Machine Learning

Name	***SRN***
Navyata Venkatesh	PES2UG23CS375
Pinisetti Sudhiksha	PES2UG23CS916

Project Title: ***Trolls, Haters or Teammates: Classifying Sentiment in PvP Game Communications.***

Problem Statement:

Player vs Player (PvP) games often experience toxic behaviour, bullying, and hateful messages through in-game chat. Traditional moderation tools rely on profanity filters or static word lists, which miss slang, evolving language, and context-dependent toxicity.
This project builds a supervised machine learning model to accurately classify messages into:

- Neutral/Positive
- Negative
- Derogatory

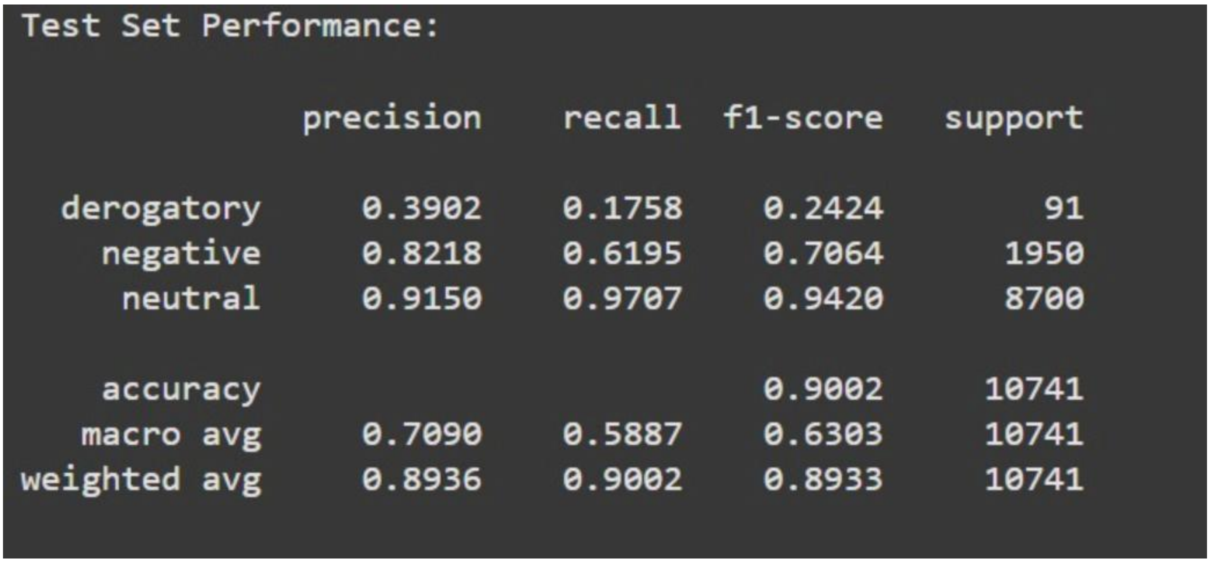
High Level Architecture:

1. Data Ingestion:
   The dataset (gametox.csv) containing real player chat messages and their labels is loaded.
2. Data Preprocessing:
   - Text is cleaned by lowercasing and removing special characters.
   - Gamer slang is normalized into standard language to improve understanding.
3. Feature Extraction:
   - The cleaned text is converted into numerical features using TF-IDF (unigrams and bigrams), which captures the importance of words and word pairs in each message.
4. Model Training:
   - A Logistic Regression classifier is trained with class weights to give more importance to less frequent derogatory messages.
5. Model Evaluation:
   - The trained model is validated and tested using classification reports and confusion matrices to measure precision, recall, and F1-scores.

6. UI:

    - A simple interactive interface is built with Gradio so anyone can enter a chat message and instantly get a predicted sentiment label.

Results:



```
Test Set Performance:

              precision    recall  f1-score   support

  derogatory     0.3902    0.1758    0.2424        91
    negative     0.8218    0.6195    0.7064      1950
     neutral     0.9150    0.9707    0.9420      8700

    accuracy                         0.9002     10741
   macro avg     0.7090    0.5887    0.6303     10741
weighted avg     0.8936    0.9002    0.8933     10741
```