Classifying Newsgroup Topics on the 20 Newsgroups Dataset

I. Dataset

The dataset used in this study is the *20 Newsgroups dataset*, which contains approximately 20,000 documents categorized into 20 different newsgroups, such as sports, politics, and technology. The dataset's business application lies in developing robust models for automatic content categorization, email filtering, and sentiment analysis in various domains. It has been widely used in research as a benchmark for text classification models [1].

II. Classification Pipeline

The classification pipeline involves several key steps:

1. *Text Preprocessing*: Text was lowercased, non-alphabetic characters were removed, tokenized, and stopwords were filtered out. Lemmatization and stemming were applied to standardize the vocabulary and reduce word variability.

2. *Text Augmentation*: Using WordNet, synonymous words were introduced to enrich the dataset semantically, improving generalization.

3. *Vectorization Pipelines*: - **TF-IDF Vectorizer**: This method assigns weights to terms based on their relative importance in a document compared to the entire corpus. - **Bag-of-Words (BoW)**: A basic word frequency count model without consideration for word importance.

For each pipeline, five models were compared: Naive Bayes, Logistic Regression, SVM, Random Forest, and K-Nearest Neighbors. SVM with the TF-IDF vectorizer had the highest mean accuracy (0.8777), while Logistic Regression performed best with BoW (0.8474). TF-IDF was superior because it adjusts for word relevance, critical for the diverse topics in this dataset.

III. Evaluation

The SVM model was fine-tuned using grid search, optimizing hyperparameters (*C*=10, *class_weight*=balanced, *gamma*=scale). The tuned model achieved a marginally improved accuracy of 0.8798, highlighting that the initial configuration was already near-optimal.

Top Features: The most important words identified, such as "atheist", "god", and "religion", clearly align with specific newsgroup topics like religion, showing the model's effectiveness at identifying topic-specific keywords. These terms are critical due to their strong association with the "alt.atheism" and similar classes.

The dataset's imbalance ratio was 1.59, below the threshold for considering it unbalanced, which justified the use of accuracy instead of balanced accuracy.

IV. Dataset Size

As the dataset size increased, so did the model's accuracy, but performance improvements diminished beyond a certain point. This suggests that after a certain amount of training data, the model had captured the dataset's main patterns. Further increasing the dataset size showed diminishing returns, which is typical for balanced datasets, indicating that expanding the dataset may not be cost-effective for the business use case.

V. Topic Analysis

Using *Latent Dirichlet Allocation (LDA)*, we assigned documents to 10 topics and trained individual classifiers for each topic. However, the two-layered classification approach resulted in low accuracy (0.1936) and balanced accuracy (0.1947). This poor performance likely stems from overlapping vocabularies between topics, making it difficult for the model to distinguish between categories like "soc.religion.christian" and "talk.politics.misc".

The topic-based classification strategy, while conceptually promising, did not significantly enhance performance in this case, primarily due to shared terminology across topics and the short length of some documents.

References

[1] "A Comparison of SVM against Pre-Trained Language Models for Text Classification Tasks", *Papers with Code*, 2021.