Plagiarism Detection System for Urdu Language

Project Proposal

Supervisor
Mr. Aamir Hussain

Submitted by

Syed Ahmed Ali
{2579-2021 / IT-21-138}

Syed Shaheer-ul-Haque
{2197-2021 / IT-21-281}

Department of Computer Science,
Hamdard University, Karachi.

[24/Jun/2024]

1. Introduction

The project aims to develop a plagiarism detection system specifically for the Urdu language. This system will be accessible through mobile applications, providing a comprehensive tool for educators, students, and professionals to ensure the originality of their Urdu texts. The project addresses the gap in existing plagiarism detection tools, which predominantly focus on widely spoken languages and often overlook regional languages like Urdu.

2. Objective

To create a user-friendly mobile application that efficiently detects plagiarism in Urdu text documents.

3. Problem Description

What: Current plagiarism detection tools primarily support widely spoken languages like English, leaving Urdu largely unaddressed. This project seeks to create a dedicated plagiarism detection system for Urdu texts.

Why: As the volume of academic and professional content in Urdu increases, ensuring its originality is essential for maintaining academic integrity and professional credibility. Existing tools are not equipped to handle the unique script and linguistic structure of Urdu, necessitating a tailored solution.

4. Methodology

To address the problem, the project will utilize natural language processing (NLP) techniques specific to Urdu, following the VNV (Verification and Validation) model to ensure the accuracy and reliability of the system. The approach includes:

- **Text Preprocessing:** Tokenizing, stemming, and lemmatizing Urdu text to prepare it for analysis.
- **Similarity Detection:** Implementing advanced algorithms such as cosine similarity and sequence matching to identify potential plagiarism.
- **Database Comparison:** Comparing the processed text against a comprehensive database of Urdu documents to detect copied content.
- **Frameworks and Libraries:** Utilizing libraries like NLTK for NLP tasks and TensorFlow for machine learning models to enhance detection accuracy.

5. Project Scope

The project will focus on developing core plagiarism detection functionality for Urdu text, excluding support for other languages and advanced features like cross-language detection. Assumptions include stable internet access for online databases and users' basic understanding of Urdu.

6. Feasibility Study

Risks Involved:

- **Data Quality:** Ensuring an accurate and comprehensive database of Urdu documents.
- **Algorithm Accuracy:** Fine-tuning algorithms to handle Urdu's nuances.
- **Resource Requirements:** Securing computational resources for NLP model training and execution.

Resource Requirements:

- **Computing Resources:** High-performance servers for database management and model training.
- **Development Tools:** IDEs, version control systems, collaboration tools.
- **Data:** Access to a large corpus of Urdu documents for training and testing.

7. Solution Application Areas

The project is valuable for academic institutions, publishers, and content creators producing Urdu text. It can be used in educational settings to check assignments, in publishing to ensure manuscript originality, and in professional environments to verify the uniqueness of reports and articles.

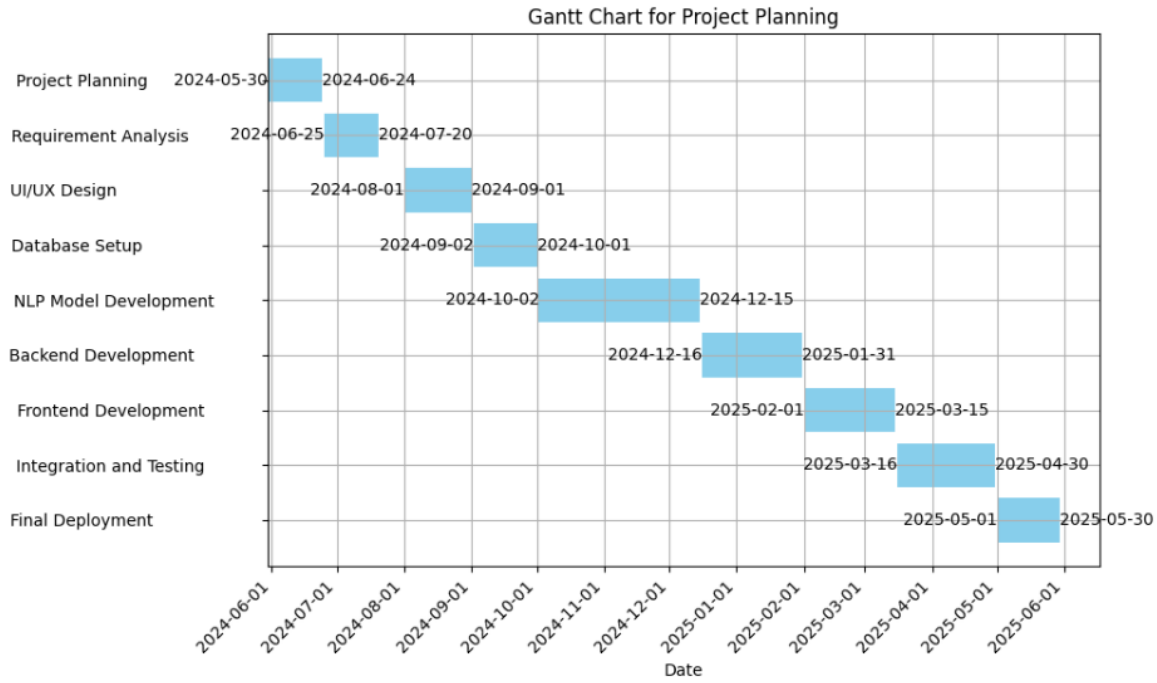
8. Tools/Technology

- **Programming Languages:** Python.
- **Libraries:** NLTK, TensorFlow, Scikit-learn, UrduHack
- **Database:** SQLLITE
- **Platforms:** FLUTTER

9. Responsibilities of the Team Members

Task/Responsibility	Syed Ahmed Ali (Team Lead)	Syed Shaheer (Team Member)	Mr. Aamir (Supervisor)	Ms. Muntaha (Co-Supervisor)
Project Planning	R	R	C	C
Requirement Analysis	R	A	C	C
UI/UX Design	A	R	C	I
Database Setup	A	R	I	C
NLP Model Development	A	R	I	C
Backend Development	R	A	C	C
Frontend Development	R	A	C	C
Integration and Testing	R	R	C	C
Final Deployment	R	A	I	C

10. Planning



11. References

- "**Plagiarism Detection in Urdu Documents using Sentence Structure Analysis**" by S. M. Akram Shah, Muhammad Ashfaq, Saira Banu, and Abdul Wahid.
- **"Urdu Plagiarism Detection using Statistical Features"** by M. Naveed Iqbal, M. Arif, S. M. Akram Shah, and Muhammad Usman.
- **"Plagiarism Detection in Urdu Language Documents Using Shallow Semantic Parsing"** by M. Yasir Khan and Huma Sarwar.