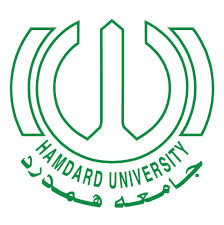
Hamdard University

Department of Computing

Final Year Project



**PLAGIARISM DETECTION SYSTEM FOR URDU LANGUAGE**

**(FYP-027/FL24)**

**Software Requirements Specifications**

Submitted by

SYED AHMED ALI (2579-2021)

SYED SHAHEER Ul HAQUE (2197-2021)

Supervisor(s)

Mr. AAMIR HUSSAIN

**Fall 2025**

**Document Sign off Sheet**

### **Document Information**

|  |  |
| --- | --- |
| **Project Title** | PLAGIARISM DETECTION SYSTEM FOR URDU LANGUAGE |
| **Project Code** | FYP-027/FL24 |
| **Document Name** | Software Requirements Specifications |
| **Document Version** | <1.0> |
| **Document Identifier** | FYP-027/FL24 |
| **Document Status** | Final |
| **Author(s)** | SYED AHMED ALI  SYED SHAHEER Ul HAQUE |
| **Approver(s)** | Mr AAMIR HUSSAIN |
| **Issue Date** | 22/MAY/2024 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Role** | **Signature** | **Date** |
| SYED AHMED ALI | Team Lead | C:\Users\DELL\Downloads\FINAL FYP\Plagiarism mom\Syed Ahmed.png | 6-July-25 |
| SYED SHAHEER UL HAQUE | Team Member 2 | C:\Users\DELL\Downloads\FINAL FYP\Plagiarism mom\Syed Shaheer.png | 6-July-25 |
| MR. AAMIR HUSSAIN | Supervisor | **C:\Users\DELL\Downloads\FINAL FYP\Plagiarism mom\Sir Aamir.png** | 6-July-25 |
|  | Co-Supervisor |  |  |

**Revision History**

|  |  |  |  |
| --- | --- | --- | --- |
| **Date** | **Version** | **Description** | **Author** |
| 11-JAN-25 | 1.0 | Final Draft of the SRS Document for Submission | Syed Shaheer/ Ahmed Ali |
|  |  |  |  |
|  |  |  |  |

**Definition of Terms, Acronyms, and Abbreviations**

|  |  |
| --- | --- |
| **Term** | **Description** |
| Plagiarism | The act of using someone else's work, ideas, or text without proper attribution, resulting in intellectual dishonesty. |
| Urdu NLP | Natural Language Processing techniques specifically designed to process and analyze the linguistic structure of the Urdu language. |
| Tokenization | The process of breaking down text into smaller units such as words or sentences for analysis. |
| Stemming | A text preprocessing technique used to reduce words to their root form (e.g., "کھانے" to "کھا"). |
| Lemmatization | A linguistic process to reduce words to their base or dictionary form, considering the context (e.g., "کھیلا" to "کھیل"). |
| Cosine Similarity | A mathematical algorithm to measure the similarity between two text documents based on their vector representation. |
| Database | A collection of Urdu documents stored digitally, used for comparison against input text to detect plagiarism. |
| UrduHack | A Python library used for processing and analyzing Urdu text for NLP tasks. |
| TensorFlow | An open-source library used for machine learning, aiding in creating models for detecting plagiarism. |
| Plagiarism Report | A document generated by the system that highlights copied text, provides similarity scores, and identifies sources of plagiarism. |
| ML Model | A machine learning model designed to identify patterns in text to detect plagiarism efficiently. |
| Preprocessing | A set of steps (e.g., tokenization, stemming) to clean and prepare Urdu text for further analysis by the system. |
| Similarity Score | A numerical value indicating how much an input document resembles other documents in the database. |
| Error Message | A notification provided by the system when an invalid input or processing error occurs. |
| Urdu Text | Input provided by the user in the Urdu language to check for plagiarism. |

# **Table of Contents**

**1**. [**Introduction** 6](#_Toc187589800)

[1.1 Purpose of Documents 6](#_Toc187589801)

[1.2 Intended Audience 6](#_Toc187589802)

[1.3 Abbreviations 6](#_Toc187589803)

[2. Overall System Description 7](#_Toc187589804)

[2.1 Project Background 7](#_Toc187589805)

[2.2 Problem Statement 7](#_Toc187589806)

[2.3 Project Scope 7](#_Toc187589807)

[2.4 Project Objectives 7](#_Toc187589808)

[2.5 Stakeholders & Affected Groups 7](#_Toc187589809)

[2.6 Operating Environment 8](#_Toc187589810)

[2.7 System Constraints 8](#_Toc187589811)

[2.8 Assumptions & Dependencies 8](#_Toc187589812)

[3. External Interface Requirements 9](#_Toc187589813)

[3.1 Hardware Interfaces 9](#_Toc187589814)

[3.2 Software Interfaces 9](#_Toc187589815)

[3.3 Communications Interfaces 9](#_Toc187589816)

[4. System Functions / Functional Requirements 10](#_Toc187589817)

[4.1 System Functions 10](#_Toc187589818)

[4.2 Use Cases 10](#_Toc187589819)

[4.2.1 List of Actors 10](#_Toc187589820)

[4.2.2 List of Use Cases 10](#_Toc187589821)

[4.2.3 Use Case Diagram 11](#_Toc187589822)

[4.2.4 Description of Use Cases 11](#_Toc187589823)

[5. Non - Functional Requirements 14](#_Toc187589824)

[5.1 Performance Requirements 14](#_Toc187589825)

[5.2 Safety Requirements 14](#_Toc187589826)

[5.3 Security Requirements 14](#_Toc187589827)

[5.4 Reliability Requirements 14](#_Toc187589828)

[5.5 Usability Requirements 14](#_Toc187589829)

[5.6 Supportability Requirements 14](#_Toc187589830)

[5.7 User Documentation 14](#_Toc187589831)

[6. References 15](#_Toc187589832)

# **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.

## **Purpose of Documents**

This document defines the software requirements for the **Plagiarism Detection System for Urdu Language**, a project aimed at providing a user-friendly tool to detect plagiarism in Urdu text using Natural Language Processing (NLP) techniques. It ensures clarity on functional and non-functional requirements for development and evaluation.

## **Intended Audience**

This document is intended for the following stakeholders:

* Development Team
* Supervisors (Mr. Aamir Hussain, Dr. Shahid Muneer)
* End Users (educators, students, and professionals working with Urdu content)
* Academic and administrative decision-makers

## **Abbreviations**

|  |  |
| --- | --- |
| **Abbreviation** | **Description** |
| NLP | Natural Language Processing |
| UI/UX | User Interface/User Experience |
| VNV | Verification and Validation |
| NLTK | Natural Language Toolkit |

# **Overall System Description**

## **Project Background**

Current plagiarism detection tools like Aplag, Turnitin, EVE2, and Copyscape, are all primarily created for widely spoken languages like English and Arabic, leaving regional languages like Urdu overlooked. This project addresses the lack of a mobile application which is Urdu-specific and can efficiently detect plagiarism with better accuracy.

## **Problem Statement**

Accessing information online is easy, but it raises concerns about plagiarism. Plagiarism detection tools help maintain the quality of academic work. Institutions like HEC focus on ensuring originality, and to prevent theft, they require proof of ownership. Comparing original and copied documents can help confirm authorship and protect intellectual property.

## **Project Scope**

The system will:

* Detect plagiarism in Urdu text documents.
* Focus exclusively on Urdu language.
* Utilize NLP and machine learning techniques for high accuracy.
* Provide a mobile-friendly interface for user accessibility.

**Excluded Features:**

* Support for cross-language plagiarism detection.
* Integration with other languages or datasets.
* Offline functionality without stable internet access for database queries.

## **Project Objectives**

* Develop a robust NLP-based system for Urdu plagiarism detection.
* Ensure high accuracy and usability through iterative testing.
* Create a mobile application with intuitive UI/UX.

## **Stakeholders & Affected Groups**

* Students and educators in academic institutions
* Urdu-language publishers and writers
* Professionals generating Urdu reports/articles

## **Operating Environment**

* Mobile devices support NLP-based applications
* Backend database hosted on high-performance servers

## **System Constraints**

* Stable internet connection required for database operations
* Processing limitations based on server resources

## **Assumptions & Dependencies**

* Availability of a comprehensive Urdu document dataset for training and testing
* Users have a basic understanding of the Urdu language

# **External Interface Requirements**

## **Hardware Interfaces**

* Mobile devices with a minimum of 4GB RAM and Android/iOS support
* Backend servers for NLP model hosting and database management

## **Software Interfaces**

* Python-based Libraries also system will connect to SQLLITE database backend utilizing NLTK and TensorFlow libraries
* SQLLITE database for storing and comparing documents

## **Communications Interfaces**

* Mobile application connected to the backend via RESTful APIs
* Data transfer through encrypted channels (HTTPS)

# **System Functions / Functional Requirements**

## **System Functions**

|  |  |  |  |
| --- | --- | --- | --- |
| Ref # | Function | Category | Priority |
| F1 | Preprocess Urdu text for tokenization | Core | High |
| F2 | Detect text similarity using NLP models | Core | High |
| F3 | Compare text against Urdu document database | Core | High |
| F4 | Provide detailed plagiarism reports | Evident | Medium |
| F5 | User account management | Evident | Medium |

## **Use Cases**

**Actors**: User, Admin

**System**: Check Plagiarism, Generate Report

### **List of Actors**

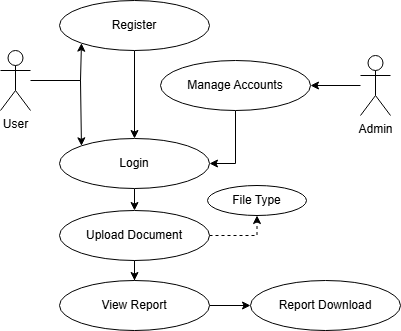
**User:** This person can register/login, upload document and view/download report.

**Admin:** This person can manage accounts.

### **List of Use Cases**

1. User can register or login into accounts to access the app, submit the desired document and if the document satisfies the system criteria, then it’ll generate a report that the user can view or download.

### **Use Case Diagram**



### **Description of Use Cases**

|  |  |
| --- | --- |
| Section: Main |  |
| **Name:** | Submit Document for Plagiarism Check |
| **Actors:** | User, Admin |
| **Purpose:** | To check for plagiarism in an Urdu text document and provide a report. |
| **Description:** | The user registers/login and uploads a PDF document containing only Urdu text for plagiarism detection. The system processes the text using NLP techniques, compares it against the database, and calculates the plagiarism percentage. Upon completion, the system generates a detailed plagiarism report. |
| **Cross References:** | Use Cases: The system must verify file type before proceeding. |
| Pre-Conditions | The system assumes that the user has inputted an Urdu text document to check for plagiarism. |
| Successful post-conditions | The system successfully identifies and reports any plagiarized content in the Urdu text. |
| Failure post-conditions | The system fails to analyze the text or detect plagiarism due to input errors or system failure. |

|  |  |  |  |
| --- | --- | --- | --- |
| Typical Course of Events | | | |
| Actor Action | | System Response | |
| 1 | The user registers an account. |  |  |
| 3 | The user login into an account. |  |  |
| 5 | The user uploads or inputs the Urdu text document into the system. |  |  |
|  |  | 2 | Generate an account. |
|  |  | 4 | Checks Credentials, successful login. |
|  |  | 6 | The system accepts the document and verifies its format. |
|  |  | 7 | The system begins preprocessing the text (tokenization, stemming, lemmatization). |
|  |  | 8 | The system compares the input text with the database of Urdu documents. |
|  |  | 9 | The system calculates the similarity score using algorithms |
| 10 | The user reviews highlighted plagiarized sections. |  |  |
|  |  | 11 | The system highlights copied text and generates a plagiarism report. |
| 12 | The user downloads the plagiarism report. |  |  |
|  |  | 13 | The system provides the plagiarism report in document to the user. |

|  |  |
| --- | --- |
| Alternative Course |  |
| Step 2: | Password criteria doesn’t match. Ask the user to enter another password to register an account |
| Step 4: | Invalid Credentials. Ask the user to try again. |
| Step 6: | Invalid file format or unsupported language entered. Indicate an error message. |
| Step 7: | Cross-Language error. Won’t proceed to check the plagiarism with and error and ask the user to enter another document. |
| Step 8: | System could not access the plagiarism database. Cancel the plagiarism detection process and notify the user. |

# **Non - Functional Requirements**

## **Performance Requirements**

* The system must process a document (max 1000 words) within 30-40 seconds.

## **Safety Requirements**

* Sensitive user data must be encrypted during storage and transmission.

## **Security Requirements**

* System access should be protected through secure login.

## **Reliability Requirements**

* 99.9% uptime for backend services.

## **Usability Requirements**

* The mobile application must have an intuitive interface with English-language support for instructions.

## **Supportability Requirements**

* Regular updates for the NLP model and database to ensure detection accuracy.

## **User Documentation**

* User guide for mobile application operation.

# **References**

* "Plagiarism Detection in Urdu Documents using Sentence Structure Analysis" by S. M. Akram Shah et al.
* "Urdu Plagiarism Detection using Statistical Features" by M. Naveed Iqbal et al.
* "Plagiarism Detection in Urdu Language Documents Using Shallow Semantic Parsing" by M. Yasir Khan et al.