Document (MS Word and/ PowerPoint)
- Phone call
- Meetings (using, meeting rooms, conference phones, Ms Teams)
- Chats (WhatsApp)

4.0 Commercialization

The Intelligent Document Formatter targets students, researchers, and academic professionals who need IEEE-compliant formatting. The solution addresses common challenges in academic publishing, making it ideal for universities and research institutions.

Focusing on academic settings, this tool offers unique features like real-time personalized formatting suggestions and automated IEEE compliance checks. It will be commercialized through a subscription model, with tiers offering varying levels of service.

To broaden its adoption, the tool will be offered as a standalone product or integrated into existing academic platforms, with potential partnerships to enhance its reach and integration.

5.0 Budgeting

The proposed Intelligent Document Formatter is a software-based solution, so the primary costs are related to cloud services and software tools. Below is the estimated budget:

Type	Cost
Cloud computing and storage	7,500 LKR
Subscription to development tools	10,000 LKR
Web hosting and domain	4,000 LKR
Miscellaneous (e.g., documentation, testing)	3,500 LKR
TOTAL	**25,000 LKR**

Table 1: Budget

This budget may vary depending on economic conditions and additional unforeseen expenses.

6.0 Summary

The Intelligent Document Formatter is designed to streamline the process of formatting academic documents to IEEE standards. By automating this process, the tool addresses the common challenges faced by researchers, academics, and students, such as the time-consuming and error-prone nature of manual formatting. The solution utilizes advanced NLP techniques and the FLAN-T5 model to handle complex formatting elements like equations, figures, tables, and citations. Additionally, it offers real-time personalized suggestions, adapting to the user's unique document structure and preferences.

The project includes a thorough analysis of technical and economic feasibility, ensuring that the solution is not only effective but also cost-efficient. The communication management plan ensures that all stakeholders are kept informed and engaged throughout the development process. With a clear budget and schedule, the project is set to deliver a tool that will significantly enhance productivity and accuracy in academic publishing. Once implemented, the Intelligent Document Formatter will reduce the manual effort required for document preparation, making it easier for users to meet the strict formatting requirements of IEEE and other academic standards.

7.0 References

1. **A. Shaw, R. Furuta, and J. Scofield, "Document formatting systems: Survey, concepts, and issues (Extended Abstract)" Tech. Rep. 80-10-02, Dept. of Computer Science, University of Washington, Seattle, WA, USA, Oct. 1980.**
2. **Shen, Y., Lee, K., & Gao, J. (2017). Unsupervised style transfer for text with sequence-to-sequence models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (pp. 788-798).**
3. **Shen, Y., Lee, K., & Gao, J. (2018). Style transfer for text generation with conditional variational autoencoders. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 542-551).**
4. **M. R. Ciosici and L. Derczynski, "Training a T5 Using Lab-sized Resources" in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2021, pp. 1234-1245.**
5. **Zhang, Y., Liu, Y., & Wu, J. (2019). Automatic text formatting with a sequence-to-sequence model. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (pp. 528-537).**
6. **Wang, S., Liu, X., & Zhang, H. (2016). Document structure understanding with recurrent neural networks. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (pp. 1414-1424).**
7. **Liu, X., Zhang, Y., & Chen, Z. (2020). Formatting text documents with a bidirectional LSTM encoder-decoder model. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (pp. 456-466).**