User Manual for Text Classification Tool using SVM and TF-IDF

Introduction: This tool is designed to perform text classification using a Support Vector Machine (SVM) classifier and TF-IDF for feature extraction. This manual explains how to set up and use the tool.

Installation:

- Follow the installation steps in the requirements.pdf to set up the dependencies and environment.
- Ensure you have a valid dataset for training and testing the model. The code assumes the dataset is in CSV format with columns Title, Body, and Class.

Usage Instructions:

1. Prepare Your Dataset:

   Ensure the dataset is in a CSV format with the following columns: Title, Body, and Class.

   The Class column should contain the labels for classification.

2. Run the Code:

   The script can be executed via the command line:

   ```
   python svm.py
   ```

   The code will preprocess the data, train the SVM model, and output evaluation results (Accuracy, Precision, Recall, F1 Score, AUC).

3. Customizing the Dataset Path:

   Update the data_file variable in svm.py to point to your dataset file.data_file = "path/to/your/dataset.csv"

Key Features:

- Text Preprocessing: Handles cleaning of HTML tags, emojis, stopwords, and special characters.
- TF-IDF Feature Extraction: Converts text into numerical features using TF-IDF with n-grams.
- SVM Model: Uses a Support Vector Machine with grid search for hyperparameter optimization.
- Cross-Validation: Implements Stratified K-fold cross-validation for robust model evaluation.
- Metrics: Outputs multiple performance metrics, including accuracy, precision, recall, F1 score, and AUC.

Example Output: After running the code, you will see output similar to:

=== Model Evaluation Results ===
Average Accuracy: 0.9231
Average Precision: 0.9123
Average Recall: 0.8987
Average F1 Score: 0.9000
Average AUC: 0.9589

Saving Results:

- The results are saved in a CSV file named evaluation_results.csv.

Troubleshooting:

- Missing Libraries: Ensure all required libraries are installed as per the requirements.pdf.
- Dataset Format: Double-check your dataset's format, especially the column names.