Progress Report 1

Team member:

Ahella mohamed ali 21100824

Ali Yasser 21100801

Hanen Ehab 21100792

Project Plan Outline

1. Project Overview

- **Objective:** To implement and evaluate adversarial attacks on an Arabic offensive language detection model using techniques like the greedy approach.
- **Scope:** The project focuses on developing a robust offensive language detection system by testing it against adversarial examples. This includes data preprocessing, model training, attack implementation, and evaluation.

2. Methodology

- **Data Preparation:**

Collect and preprocess the Arabic offensive language dataset.

The **data** includes: the training, development, and test data sets.

The **preprocessing** includes:

  - **Normalization**
  - **Diacritic Removal**
  - **Punctuation Removal**
  - **Stop Words Removal**
  - **Tokenization**
- **Model Training:**

Use a pre-trained BERT model fine-tuned for the specific task of offensive language detection.

Fine-Tuning:

- **Model:** `qarib/bert-base-qarib`
- **Learning Rate:** 2e-5
- **Batch Size:** 64
- **Epochs:** 1

The model is fine-tuned on the labeled Arabic tweet dataset using these parameters to balance training efficiency and performance. The choice of learning rate and batch size is based on empirical best practices for BERT-based models.

- **Adversarial Attack Implementation:**

Attack Strategy: We employ a black-box greedy adversarial attack approach, consisting of three stages:

  - **Token Masking:** Replace a word in the text with a [MASK] token.
  - **Unmasking:** Use the BERT model to predict possible substitutes for the masked token.
  - **Substitute Selection:** Choose the most similar substitute using a pre-trained Fast Text model to find the closest word vector.

Purpose of the Attack: To evaluate the classifier's vulnerability to perturbations and understand how such attacks can degrade its performance.

- **Evaluation Metrics:**

Use accuracy, precision, recall, and F1 score to measure the model's performance.

These metrics are chosen to provide a comprehensive evaluation of the model's performance, capturing its ability to classify offensive language accurately and reliably.

3. Timeline and Gantt Chart

Phases and Milestones:

Task 1

- Data Preparation: Complete data collection and preprocessing.

Task 2

- Model Training: Fine-tune the BERT model and validate its performance.

Task 3

- Adversarial Attack Development: Implement the greedy approach and test it.

Task 4

- Evaluation and Analysis: Evaluate the model's robustness and document findings.

Task 5

- Project improvements: Add one more model & improve the performance (InProgress)

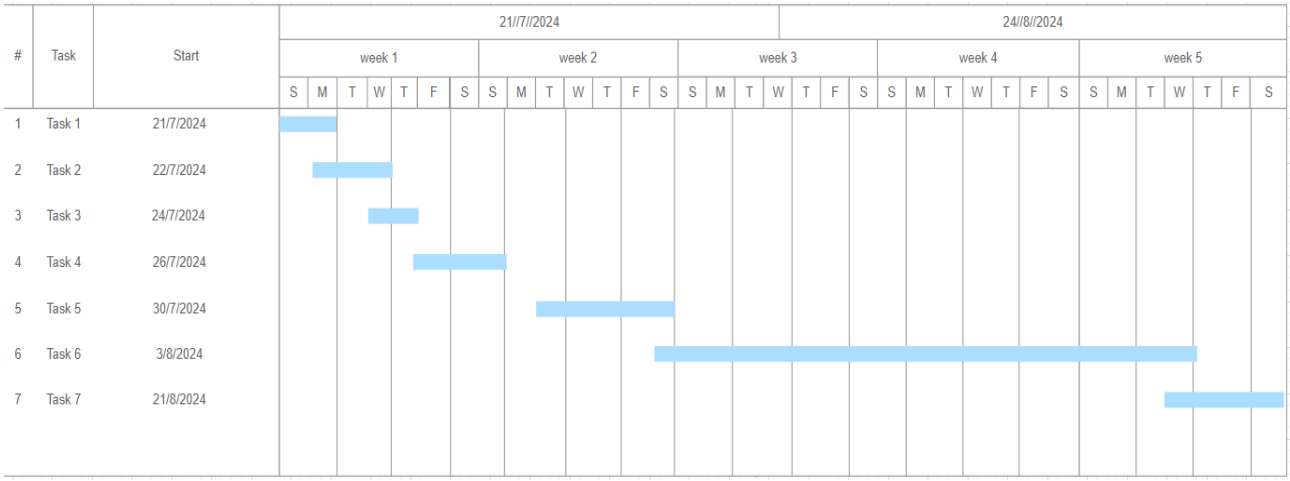
Task 6

- Project app: Implement the project using a useful application (InProgress)

Task 7

- Final Report: Compile all findings & implementation into a final report (InProgress).

Gantt Chart:



4. Progress and Accomplishments

- **Completed Tasks:**

Collected and preprocesses the dataset.

Fine-tuned the BERT model for offensive language detection.

- **Ongoing Work:**

Implementing new model and refining the greedy adversarial attack approach.

- **Future Work:**

Evaluate the new model's performance under adversarial attacks.

Turn this project into a useful app and then compile results and write the final report.