Machine Learning with Python-Practice Exercise

MLA- AIML Batch

NLP Practice Assignments

Day 2

You are part of a team developing a text classification system for a news aggregator platform. The platform aims to categorize news articles into different topics automatically. The dataset contains news articles along with their corresponding topics. Perform only the Feature extraction techniques.

Dataset Link: https://www.kaggle.com/datasets/therohk/million-headlines

Data Exploration: Begin by exploring the dataset. What are the different topics/categories present in the dataset? What is the distribution of articles across these topics?

Bag-of-Words (BoW): Implement a Bag-of-Words (BoW) model using CountVectorizer or TF-IDF to transform the text data into numerical features. Discuss the advantages and limitations of BoW in this context. Apply both unigram and bigram techniques and compare their effects on classification accuracy.

N-grams: Explore the use of N-grams (bi-grams, tri-grams) in feature engineering. How do different N-gram ranges impact the performance of the classification model?

TF-IDF: Apply TF-IDF (Term Frequency-Inverse Document Frequency) to the text data. Describe how TF-IDF works and its significance in capturing the importance of words across documents. Compare the results of TF-IDF with the BoW approach.

One-Hot Encoding: Investigate the application of One-Hot Encoding to encode categorical variables or labels. Can One-Hot Encoding be used directly for text classification? Why or why not?

Deliverables:
Present insights gathered from data exploration and discuss the impact of different feature engineering techniques (BoW, N-grams, TF-IDF, One-Hot Encoding). Provide recommendations for the best feature engineering strategy.