Progress report 2

Team member:

Ahella mohamed ali 21100824

Hanen Ehab 21100792

Ali Yasser 21100801

1. Project Overview

This project involves developing a robust machine learning pipeline for detecting offensive language in Arabic text. The main objectives include data preprocessing, model training using BERT-based transformers, and building an interactive user interface using Gradio. The project is divided into distinct phases, each focusing on different aspects of the system development lifecycle.

2. Objectives

- **Data Collection and Preparation:** Acquire and preprocess relevant datasets for training and evaluation.
- **Model Selection and Configuration**: Implement and fine-tune BERT-based models for offensive language detection.
- **Preprocessing and Feature Engineering:** Clean and preprocess Arabic text, including stop word removal and text normalization.
- **Model Training and Evaluation:** Train the model on the processed dataset and evaluate its performance.
- **GUI Development:** Build a user-friendly interface for the model using Gradio, allowing users to interact with the model and receive predictions.
- **Deployment and Testing:** Deploy the model and GUI, followed by thorough testing to ensure reliability.

3. Accomplished Progress

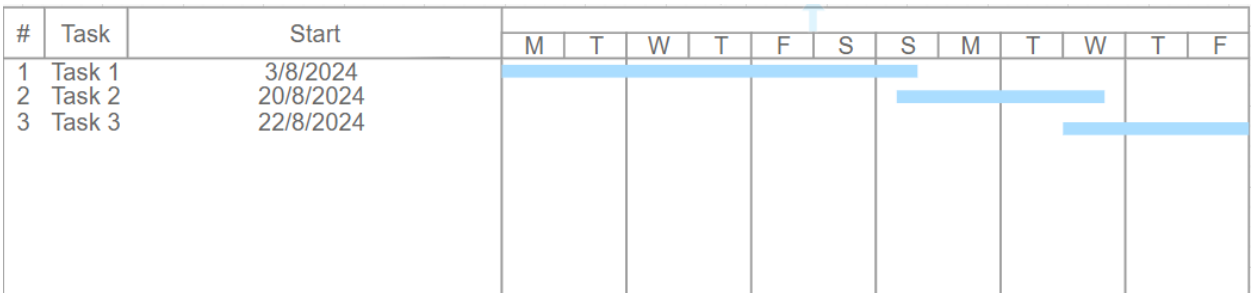
- **Data Collection:** Successfully downloaded and prepared the Arabic datasets from relevant sources.
- **Model Configuration:** Implemented BERT-based models and established pipelines for sentiment and offensive language detection.
- **Text Preprocessing:** Performed extensive preprocessing, including text cleaning, stop word removal, and feature engineering.
- **GUI Development:** Developed a Gradio-based interface, enabling user interaction with the model.

- **Model Training:** Completed initial model training and testing on the provided datasets.

4. Remaining Work

- **(Task1) Advanced GUI Customization:** Enhance the interface with additional features, modern themes, and improved user interaction.
- **(Task2) Testing and Validation:** Conduct comprehensive testing to validate model accuracy and interface usability.
- **(Task3) Documentation:** Prepare detailed documentation covering and compile all findings & implementation into a final report

5. Gantt chart



6. Challenges and Mitigations

Memory Usage: Initial issues with memory usage during model loading were addressed by optimizing data loading processes and utilizing more efficient hardware resources.

Dataset Imbalance: Managed dataset imbalance using techniques such as oversampling to improve model performance.