RAJALAKSHMI
ENGINEERING COLLEGE
An AUTONOMOUS Institution
Affiliated to ANNA UNIVERSITY, Chennai

25 years OF ACADEMIC EXCELLENCE

Department of Computer Science and Engineering

FAKE JOB POSTING PREDICTOR

Mrs. M. Divya M.E.,
Asst.Prof, Department of CSE
Rajalakshmi Engineering College

220701216
Rakhul Prakash S B

Problem Statement and Motivation

Problem

Online job portals face a growing threat from fraudulent job postings, which exploit job seekers and undermine trust, while traditional detection methods fail to scale against such sophisticated scam.

Motivation

There is an urgent need for an automated, intelligent, and scalable solution using Machine Learning and NLP to detect fake job postings and ensure a safer, more trustworthy online recruitment environment.

Existing System

1. Manual Moderation & Rule-Based Filters
2. Basic NLP & Simple Classifiers

The Need for Advancement: These limitations highlight the necessity for more sophisticated, data-driven approaches like advanced ensemble ML models and deeper feature engineering.

Objectives

Develop a Robust Detector 

Leverage Rich Data 

Optimize for Performance 

Ensure Practicality 

Demonstrate & Validate 

Abstract

Utilizing a real-world dataset, the methodology encompasses comprehensive data preprocessing, including text cleaning and TF-IDF vectorization for textual features, combined with engineered features like suspicious keyword counts, link presence, and title length.

Categorical data is handled via one-hot encoding, and class imbalance is addressed using RandomOverSampler.

A **Random Forest classifier**, chosen for its robustness with high-dimensional and sparse data, forms the core of the predictive model.

Proposed System

Preprocessing Pipeline:

- Automated text cleaning (SpaCy).
- TF-IDF vectorization for textual features.
- One-Hot encoding for categorical features

Strategic Feature Engineering:

- Combines key text fields (title, description, etc.).
- Extracts custom scam indicators: title_length, num_links, suspicious_words_count.

Imbalance Correction:

- RandomOverSampler to boost minority (fake) class representation in training

Proposed System

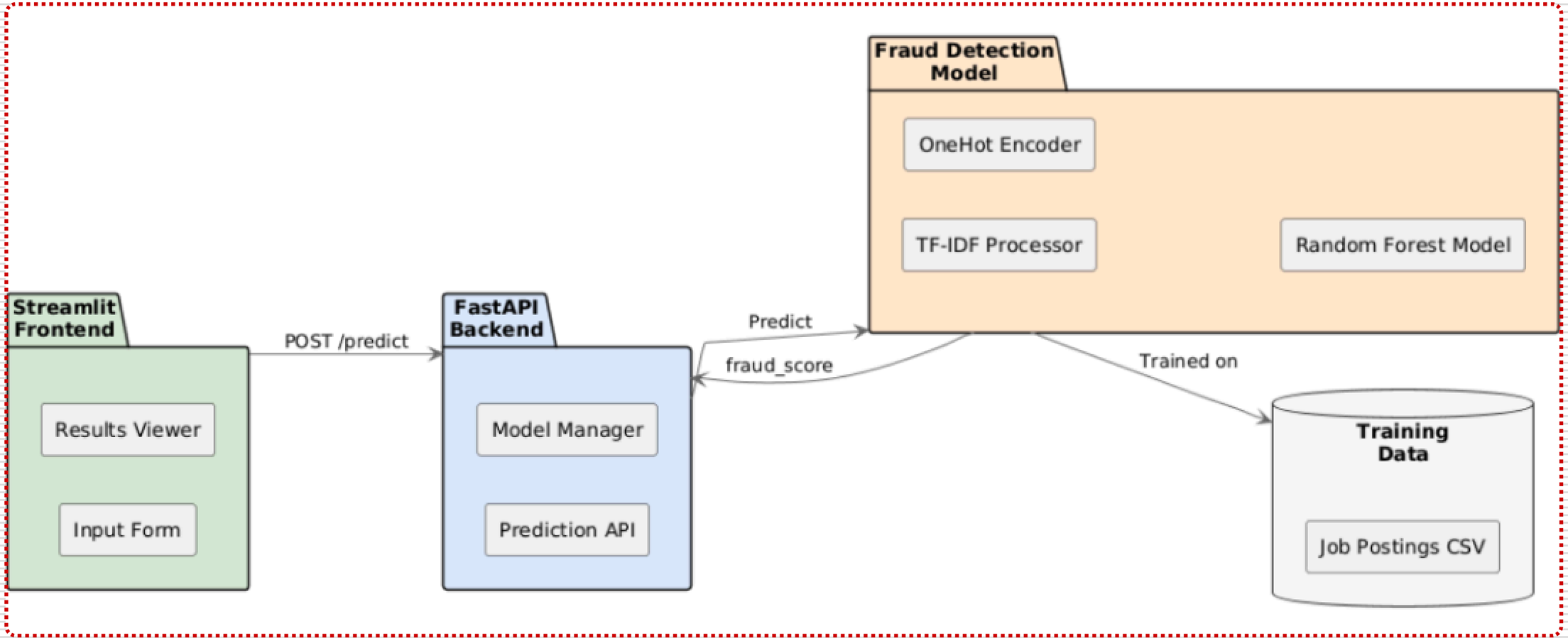
Preprocessing Pipeline:

- RandomForestClassifier for robust and accurate prediction on complex, high-dimensional data.

Strategic Feature Engineering:

- Delivers "Fake" or "Real" classification for job postings.
- Provides confidence probability scores.

System Architecture



List of Modules

Data Preprocessing & Preparation Module

Feature Engineering Module

Class Imbalance Handling Module

Machine Learning Model Module

Prediction & Evaluation Module

User Interface & API Module

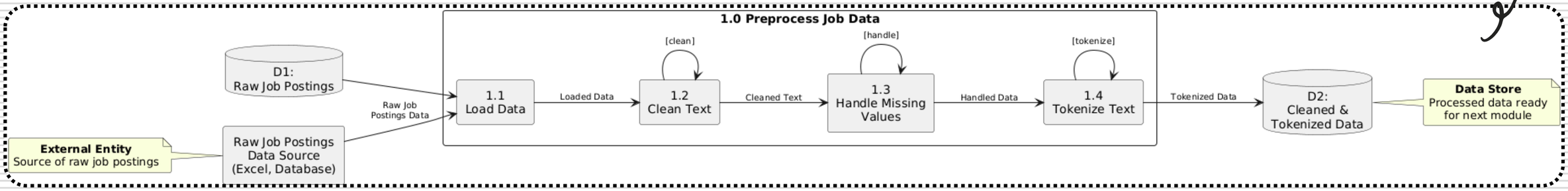
Functional Description for each modules with DFD and Activity Diagram

1. Data Preprocessing & Preparation Module

Functional Description

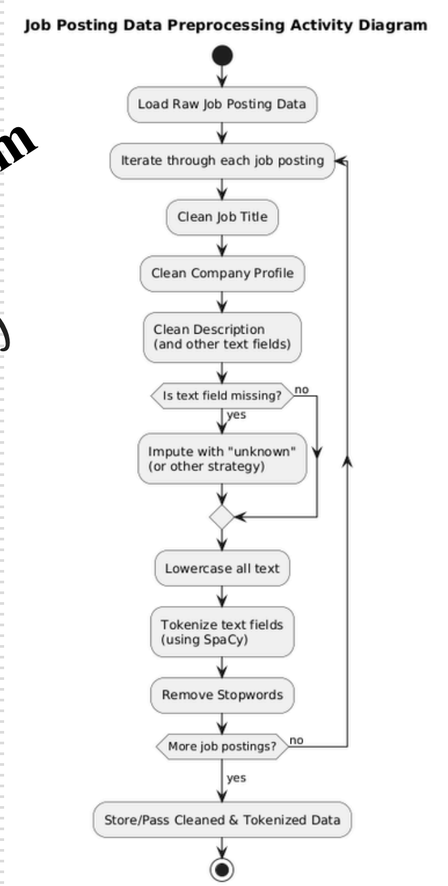
This module is responsible for taking the raw job posting data and transforming it into a clean, structured format suitable for feature engineering

DFD



Functional Description for each modules with DFD and Activity Diagram

Activity Diagram



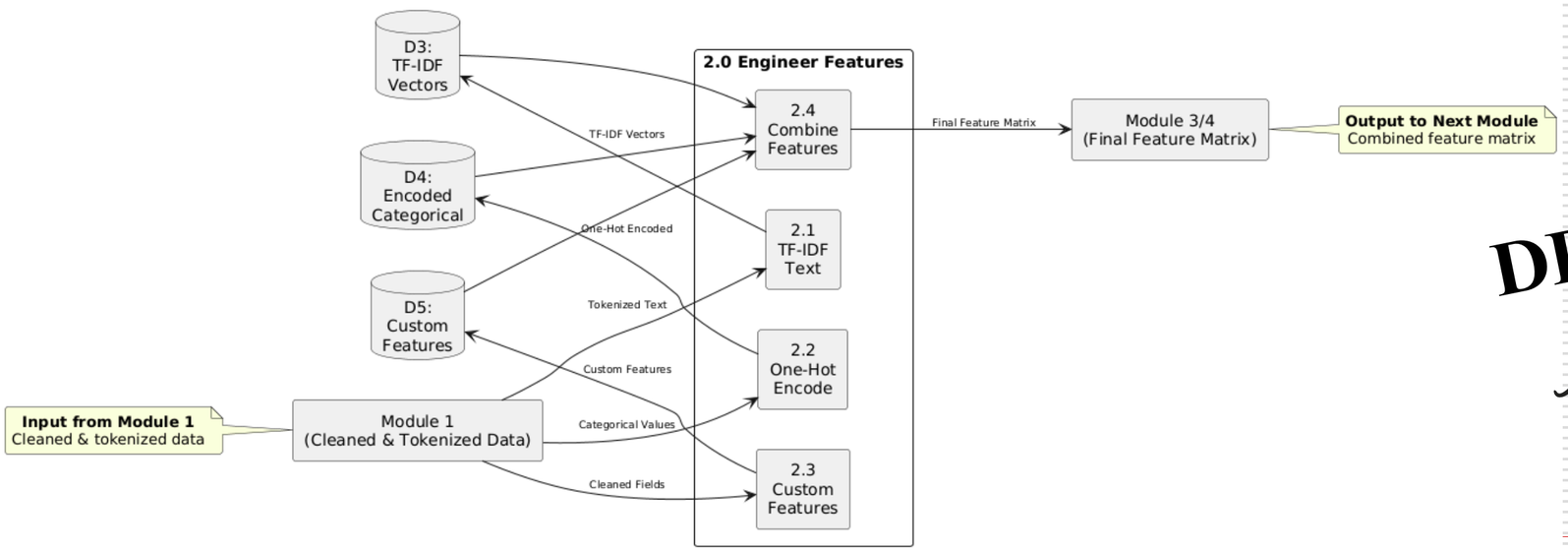
- Start Node
- Action: Load Raw Job Posting Data
- Action: Iterate through each job posting
  - Action: Clean Job Title
  - Action: Clean Company Profile
  - Action: Clean Description (etc. for all text fields)
  - Decision Node: Is text field missing?
    - Yes -> Action: Impute with "unknown" (or other strategy)
    - No -> (continue)
  - Action: Lowercase all text
  - Action: Tokenize text fields (e.g., using SpaCy)
  - Action: Remove Stopwords
- Action: Store/Pass Cleaned & Tokenized Data
- End Node

Functional Description for each modules with DFD and Activity Diagram

2. Feature Engineering Module

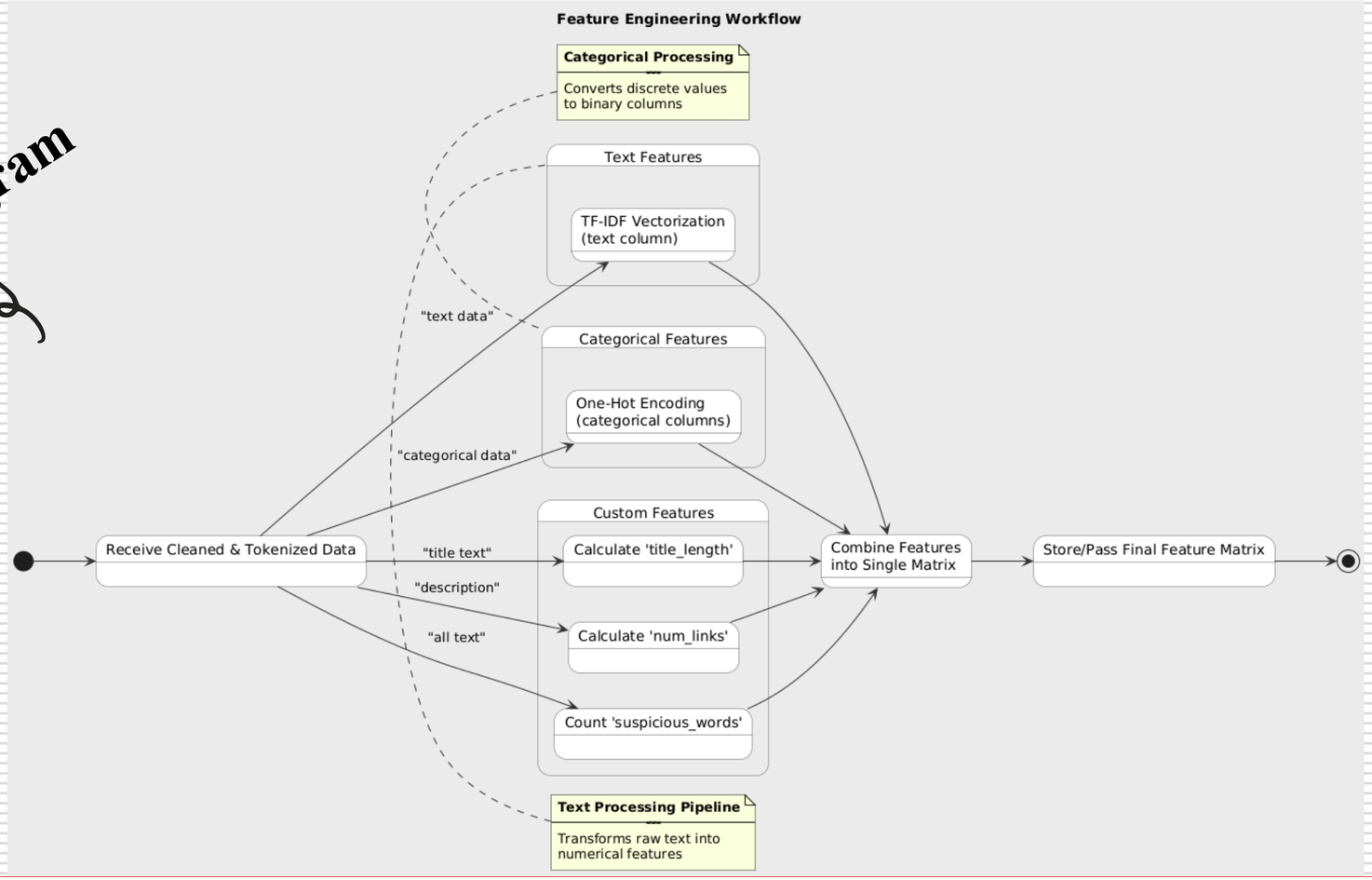
Functional Description

This module takes the cleaned and tokenized data and transforms it into a numerical feature set that the machine learning model can understand.



Functional Description for each modules with DFD and Activity Diagram

Activity Diagram

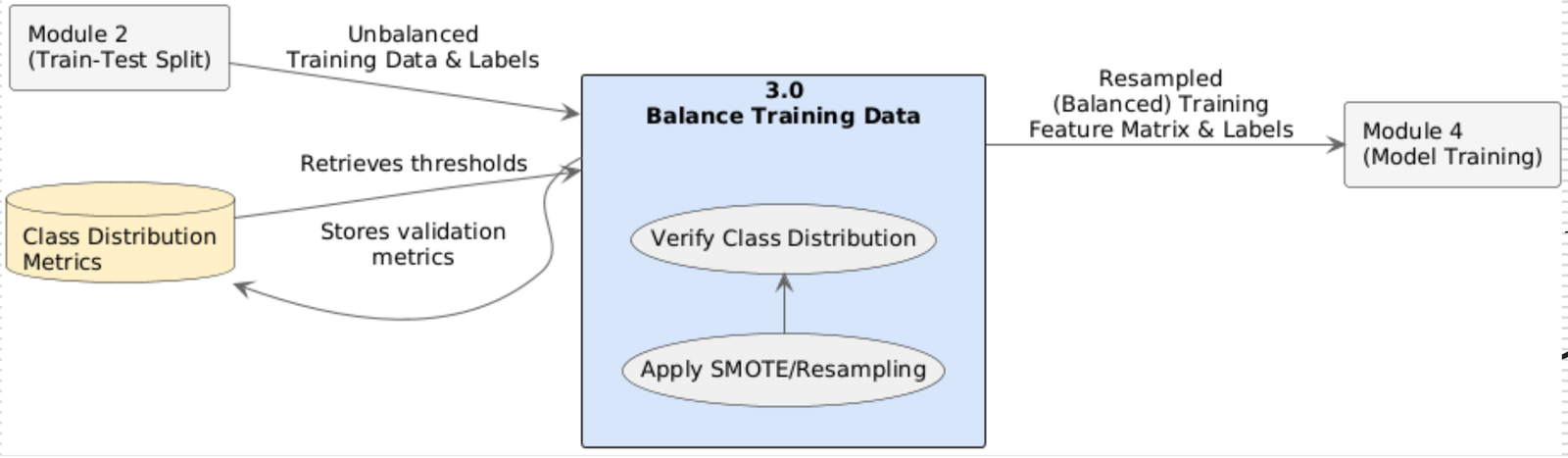


Functional Description for each modules with DFD and Activity Diagram

3. Class Imbalance Handling Module

Functional Description

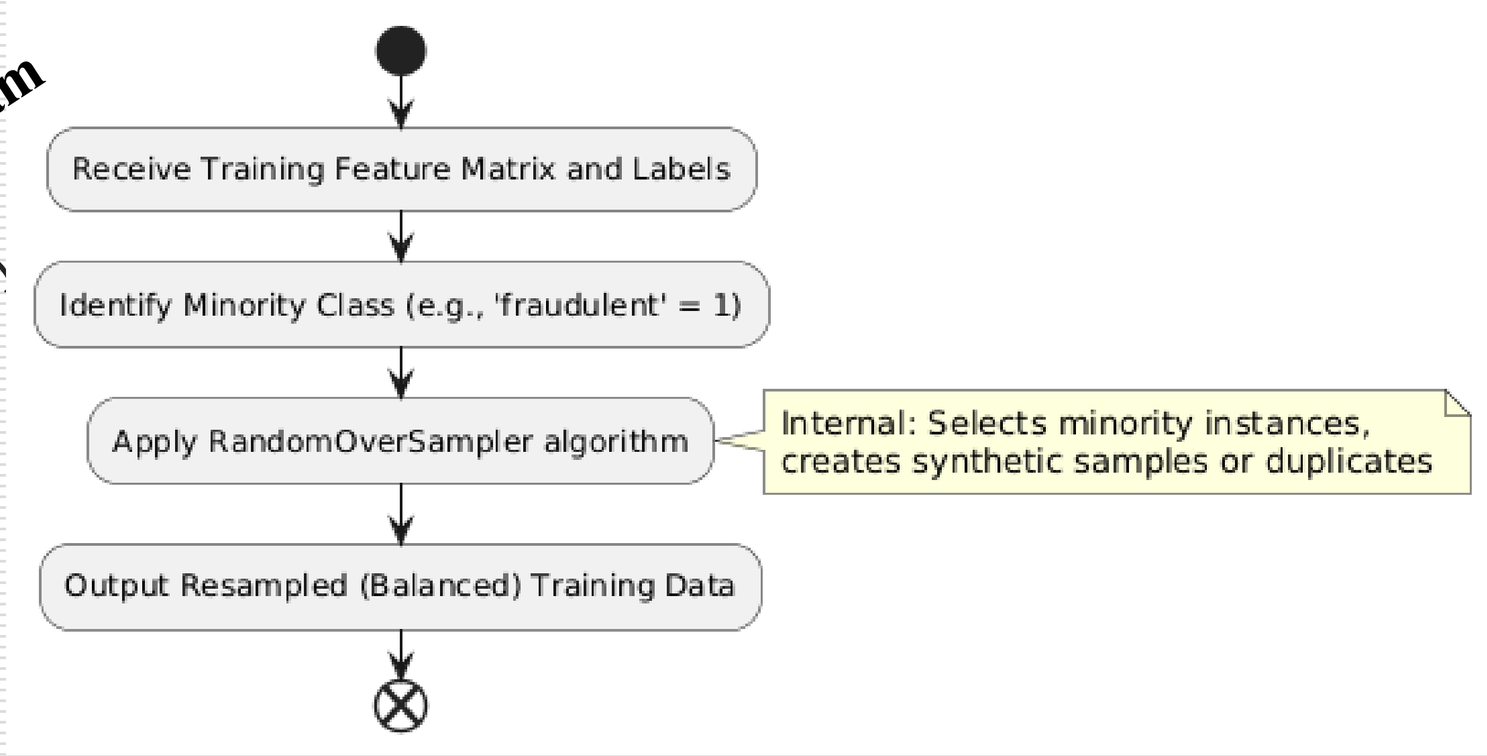
This module addresses the common issue of class imbalance in fraud detection datasets, where fraudulent instances are typically far fewer than legitimate ones.



DFD

Functional Description for each modules with DFD and Activity Diagram

Activity Diagram

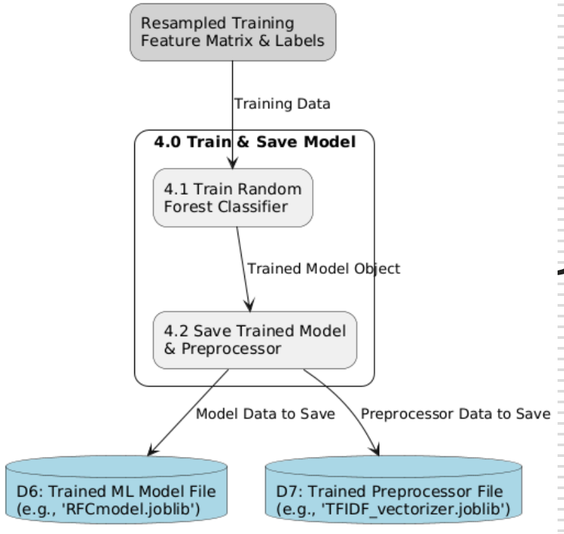


Functional Description for each modules with DFD and Activity Diagram

4. Machine Learning Model Module

Functional Description

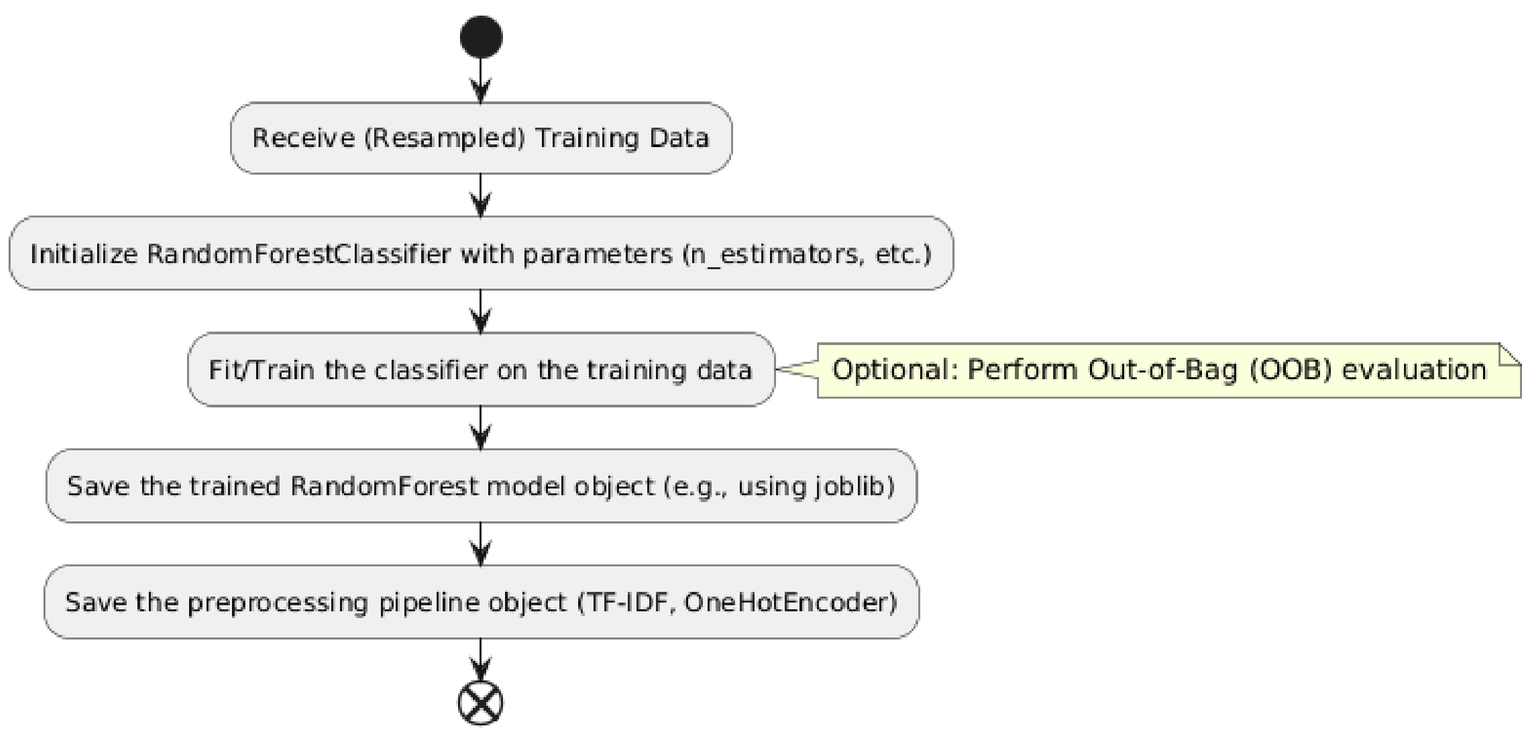
This module is the core of the prediction system. It takes the (potentially resampled) training data to train the chosen RandomForestClassifier.



DFD

Functional Description for each modules with DFD and Activity Diagram

Activity Diagram

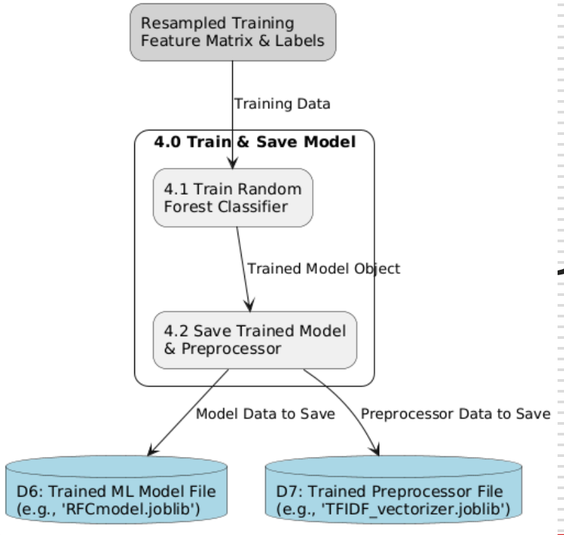


Functional Description for each modules with DFD and Activity Diagram

5. Prediction & Evaluation Module

Functional Description

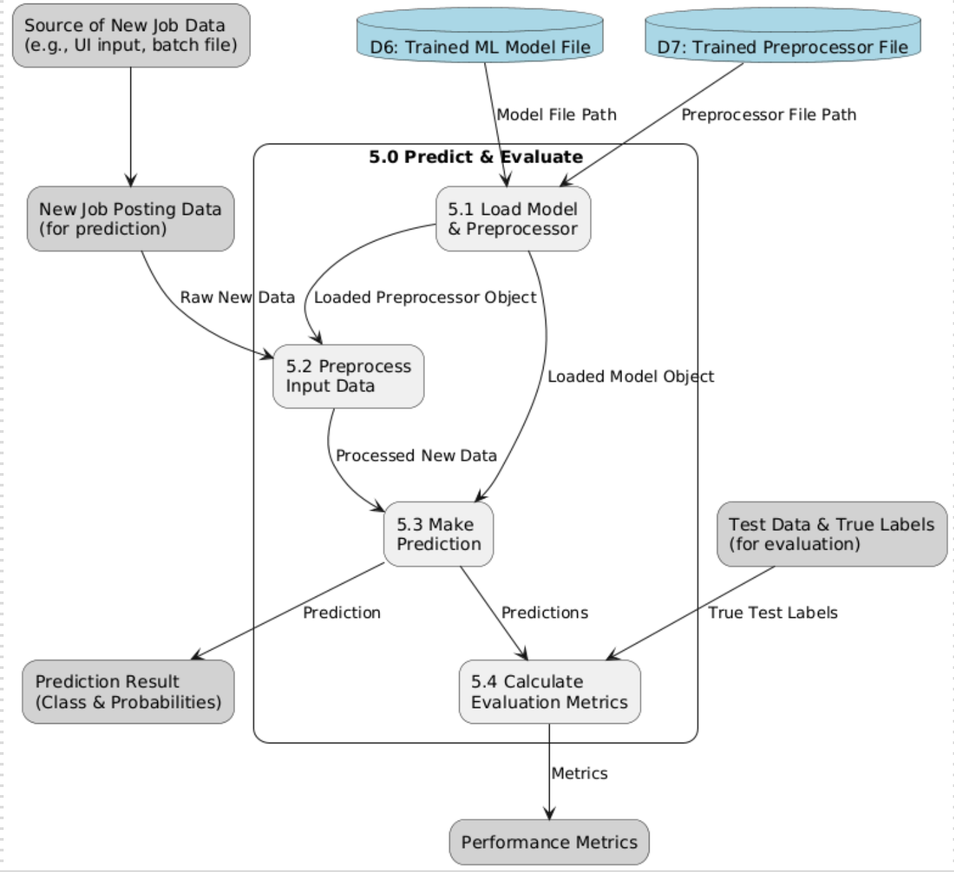
This module uses the trained model and preprocessor to make predictions on new, unseen job posting data. It first loads the saved model and preprocessing components



DFD

Functional Description for each modules with DFD and Activity Diagram

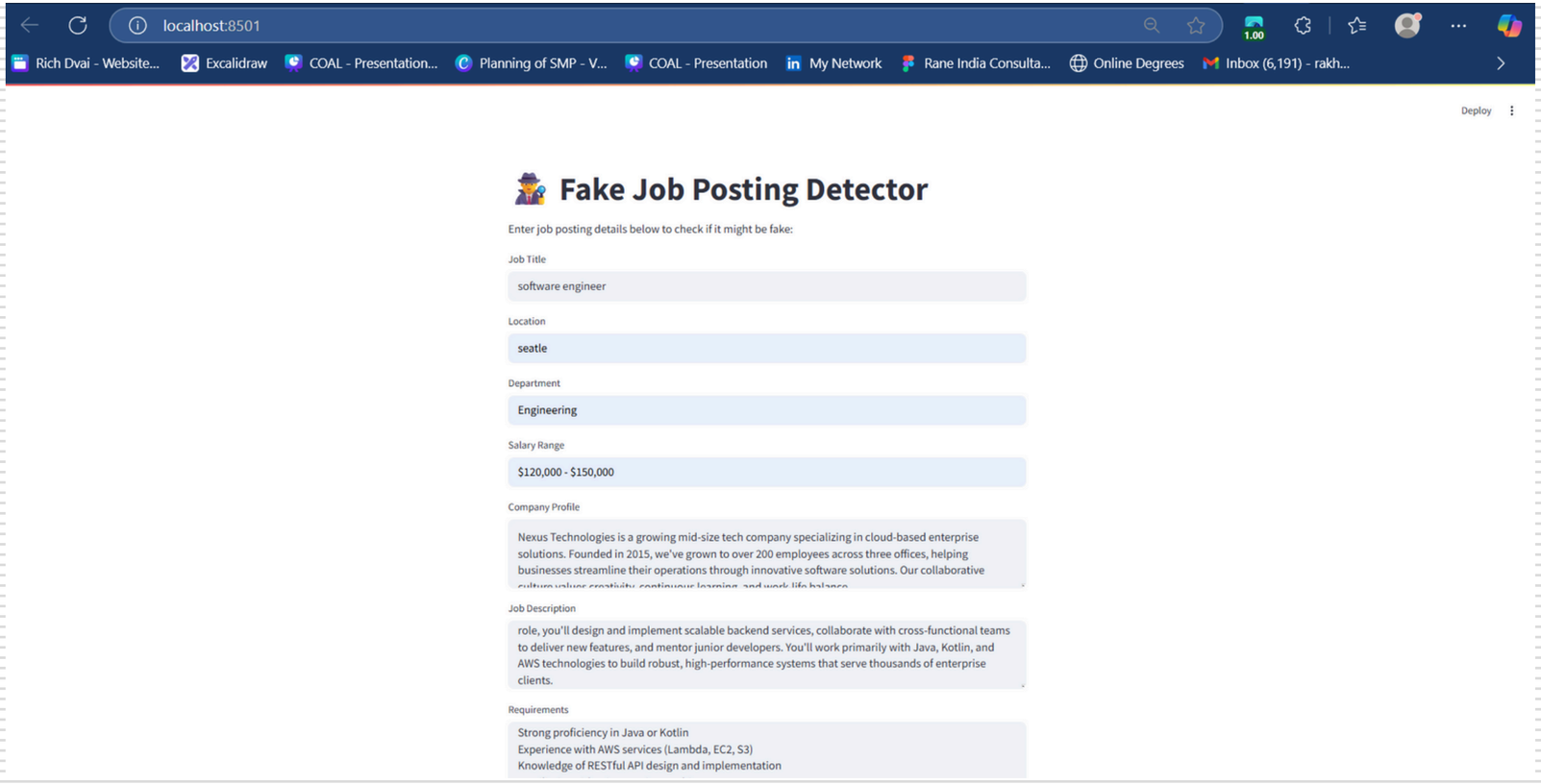
Activity Diagram



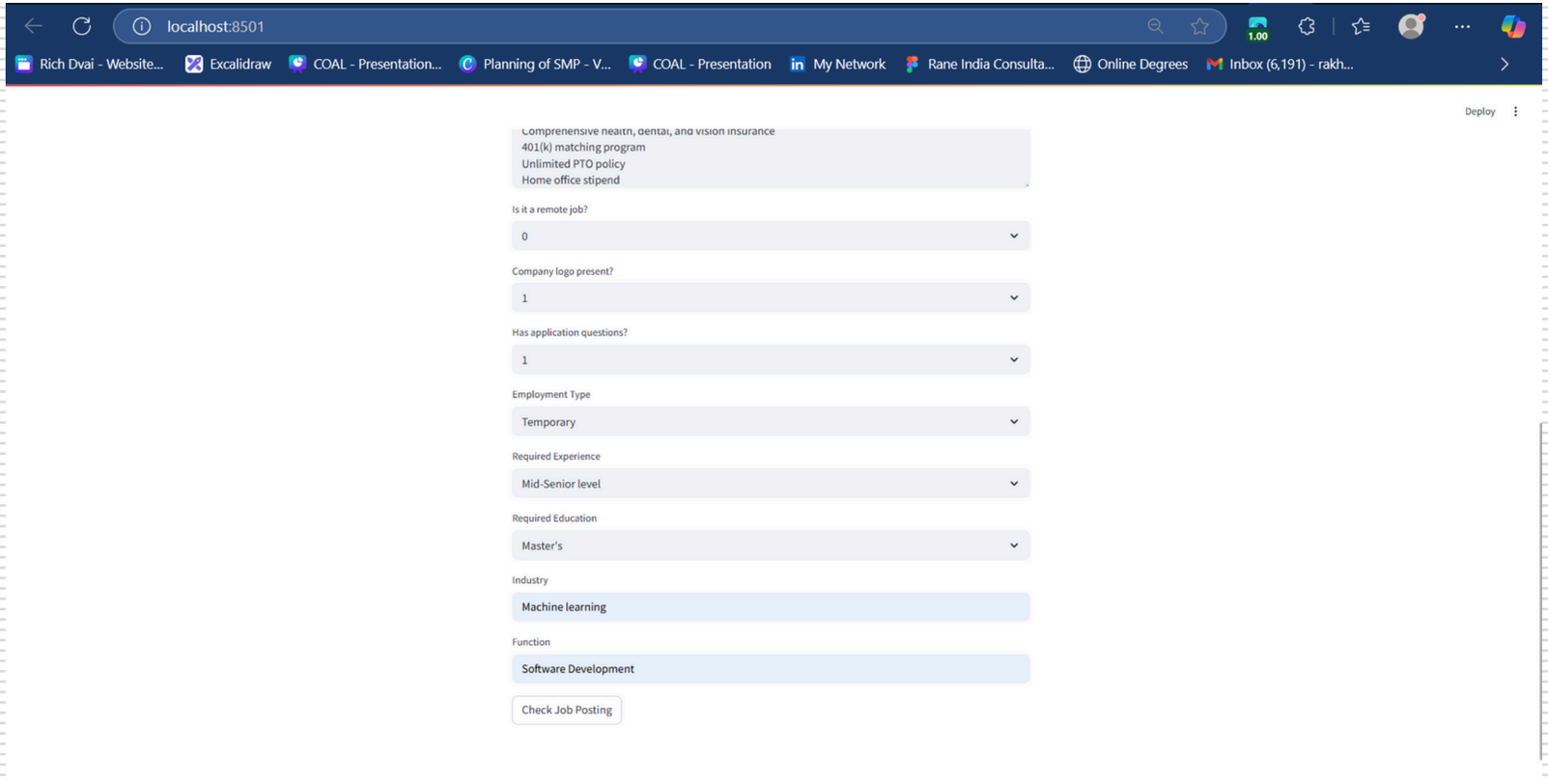
Implementation & Results of Module

Word Clouds

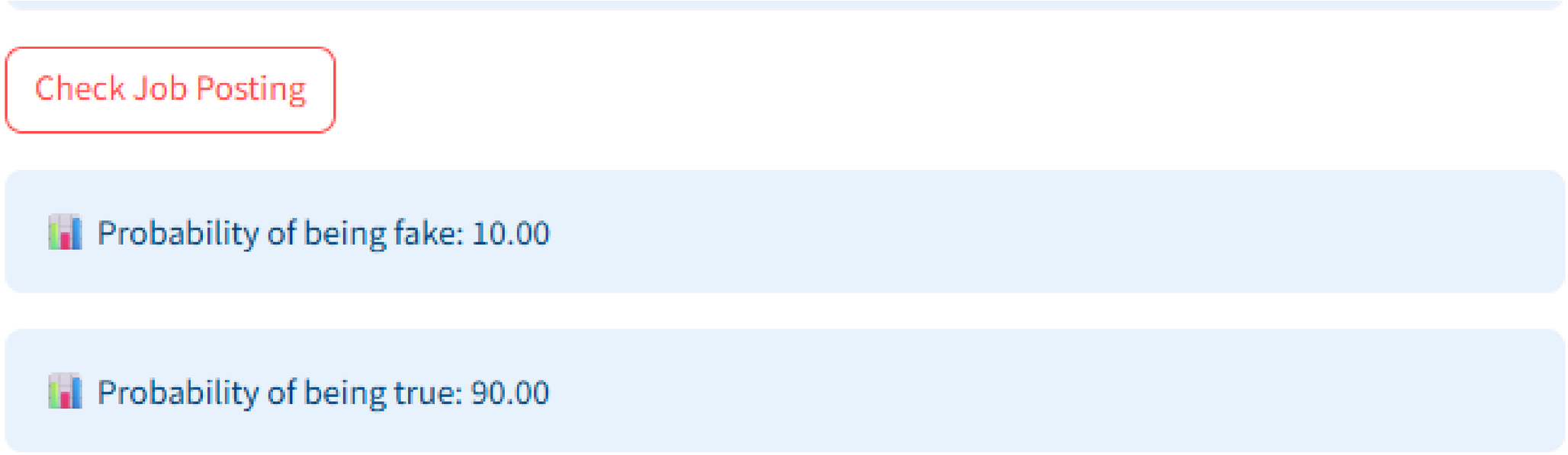
Implementation & Results of Module



Implementation & Results of Module



Implementation & Results of Module



Conclusion & Future Work

- **Thus our Fake Job Posting Predictor offers a robust, user-friendly solution for automated fake job posting detection, enhancing online platform safety for all users. This advancement paves the way for more secure job searching.**

Next Steps:

- Contextual NLP (BERT)
- Explainable AI (LIME/SHAP)
- Expanded Data (metadata, networks)
- Continuous Learning
- Scalable Deployment (containers, API)
- Multi-Language Support

References

Ahmed, A. S., Hossain, M. S., & Rahman, M. A. (2018). "Detecting Fake Job Postings Using Machine Learning Techniques."

Subramaniyaswamy, A., et al. (2020). "An Intelligent System for Fake Job Detection Using Machine Learning Techniques."

Subramaniyaswamy, A., et al. (2020). "An Intelligent System for Fake Job Detection Using Machine Learning Techniques."

Islam, M. R., Rahman, M. M., & Sadi, S. F. M. (2020). "Detecting Fraudulent Job Postings with Natural Language Processing."

Caruana, R., & Niculescu-Mizil, A. (2006). "An Empirical Comparison of Supervised Learning Algorithms."

Thank You