Overviews

The project focuses on detecting hate speech in Arabic tweets using a fine-tuned BERT model and evaluating the model's robustness against adversarial examples. Significant steps have been taken to advance the project's goals.

Work Completed

1. **Model Selection and Fine-Tuning**
    - We selected the 'qarib/bert-base-qarib' model for its effectiveness in Arabic text processing.
    - The model has been fine-tuned on a dataset of Arabic tweets labeled for hate speech, ensuring it can accurately identify hateful content.
2. **Data Preparation**
    - Arabic text data was preprocessed, including tokenization, removal of stop words, and handling of prefixes/suffixes, to improve model performance.
    - A list of Arabic stop words was utilized to clean the data, focusing the model on meaningful content.
3. **Adversarial Attack Design**
    - A blackbox greedy approach was designed to test the model's robustness:
        1. **Word Masking:** Identifying and masking target words in the text.
        2. **BERT Substitution:** Using BERT to suggest alternative words for the masked terms.
        3. **Similarity Matching:** Selecting the most similar word to replace the original, aiming to create adversarial examples that challenge the model's accuracy.
4. **User Interaction**
    - Code was developed to allow users to input a line of text, which the model then classifies as hate speech or not. If hate speech is detected, the specific hateful word is identified.
5. **Challenges and Adjustments**
    - Initial testing revealed some false positives and issues with movement sensitivity in the Serbot integration, which were addressed by adjusting the detection confidence threshold and refining the logic for centering detected persons.

Next Steps

- **Further Testing:** Continue to test the model with additional adversarial examples to assess its robustness and make necessary adjustments.
- **Integration:** Integrate the hate speech detection system with the Serbot project, ensuring smooth interaction between the detection model and the robot's behavior.
- **Optimization:** Focus on optimizing the processing pipeline for better performance and accuracy in real-world scenarios.