**Urdu Back Translation Text Reuse Detection**

Final Year Project Proposal

Session 2021-2025

A 4th Year Student

A project submitted in partial fulfillment of the

COMSATS University Degree

of

BSc. (Hons.)BS in Software Engineering (CUI)



Department of Computer Science

COMSATS University Islamabad, Lahore Campus

30 September 2024

**Project Registration**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Project ID (for office use) | | |  | | | | |
| Type of project | | | [✔] **Traditional** [ ] Industrial [ ] Continuing | | | | |
| Nature of project | | | [ ] **D**evelopment [ ✔] **R**&**D** | | | | |
| Area of specialization | | | Machine Learning, Natural Language Processing | | | | |
| **Project Group Members** | | | | | | | |
| Sr.# | Reg. # | Student Name | | CGPA | Email ID | Phone # | Signature |
| (i) | FA21-BSE-114 | Muhammad Umer Aamir | | 3.11 | FA21-BSE-114@cuilahore.edu.pk | +92 335 9119222 |  |
| (ii) | FA21-BSE-053 | Usama Tufail | | 3.50 | FA21-BSE-053@cuilahore.edu.pk | +92 333 0469452 |  |
| (iii) | FA21-BSE-120 | Malik Ashas | | 2.65 | FA21-BSE-120@cuilahore.edu.pk | +92 315 4816309 |  |
| **Declaration:** FYP group members have cleared all prerequisite courses For FYP-I as per their degree requirements.  For BS(Computer Science)  (CSC241 Object Oriented Programming, CSC291 Software Engineering Concepts, CSC371 Database Systems-I, HUM102 Report Writing Skills)  For BS(Software Engineering)  (CSC241 Object Oriented Programming, CSE291 Introduction to Software Engineering , CSC371 Database Systems-I , HUM102 Report Writing Skills) | | | | | | | |

# Plagiarism Free Certificate

This is to certify that, I am Muhammad Umer Aamir S/D/o Chaudhary Aamir Riaz, group leader of FYP under registration no CIIT/FA21-BSE-114/LHR at the Computer Science Department, COMSATS Institute of Information Technology, Lahore. I declare that my FYP proposal is checked by my supervisor and the similarity index is \_\_\_\_\_\_\_\_% that is less than 20%, an acceptable limit by HEC. The report is attached herewith as Appendix A.

Date: 30-09-2024 Name of Group Leader: Muhammad Umer Aamir Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_

Name of Supervisor: Dr Muhammad Sharjeel

Designation: Assistant Professor

Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Table of Contents

[Final Year Project Proposal 1](#_Toc178604339)

[Session 2021-2025 1](#_Toc178604340)

[Plagiarism Free Certificate 2](#_Toc178604341)

[1 Introduction 5](#_Toc178604342)

[2 Success Criterion 6](#_Toc178604343)

[3 Related work 6](#_Toc178604344)

[4 Project Rationale 7](#_Toc178604345)

[4.1 Aims and Objectives 7](#_Toc178604346)

[4.2 Scope of the Project 7](#_Toc178604347)

[4.3 Deliverables 8](#_Toc178604348)

[5 Proposed Methodology and Architecture 8](#_Toc178604349)

[5.1 Model Architecture 8](#_Toc178604350)

[5.2 Application Architecture 9](#_Toc178604351)

[6 Individual Tasks 9](#_Toc178604352)

[7 Gantt Chart 11](#_Toc178604353)

[8 Tools and Technologies 11](#_Toc178604354)

[8.1 Programming language: 11](#_Toc178604355)

[8.2 Data preprocessing: 11](#_Toc178604356)

[8.3 ML and DL libraries: 11](#_Toc178604357)

[8.4 Data visualization: 12](#_Toc178604358)

[8.5 UI design 12](#_Toc178604359)

[8.6 Web application development: 12](#_Toc178604360)

[8.7 Mobile application development: 12](#_Toc178604361)

[9 References 12](#_Toc178604362)

**Project Abstract**

In this modern digital world, it is so easier to reuse, copy, or rephrase text due to the huge amount of free data available on the internet. This can lead to concern with text being reused without proper attribution or credit. Urdu is a low-resource language used less commonly in digital contexts and NLP research, which causes additional challenges in Urdu text reuse and plagiarism detection systems. Also, the writer cannot get proper credit or attribution. Currently, there is no automated tool or service provided to detect text reuse in the Urdu language.

To address the challenge, we propose research and development project for the detection of text reuse and plagiarism in the Urdu language, using machine learning and deep learning architectures. In this FYP, the main objective is to develop a number of Urdu text reuse corpora using a semi-automated back translation approach. The secondary objective is the design and assessment of several Machine Learning and Deep Learning algorithms using the proposed corpora. The process involves preparing corpora, utilizing translation models, algorithmic evaluation, and integrating detection processes in the mobile and web applications. We believe the outcome of this FYP will serve as an important resource for the Urdu language NLP community, with practical applications in text reuse and plagiarism detection at both the academic and professional levels.

# Introduction

Text reuse detection is the task of determining two texts whether they possess same meaning by using different words or grammar. Text reuse and plagiarism detection become an important concept in field of research. This includes academics, content development, professional documentation etc. The tasks to detect the reused content in Urdu language is difficult, challenging and tough because it is a low-resource language. Also, there is unavailability of large, annotated corpora, advanced machine learning models and resources. On the internet, there is not a single Urdu language text reuse detection tool available. The conventional methods for text reuse and plagiarism detection are based on lexical (relating to the words or vocabulary) and syntactical (relating to the grammatical arrangement of words) similarities which fails to detect subtlety in meaning and contextual variation typical for rich uniquely crafted languages like Urdu. In the last couple of years, with the development of, Machine Learning (ML), Deep Learning (DL), and Large Language Models (LLMs), the Natural language Processing (NLP) has gained a boom. There are chances for developing models that might understand semantic similarities deeper. Yet, the main obstruction is the unavailability of specialized corpus of Urdu language and algorithms made according to its needs.

Text reuse detection around the world is leading to a range of challenges. For example, Mikołaj Piniewski [[1]](#_References), a researcher noted that the peer-review feedback received, is identical to those that were already available on the Web [[2]](#_References). Paraphrasing tools also disturb the integrity of academia, Gen-z students and researchers frequently use text sipping tools for AI paraphrasing on AI generative content and which produce alternate version with different words or sentence structure, leading professor and educators to think that the work is entirely original. CNET, a well-known tech news website, has attempted to publish articles generated by artificial intelligence (AI). This shows that the use of these tools has contributed to concerns about content originality and authenticity in online spaces. A research on paraphrasing tools in English as foreign Language (EFL) writing class in an Indonesian university students [[3]](#_References) shows the use of paraphrasing tool quilt bot [[4]](#_References) as:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **No.** | **Statements** | **SA** | **A** | **N** | **SD** | **D** |
| 1 | Generally, I use QuillBot to paraphrase my writing. | 85% | 10% | 5% | 0% | 0% |
| 2 | There are many features in QuillBot that are very useful for improving my writing skill. | 60% | 35% | 5% | 0% | 0% |
| 3 | Using QuillBot helps me to improve my grammar skills. | 30% | 50% | 20% | 0% | 0% |
| 4 | Using QuillBot helps me to increase my vocabulary. | 30% | 60% | 10% | 0% | 0% |
| 5 | Using QuillBot helps me use proper language for paraphrasing my writing. | 40% | 40% | 20% | 0% | 0% |
| 6 | I can write confidently and get better paraphrasing results using QuillBot. | 50% | 40% | 10% | 0% | 0% |

Figure 1 : Strongly Agree (SA), Agree (A), are Neutral (N), Strongly Disagree (SD), or Disagree (D)

In this FYP, we aim to develop nine text reuse detection corpora for the Urdu language using semi-automated approach. The approach involves translating an Urdu text into several different languages and then back to the Urdu language. This process changes the text significantly, making it a paraphrase version of the original text. Five corpora will be created by translating Urdu text into most popular languages while four will be created using different language families. Back-translation approach involves translation from a source language A to a target language B and then its translation back to the source language A. It can be of help in generating text reuse cases that capture different contextual and semantic meanings which are critical for training models. Secondly, we will experiment with ML, DL, LLM-based algorithms to assess their strength across multiple Urdu text reuse corpora that shall be created during the project. Mobile and web-based applications will be developed which will be able to detect text reuse and plagiarism amongst other features like OCR and text-to-speech in Urdu. Rest API’s will be useful tool in academic and professional system where access to plagiarism detection tools for languages like Urdu is rather low. This project also uses modern architecture, including transformers and other deep neural networks, to develop a robust system for the Urdu NLP community that could be useful in semantics-based text similarity detection and finding out highly similar or copied texts.

This project also contributes to the broader research effort on low-resource languages by publishing the corpora and models generated within the research community.

# Success Criterion

The ultimate success of project is to make mobile and web based text reuse detection application that can easily integrate in any app using REST API. Our project will make contribution to the domain of NLP in the context of the Urdu language. The success of this final year project is based on set of realistic and verifiable criteria as defined following:

1. **Creation of Text Reuse Corpora for the Urdu language Using Back Translation:**

A significant work is the construction of 9 large scale Urdu text reuse corpora. We will construct the corpora using back translation technique. The 5 corpora will be made by utilizing a 1-3-5-7-9 languages translation method. This process involves translating text from one language to other and then back to its original (e.g., Urdu to English, English to Arabic, and then back in Urdu). The 4 corpora will be made by back translating text using 4 most popular language families. These corpora will have source and reused text classified into 3 classes such as Partially Derived (PD), Wholly Derived (WD), and Non-Derived (ND).

1. **Successful Design of ML and DL Architectures:**

The second success of this final year project will be the design, implementation, and testing of models based on ML and DL architectures. These models must effectively detect text reuse for the Urdu language. The models will be evaluated based on performance metrics like precision, recall, and F1 score.

1. **Development of Mobile and Web Applications:**

The success of this final year project will also be based on the development of a user-friendly mobile and web application that will provide text reuse detection in Urdu language. Both web and mobile application will offer modern functionalities like cloud-based history access, provides REST API for integration in other app and a modern interface with features like OCR that ensure ease of use for academic, professional, and personal use cases.

# Related work

In the context of text and plagiarism detection, many researchers have explored the use of Deep Neural Networks (DNNs) for various languages. However, for the Urdu language there is limited work exists for text reuse detection as in comparison to widely used languages like English.

In related studies, neural network-based techniques such as Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN) have been utilized to handle linguistic variations [[5]](#_References) . While these models perform well, the unique challenges presented by Urdu’s rich morphology and syntactic structure require specific attention, making this research crucial.

The use of transfer learning for text reuse detection has gained popularity due to its efficiency in handling smaller datasets. Pre-trained sentence transformer models like BERT [[6]](#_References) have been adapted for paraphrase detection, providing promising results in popular languages.

This study extends the application of transfer learning and in low-resource languages like Urdu, offering an approach for paraphrase detection while overcoming challenges posed by the limited availability of annotated data.

Concludingly, above mentioned researchers work on Urdu language paraphrasing models, but these are not well efficient. Also, there is no application developed to detect the paraphrasing score in Realtime.

# Project Rationale

This final year project focuses on text reuse detection within the Urdu language. Currently, there is a significant lack of efficient and effective tools for detecting text reuse in Urdu language, both in academic and professional contexts. Urdu is the 10th most widely spoken language globally [[7]](#_References) , yet there is minimal research and development in the domain of NLP for Urdu, especially regarding text reuse detection. Text reuse is considerably more prevalent in the Urdu language compared to others. Many individuals often use content from other sources, paraphrase it, and present it as their own work and due to lack of tools in become undetectable. There are many complexities in detection of paraphrasing in Urdu language such as its distinctive structure, vocabulary and syntactic styles.

By developing this system, we aim to address the issue of text reuse and promote originality in content creation within the Urdu-speaking community. This will help control the use of plagiarized content, maintain academic integrity, prevent intellectual property theft, and ensure the authenticity of Urdu content.

## Aims and Objectives

* Utilize ML and DL methods (e.g., sentence transformers, CNN, LSTM, etc.) to predict text reuse in Urdu documents.
* To create 9 different Urdu text reuse corpora using back translation approach.
* To conduct various experiments using ML and DL methods on variety of corpora and evaluate the comparative results.
* To contribute to the advancement of research on low resource Urdu language by developing text reuse detection tool.

## Scope of the Project

* Develop nine Urdu corpora using back translation approach for text reuse detection research.
* Train and evaluate different machine learning algorithms on the proposed corpora for Urdu text reuse detection.
* Employ modern deep learning techniques for effective content analysis and classification.
* Conduct a series of experiments to identify the most accurate and efficient classification model for Urdu text reuse detection.
* Integrate the trained model into both mobile and web applications for seamless use.

## Deliverables

The primary deliverable will consist of a gold standard Urdu text reuse corpus created using back translation approach, along with the most efficient classification model(s) integrated into a mobile and web-responsive application.

* Project Documentation
* Trained ML Model or Deep Learning architecture
* A Comprehensive cyclic translation dataset
* Mobile and web app that checks for paraphrasing
* Final Project Presentation to communicate project scope

# Proposed Methodology and Architecture

This project will follow the agile model. We will develop both web and mobile applications with rich features such as providing REST API’s to easily integrate the paraphrase and plagiarism detection in the other applications (e.g., blog apps). The further details of our application and model architecture is as following:

## Model Architecture

Initially, we will collect the data form different sources and compile the corpus. Our corpus will consist of 2 input attributes which are ‘text1’ and ‘text2’ and 1 target output with 3 classes which are PD (Partially Derived), ND (Non Derived), and WD (Wholly Derived). Then we will apply pre-processing techniques to remove any noise. After pre-processing of data, we will extract features from the text using sentence transformers. After feature extraction, we will apply ML and DL approaches on that data and train models. After the model training, we will evaluate that model and calculate accuracy, recall, precision, and F1 score. We will re-train the model until we achieve the desired results. Below is the illustration of the model pipeline:

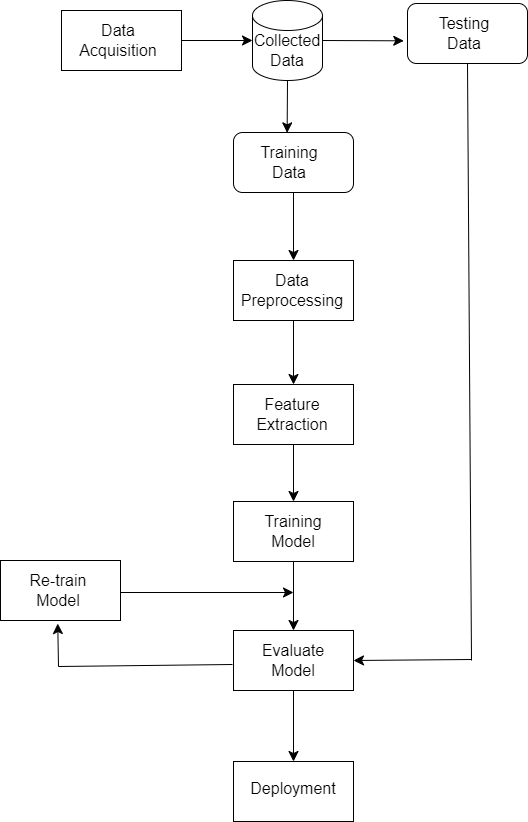


Figure 2 : Model pipeline diagram

## Application Architecture

We will utilize the client server architecture in our project. In the backend, we have 3 servers. The database will be hosted on the database server. The trained and evaluated model will be hosted on the ML server. The main server will be the core of our system which will contain the backend files of our system and will be integrated with the rest of the servers. The database and ML servers will connect to the main server. We will make web application and mobile application which will integrate to the backend server. Below is the system diagram of our architecture.

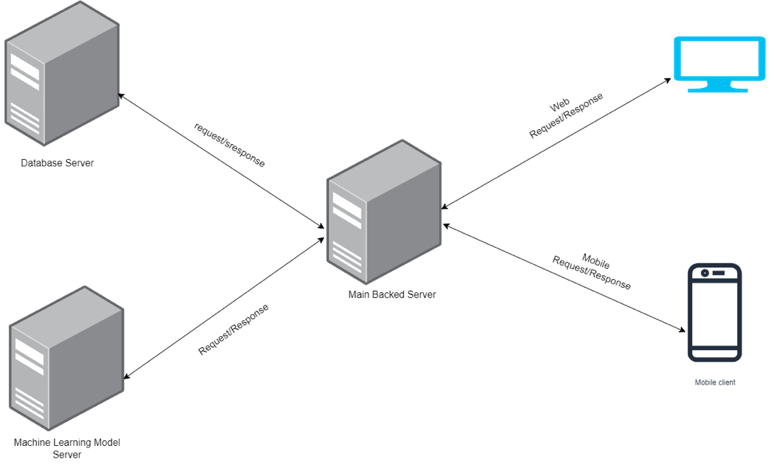


Figure 3 : Architecture Diagram

# Individual Tasks

|  |  |  |
| --- | --- | --- |
| **Team Member** | **Activity** | **Tentative Date** |
| Muhammad Umer Aamir, Ashas Abbas, Usama Tufail | Project research | 12-Sep-2024 |
| Muhammad Umer Aamir, Usama Tufail | Project Documentation | 10-Nov-2024 |
| Ashas Abbass | Data Gathering and preprocessing | 25-Nov-2024 |
| Usama Tufail | Feature extraction | 20-Dec-2024 |
| Muhammad Umer Aamir | Feature engineering | 31-Dec-2024 |
| Muhammad Umer Aamir, Usama Tufail | Model selection and development | 10-Apr-2024 |
| Malik Ashas | Model evaluation | 20-Apr-2024 |
| Muhammad Umer Aamir | Backend Development | 25-May-2024 |
| Ashas Abbas | User interface Design | 5-June-2024 |
| Muhammad Umer Aamir, Usama Tufail, Muhammad Ashas | Integration | 15-June-2024 |
| Muhammad Umer Aamir | System Deployment | 20-June-2024 |
| Usama Tufail | Testing mobile and web application and model | 30-June-2024 |
| Muhammad Umer Aamir, Malik Ashas, Usama Tufail | Presentation | 31-June-2024 |

# Gantt Chart

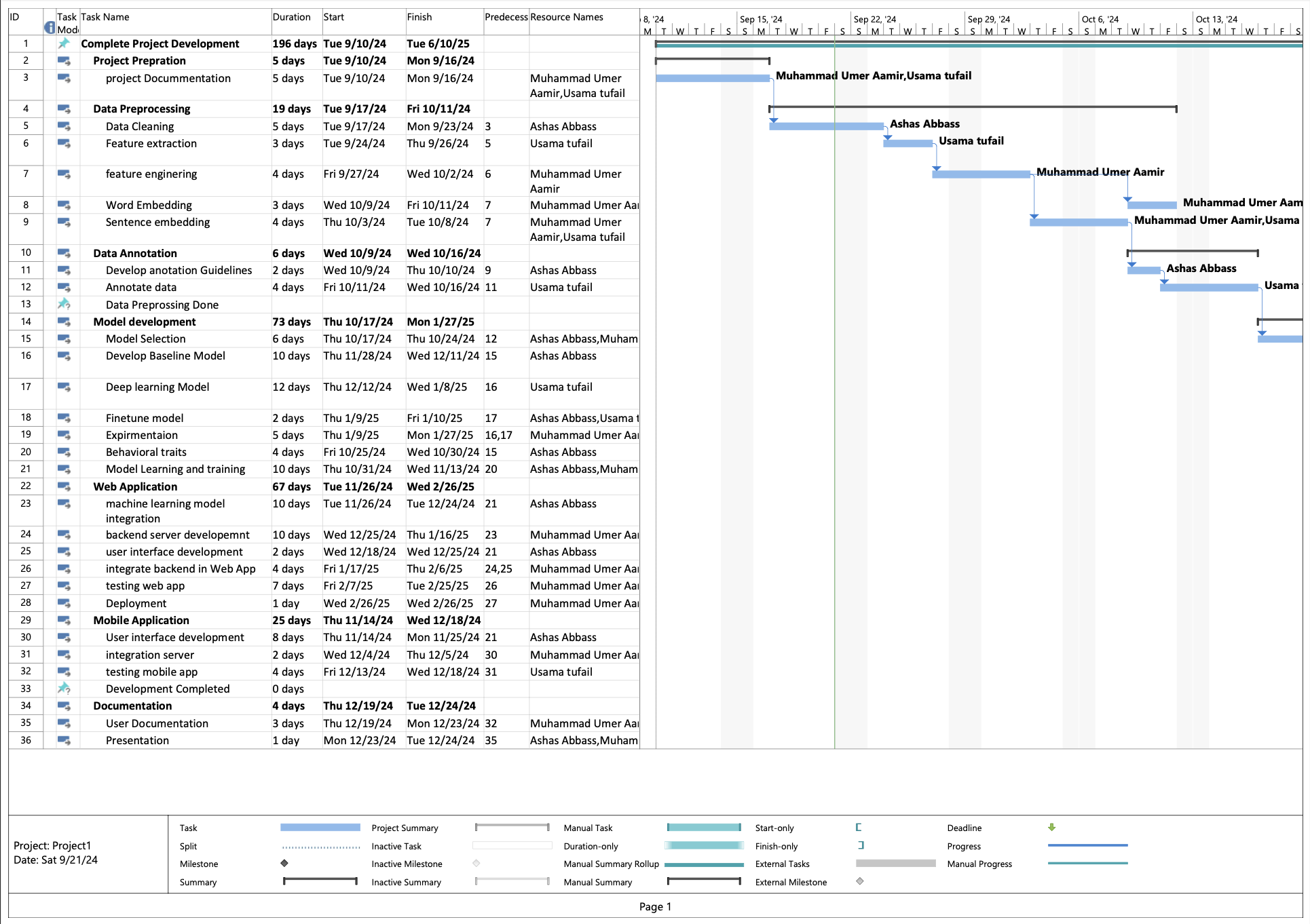


Figure 4 : Gantt Chart

# Tools and Technologies

## Programming language:

We will use Python as the primary programming language in project development. Python is popular and widely used language for the development of ML and DL models.

## Data preprocessing:

* **Pandas:** For efficient data manipulation and preprocessing tasks, such as handling missing data, standardizing formats, and extracting relevant features from the dataset(s).
* **NumPy:** For numerical operations and array manipulations, providing a foundation for various mathematical operations involved in data preprocessing.

## ML and DL libraries:

We will use the Python libraries such as sci-kit-learn, TensorFlow, and PyTorch etc. that are crucial for developing and deploying ML and DL models.

* **Scikit-Learn:** For implementing the classification algorithms. It will be used for implementing and fine-tuning the classification models.
* **TensorFlow** and **PyTorch:** For implementing complex neural network architectures, providing flexibility and scalability in model development.

## Data visualization:

* **Matplotlib** and **Seaborn:** For creating visualizations to aid in data exploration, analysis, and presentation of results

## UI design

* **Figma:** For prototype design.

## Web application development:

* **NextJS:** for web server development, as it provides multiple tools for user data handling, authentication system and manage database.
* **React JS:** a fast and modern JavaScript library to create modern milimistic user interface using component-based approach.
* **Node JS:** aJavaScript runtime environment that allows server side execution by providing features like asynchronously and micro services web server

## Mobile application development:

* **React Native:** For application development, as it provides tools and a framework for cross-platform applications.

# References

[1] <https://scholar.google.com/citations?user=FctmX1AAAAAJ&hl=en>

[2] <https://www.nature.com/articles/d41586-024-01312-0>

[3] <https://www.researchgate.net/publication/370672507_Utilizing_Artificial_Intelligence-based_Paraphrasing_Tool_in_EFL_Writing_Class_A_Focus_on_Indonesian_University_Students'_Perceptions>

[4] <https://quillbot.com/blog/frequently-asked-questions/what-does-quillbot-do/>

[5][https://www.researchgate.net/publication/371144244\_Urdu\_paraphrase\_detection\_A\_nove l\_DNN-based\_implementation\_using\_a\_semi-automatically\_generated\_corpus](https://www.researchgate.net/publication/371144244_Urdu_paraphrase_detection_A_nove%20l_DNN-based_implementation_using_a_semi-automatically_generated_corpus)

[6] <https://huggingface.co/docs/transformers/en/model_doc/bert>

[7] <https://en.m.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers>

[8]<https://www.pactranz.com/backtranslation/#:~:text=Back%20translation%20is%20a%203,meaningful%20differences%20between%20the%20two>

