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.

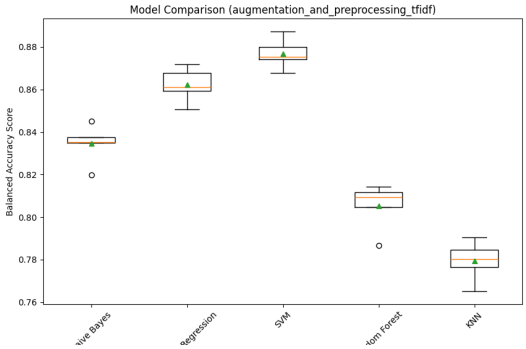

Fig. 1. Model comparison using TF-IDF vectorizer. Support Vector Machine (SVM) performs best with mean accuracy of 0.8777.

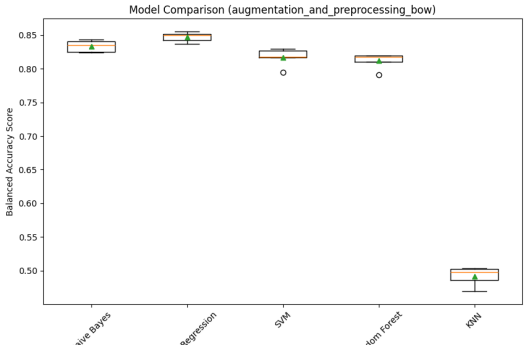

Fig. 2. Model comparison using Bag of Words (BoW). Logistic Regression performs best with mean accuracy of 0.8474.

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.

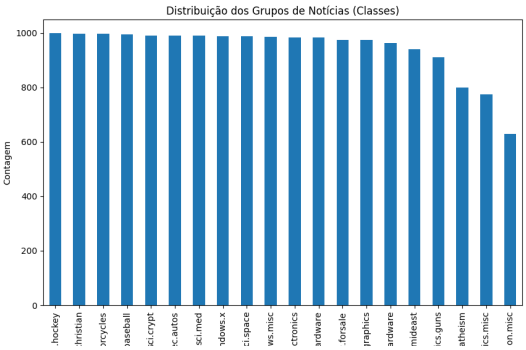

Fig. 3. Distribution of documents across different newsgroups. The dataset contains 20 different categories with varying document counts.

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.

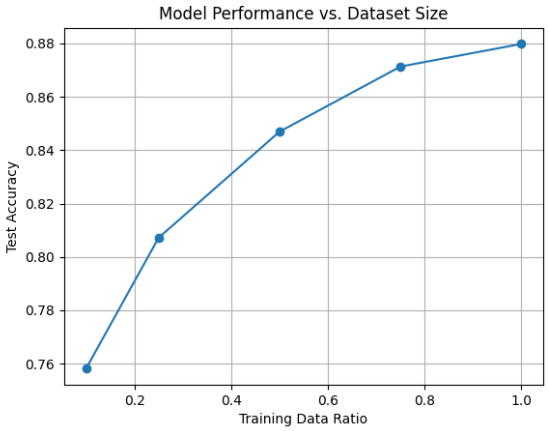

Fig. 4. Comparison of model performance vs dataset size. The accuracy improves with dataset size but shows diminishing returns beyond a certain point.

V. Topic Analysis

To gain deeper insights into the dataset, a Latent Dirichlet Allocation (LDA) topic modeling approach was applied. This method allows us to explore whether categorizing documents into broader topics before classification could improve performance.

A. Topic-Based Classifier Implementation

- *Topic Assignment*: Using a TF-IDF vectorizer, the text was transformed, and LDA was employed to assign each document to one of the 10 predefined topics.
- *Topic-Specific Classifiers*: For each topic, a Logistic Regression classifier was trained on the documents belonging to that specific topic. This allows the classifier to specialize in distinguishing between classes within the context of each topic.
- *Classification Process*: For the test data, documents were first assigned to topics using the LDA model, and then a topic-specific classifier was used for final classification.

B. Results

The classification accuracy achieved with the topic-based approach was 0.8610, an improvement over the baseline model. This suggests that topic-specific classifiers can be more effective in distinguishing between closely related categories within a topic.

The results show that the classifier performed particularly well in topics related to technology, religion, and sports, with

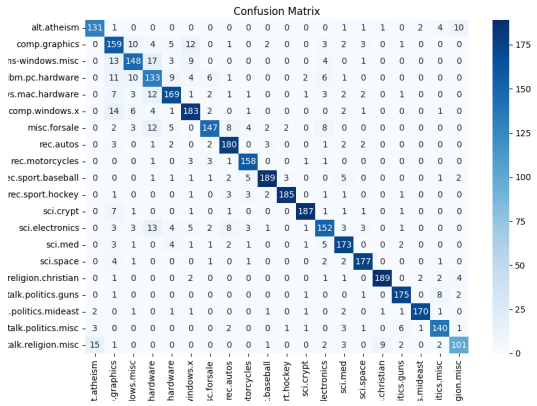

Fig. 5. Classification Report for the topic-based classification approach. Overall accuracy is 0.8610 with performance varying across topics.

precision and recall scores both high for these categories (e.g., precision of 0.94 for topic 10). However, some topics such as those related to political discussions exhibited lower recall due to overlap in vocabulary across related topics (e.g., topic 1 with a recall of 0.79).

The topic-based classification strategy demonstrates that errors are not uniformly distributed across all topics. Certain topics showed higher classification effectiveness, and the two-layer classifier combining topic modeling with classification resulted in improved performance compared to a single-layer classifier.

References

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