# Replication Instructions for SVM + TF-IDF Model Evaluation

2728711 WEI LI

**Introduction**: This document provides step-by-step instructions to replicate the results obtained using the SVM + TF-IDF approach for text classification. Follow the instructions below to replicate the reported results.

**Step 1: Environment Setup**:

- Set up a Python environment following the requirements.pdf. Ensure that all dependencies are installed.

**Step 2: Dataset**:

- Download the dataset used in the experiment (e.g., "incubator-mxnet.csv") or use a similar CSV dataset with columns: Title, Body, and Class.
- The Class column should contain the labels for classification.

**Step 3: Preprocessing**:

- The code will automatically preprocess the dataset by removing HTML tags, emojis, stopwords, and cleaning the text using the function clean_str.
- The TF-IDF vectorizer will be applied to the cleaned text, and features will be extracted for training the model.

**Step 4: Model Training**:

- It applies SVM to find the best hyperparameters (C, kernel, gamma).

**Step 5: Evaluation**:

- The performance of the model is evaluated using the following metrics:

  **Accuracy**  
  **Precision**  
  **Recall**  
  **F1 Score**  
  **AUC (Area Under the ROC Curve)**

- The evaluation results will be printed on the screen and saved to evaluation_results.csv.

**Step 6: Results**:

- After running the code, check the output for the average performance metrics. These metrics should be similar to the ones reported in the original experiment.

**Step 7: Reproducibility**:

- Ensure the dataset is the same, the environment is set up as per the requirements.pdf, and the code is executed as instructed.
- If any discrepancies arise, check that the preprocessing steps are correctly followed and that the same model parameters are used.

**Conclusion**: By following these steps, you should be able to replicate the reported results and verify the performance of the model.