**Week3 Weekly Summary Report**

1. **Summary of Weekly Progress:**

During week three, I also created the foundation of the preprocessing pipeline of the cyberbullying detecting project. The raw social media Twitter text was ingested and filtered on incomplete or missing entries. There was removal of duplicated posts and in-relevant posts like hashtag only tweets and spam URLs. Other textual data was normalised by lowering, expansion of contractions, URLs and symbols, tokenisation, lemmatisation and stopword elimination. Labels were represented using numbers and other engineered features were added, such as statistical counts, sentiment scores and abusive keyword flags. TF-IDF was used to transform the processed text into vectors, and the data was divided into a stratified training (70%), validation (15%), and test (15%) data. A logistic regression was used as a sanity check to test this model after which the validation accuracy was about 0.91. Each procedure of preprocessing was recorded to make it repeatable.

1. **Functions/Procedures Implemented:**

Some of its most important parts are missing and incomplete data management, removal of duplicates and irrelevant text, text normalizing, tokenization and lemmatization, encoding of labels, feature engineering (statistical, sentiment, abusive keyword flags) and splitting the dataset. An order of operations is achieved by a master pipeline function. Each of the functions was tested on sample data and confirmed to perform correctly on the complete dataset without run time error.

**Challenges Faced & Problem-Solving Approach:**

There were dependency problems in Colab, such as the lack of pre-installed packages like emoji and English model of spaCy. To enable quick execution optional imports were enclosed in safe import blocks that ethically disable associated features when packages are absent that make the code to operate in the various environments without making manual adjustment.

1. **Feedback from Supervisor & Resolution:**

Supervisor asked that there should be better logging and reproducibility. I therefore included hard random seeds so that splits are consistent, displayed compact statistics with every step, and created a JSON file with the summary of the datasets of each split, the number of features, and validation metrics. These innovations paid off with respect to the feedback it was receiving.

1. **Plan for Next Week:**

I intend to incorporate word embedding concepts like Word2Vec, BERT sentence embeddings, test the baseline of the classifier like SVM and XGBoost, and begin the hyperparameter optimization process. Writing the interim results chapter will also be started.

1. **Additional Notes or Requests:**

The choice of the desirable embedding technique is under confirmation. A bigger labeled dataset will be availed after the completion of ethical clearance processes.

**GitHub Screenshots:**

A screenshot of a computer program

AI-generated content may be incorrect.A computer screen with white text

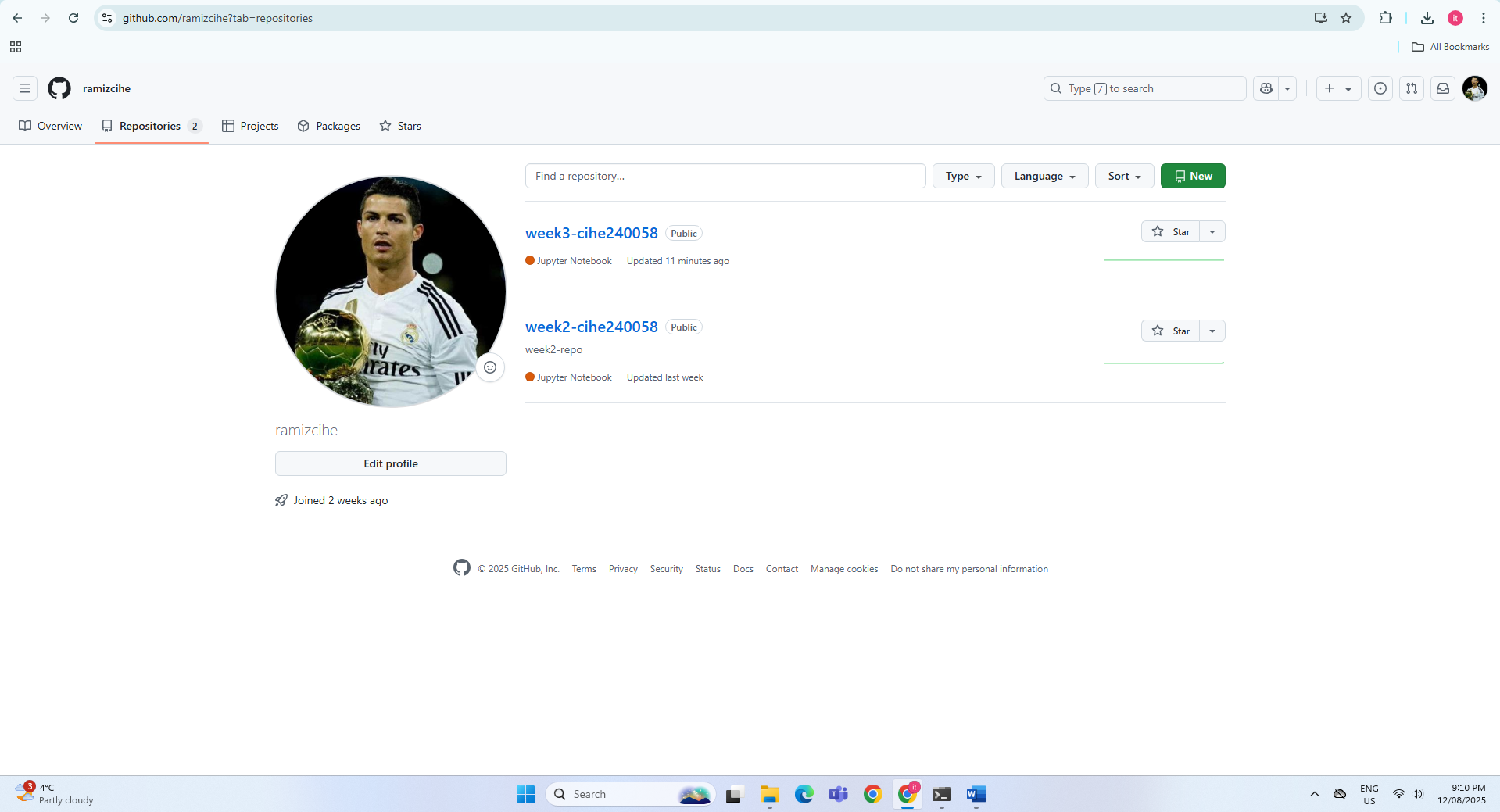
AI-generated content may be incorrect.A computer screen with white text

AI-generated content may be incorrect.A screenshot of a computer program

AI-generated content may be incorrect.A screenshot of a computer program

AI-generated content may be incorrect.A screenshot of a computer program

AI-generated content may be incorrect.

 A screenshot of a computer

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect.