ECON6087: Assignment 1
Chinese Text Classification with Machine Learning

Machine Learning for Economists

Due Date: [Insert Date]

Introduction

In the tutorial, we classified English news using the AG News dataset. In this assignment, you will apply the same techniques to a Chinese dataset: **TNEWS** (Toutiao News) from the CLUE benchmark. You will handle the unique challenge of Chinese word segmentation using the `jieba` library.

Requirements

- You must use Python for this assignment.
- Required libraries: `pandas`, `scikit-learn`, `datasets`, `jieba`.
- Submit your code as a Jupyter Notebook (.ipynb) or a Python script (.py).

Task 1: Data Preparation

1. Load the **TNEWS** dataset using the command: `load_dataset("clue", "tnews")`.
2. Convert the `'train'` and `'validation'` splits into pandas DataFrames. (We use the validation set as our test set for this assignment).
3. Inspect the data. Display the first 5 rows of the training dataframe.

Task 2: Chinese Segmentation with Jieba

Since Chinese text does not use spaces to separate words, you must segment the text before vectorization.

1. Import the `jieba` library.
2. Create a function that takes a Chinese string and returns a space-separated string of words using `jieba.cut`.
3. Apply this function to the text column of both your training and testing (validation) dataframes.
4. Create a new column (e.g., `text_cut`) to store these segmented strings.

Task 3: Bag of Words Classification

1. Use `CountVectorizer` to convert the segmented text (`text_cut`) into a numerical matrix.
2. Limit the vocabulary size to the top **5,000** features.
3. Train a **K-Nearest Neighbors (KNN)** classifier ($k = 5$) on the training data.
4. Report the classification accuracy on the test set.

Task 4: TF-IDF Classification

1. Use `TfidfVectorizer` to convert the segmented text (`text_cut`) into a numerical matrix.
2. Keep the vocabulary limit at **5,000** features.
3. Train a **K-Nearest Neighbors (KNN)** classifier ($k = 5$) on the training data.
4. Report the classification accuracy on the test set.

Task 5: Comparison and Analysis

- Compare the accuracy of BoW vs. TF-IDF. Which one performed better?
- Explain why one method might be better suited for short news headlines (TNEWS) than the other.