

RIPHAH INTERNATIONAL UNIVERSITY, ISLAMABAD



BSAI-3
(Fall 2025)

Faculty of Computing
Subject: AI

Members

Name: Muhammad Afaq Shoaib, Saghir Ali
Sap ID: 66285, 67005

Contents

Project Report	2
DATASET	2
OBJECTIVES	2
METHODOLOGY	2
EXPECTED RESULTS	3

AI Project Report: Riphah University Intelligent Chatbot

Project Report

PROBLEM STATEMENT

University administrative departments frequently face a high volume of repetitive inquiries regarding fee structures, admission deadlines, and academic programs. Manual responses are time-consuming and prone to delays. This project addresses the need for an automated, scalable solution to provide instant, accurate information to students and visitors, thereby reducing the workload on administrative staff and improving user experience.

DATASET

The model is trained on a custom-curated dataset aggregated from two JSON sources ('DatasetAIP.json' and 'DatasetAIP2.json').

- **Structure:** The dataset consists of 951 unique records after deduplication.
- **Features:** Each record includes unique identifiers, natural language questions, corresponding answers, categories (e.g., fee_structure, skills_development), intents, and extracted entities.
- **Domain:** The content is specific to Riphah International University, covering topics such as admissions, campus facilities, and academic regulations.

OBJECTIVES

- To develop a robust Question-Answering (QA) system capable of understanding natural language queries.
- To implement a hybrid retrieval mechanism that combines keyword matching, semantic understanding, and machine learning classification.
- To handle class imbalances in the training data to ensure accurate intent detection across varied query types.
- To deploy a user-friendly web interface for real-time interaction.

METHODOLOGY

The project employs a multi-stage pipeline to process queries and retrieve answers:

1. Preprocessing: Text inputs are normalized using a custom pipeline that includes spelling correction ('pyspellchecker'), abbreviation expansion (e.g., "cs" to "computer science"), and lemmatization using 'spaCy'.

2. Feature Extraction:

- **TF-IDF Vectorization:** Utilizes both word-level (1-2 n-grams) and character-level (3-5 n-grams) vectors to capture lexical patterns.

- **Semantic Embeddings:** Uses 'SentenceTransformer' (model: multi-qa-mpnet-base-dot-v1) to generate deep semantic vector representations of questions.

3. Model Training:

- **Class Imbalance Handling:** Applied SMOTE (Synthetic Minority Over-sampling Technique) to balance dataset categories.
- **Classification:** A VotingClassifier aggregates predictions from Multinomial Naive Bayes, Complement Naive Bayes, Logistic Regression, and Calibrated Linear SVC to predict user intent.

4. Hybrid Inference Logic: The system selects the best answer based on a priority hierarchy:

1. **Semantic Search:** High cosine similarity scores (> 0.66) trigger semantic answers.
2. **Fuzzy Matching:** 'RapidFuzz' handles typos and close matches (> 72 threshold).
3. **ML Classification:** If the classifier confidence is high (> 0.60), the answer is retrieved from the predicted category.

EXPECTED RESULTS

- **Performance:** The Logistic Regression component achieved the highest individual accuracy ($\approx 75.5\%$), with the ensemble Voting Classifier providing a balanced accuracy of $\approx 67\%$ on test data.
- **Robustness:** The hybrid approach effectively handles exact matches, typos, and semantically similar but phrased-differently queries.
- **Deployment:** A functional, interactive web interface built with Gradio, featuring a dark-themed UI with a news ticker and chat window, successfully providing real-time answers to university-related queries.