Program: MS Data Science | Subject: Advanced Natural Language Processing
Group: A & B | Date: 9-11-2024
Exam: Mid-Term (Fall 2024) | Time: 9-11-2024 to 16-11-2024

Instructions: Answer the following questions based on the provided dataset of Urdu movie reviews. Implement and demonstrate your code, along with results, for each question.

Q-1: Data Loading and Preliminary Analysis (10 Marks)

a. Load the Urdu movie reviews dataset into a Pandas DataFrame.
    a. a. Show the first few rows of the dataset.
    b. b. Display the column names and their data types.
b. Perform a preliminary analysis:
    a. a. Check for null values and handle any missing data if present. Explain your approach.
    b. b. Identify and analyze the distribution of the classes (positive, negative, etc.) in the dataset. Are they balanced?

Q-2: Data Visualization (10 Marks)

a. Generate and display a **word cloud** from the Urdu movie reviews text to visualize commonly used words.
b. Based on the word cloud, describe any key insights about the most frequently appearing words. Do you notice any themes or patterns?

Q-3: Data Preprocessing (20 Marks)

List and apply necessary preprocessing tasks to clean and prepare the Urdu text for feature extraction:

a. **Text Normalization**: Remove punctuation, special characters, and numbers.

b. **Tokenization**: Split the Urdu text into individual words.

c. **Stopword Removal**: Remove Urdu stopwords to improve classification accuracy.

d. **Stemming or Lemmatization** (if applicable): Briefly describe what stemming or lemmatization is and how it might help in the context of Urdu text processing.

Q-4: Feature Extraction Techniques (20 Marks)

a. Extract features from the pre-processed text using:
    a. **Unigrams** (single words)
    b. **Bigrams** (two consecutive words)
    c. **Trigrams** (three consecutive words)
    d. **Unigrams + Bigrams**
    e. **Unigrams + Bigrams + Trigrams**

b. Explain each of the above feature extraction methods briefly. What information does each capture?
c. Use **TF-IDF (Term Frequency-Inverse Document Frequency)** as an additional feature extraction method and explain why it might be helpful in text classification.

Q-5: Classification with Machine Learning Algorithms (20 Marks)

Implement the following machine learning algorithms for classifying Urdu movie reviews:

- Naïve Bayes
- Support Vector Machine (SVM)
- Decision Tree

- Random Forest
- k-Nearest Neighbors (k-NN)

2. Use **Stratified K-Fold Cross-Validation** (with k=5) to evaluate each classifier. Explain why stratification is used and its importance in this context.
3. For each algorithm, report the following performance metrics:
    - Accuracy
    - Precision
    - Recall
    - F1 Score

Q 6: Comparative Analysis and Visualization (20 Marks)

1. Plot bar charts to compare the performance (Accuracy, Precision, Recall, F1 Score) of the different machine learning algorithms you applied. Ensure that each chart is well-labeled and contains a legend if needed.
2. Based on the results, provide a detailed analysis:
    - a. Which algorithm performed the best for classifying Urdu movie reviews, and why do you think it performed better?
    - b. Identify any challenges you encountered when using each algorithm.
    - c. What additional steps could you take to improve the model's performance?

Q 7: Choose an ensemble classification method (e.g., **Bagging**, **Boosting**, or **Voting**). Explain why you selected this ensemble method and how it can potentially improve the classification performance for the Urdu movie review dataset. **(20 Marks)**

Implementation:

- Implement your chosen ensemble method using a combination of at least three of the classifiers from Question 5 (e.g., Naïve Bayes, SVM, Random Forest, etc.).

Analysis:

- Discuss the results. Did the ensemble classifier outperform individual classifiers? Why or why not? Include a discussion on any trade-offs or limitations observed with the ensemble approach.