Practical Machine Learning
Project 2: Unsupervised Learning

Dilirici Mihai
Group 411

Contents

1 Introduction

This document provides an overview of the clustering and classification project, which aims to group comments into predefined topics: Biology, Chemistry, and Physics. The project approaches two unsupervised clustering algorithms, **K-Medoids** and **Latent Dirichlet Allocation (LDA)**, to identify patterns in the data and evaluate their effectiveness. Additionally, the results from these unsupervised models are compared against a **supervised classification baseline** and a **random baseline**.

The methodology involves several key steps, including dataset preparation, vectorization of text data, clustering using K-Medoids and LDA, and comprehensive evaluation of clustering and classification performance using metrics such as Silhouette Score, Adjusted Rand Index (ARI), F1 Score, and Accuracy. This multi-faceted approach aims to provide a thorough understanding of how well unsupervised models perform relative to supervised and random baselines in clustering textual data into meaningful groups.

2 Dataset

The dataset consists of comments labeled with one of three topics: Biology, Chemistry, or Physics. The dataset is divided into training, validation, and test sets. The training dataset is used to train the clustering model, the validation dataset is used to tune hyperparameters, and the test dataset evaluates the final model.

2.1 Training Dataset Distribution

The training dataset contains the following distribution of topics:

- Biology: *3500 comments*
- Chemistry: *3000 comments*
- Physics: *2500 comments*

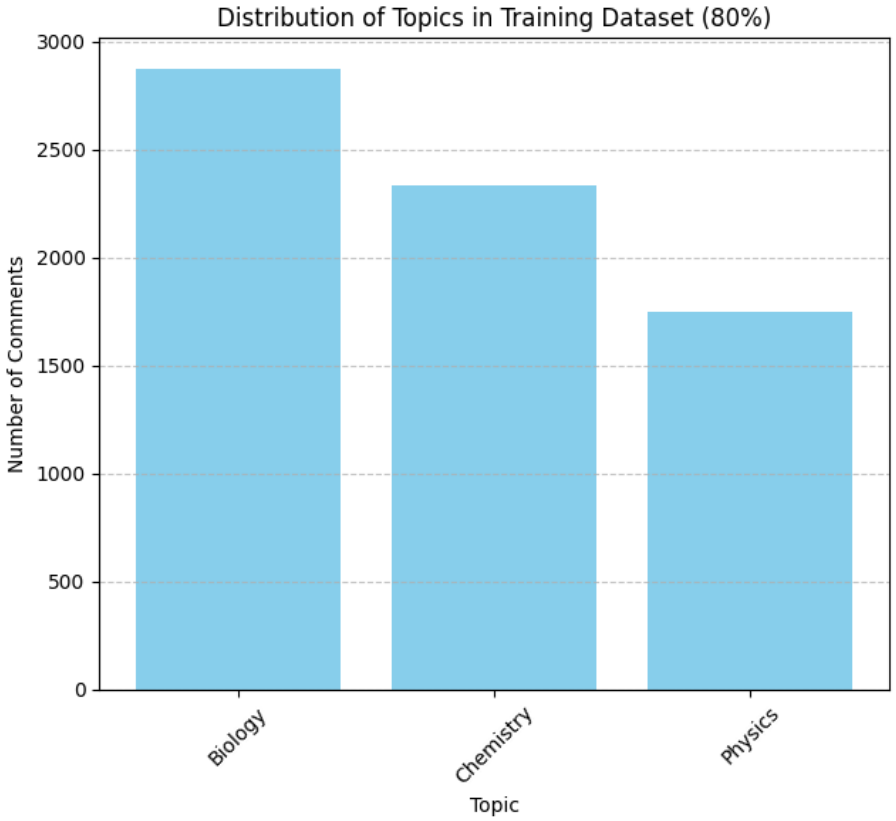

Figure 1: Training Data Distribution

2.2 Validation Dataset Distribution

The validation dataset, obtained by splitting the original training dataset, contains the following distribution of topics:

- Biology: *700 comments*
- Chemistry: *600 comments*
- Physics: *500 comments*

2.3 Test Dataset Distribution

The test dataset contains an equal number of comments for each topic to ensure balanced evaluation. The topic distribution is as follows:

- Biology: *500 comments*
- Chemistry: *500 comments*
- Physics: *500 comments*

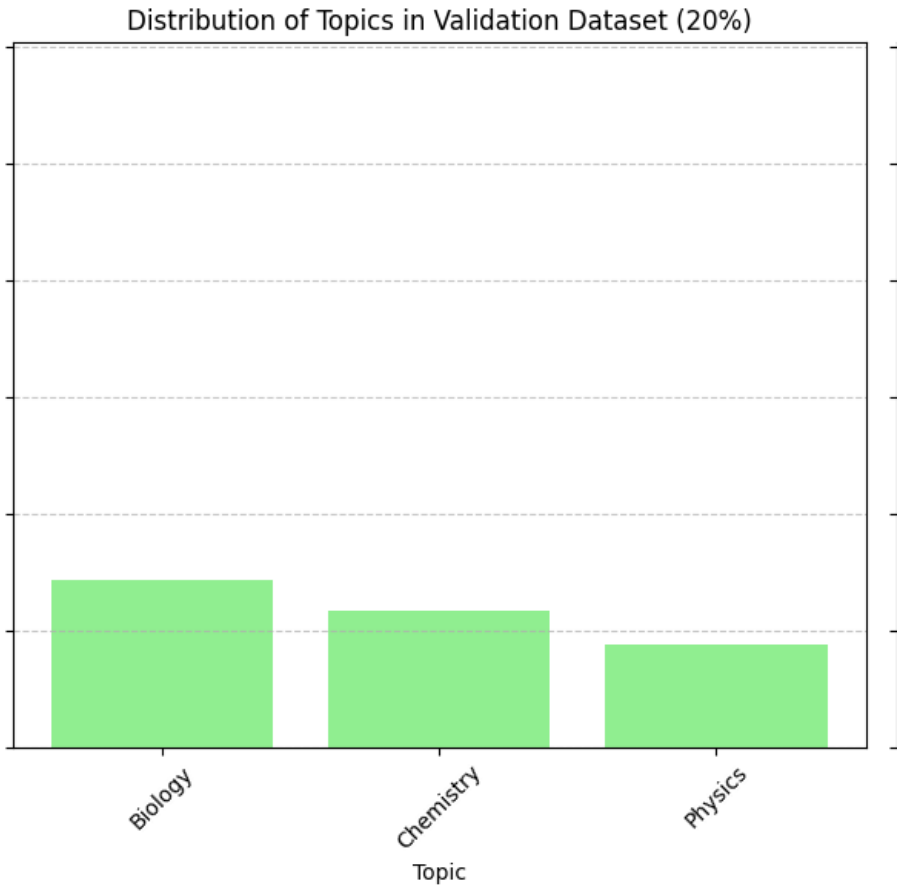

Figure 2: Validation Data Distribution

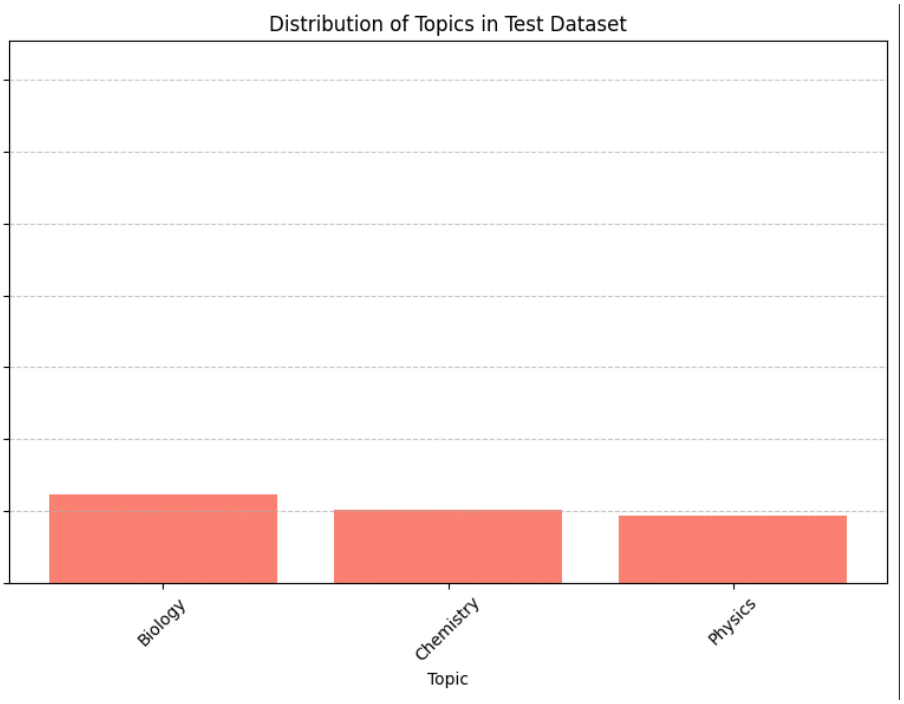

Figure 3: Test Data Distribution

2.4 Text Length Distribution

An analysis of the text length distribution in the dataset reveals significant variation in the lengths of the comments. Most comments are relatively short, with a few outliers having a large number of characters. This information is useful for preprocessing and understanding the nature of the dataset.

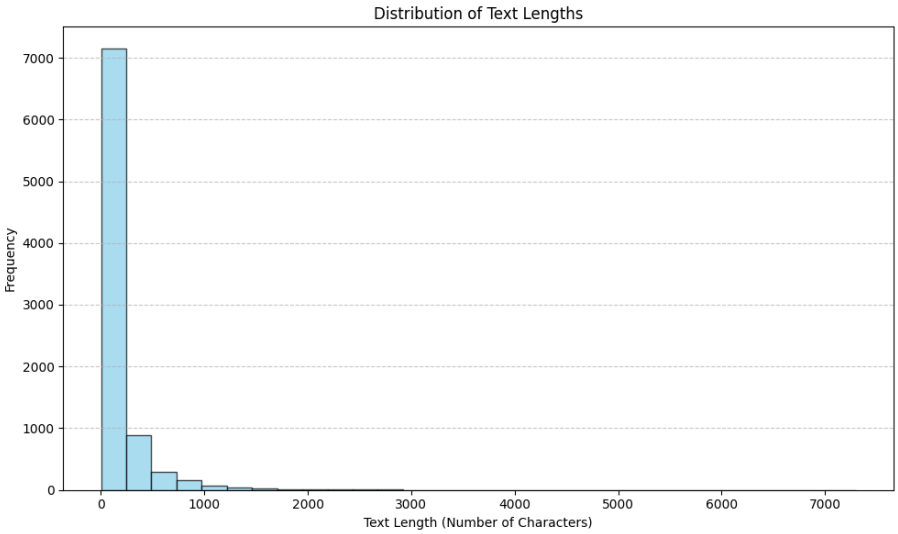

Figure 4: Distribution of Text Lengths

2.5 Word Clouds

To better understand the most frequent terms in the dataset, word clouds were generated for the training, validation, and test datasets. These visualizations provide insights into the dominant words and their relative frequencies across different datasets.

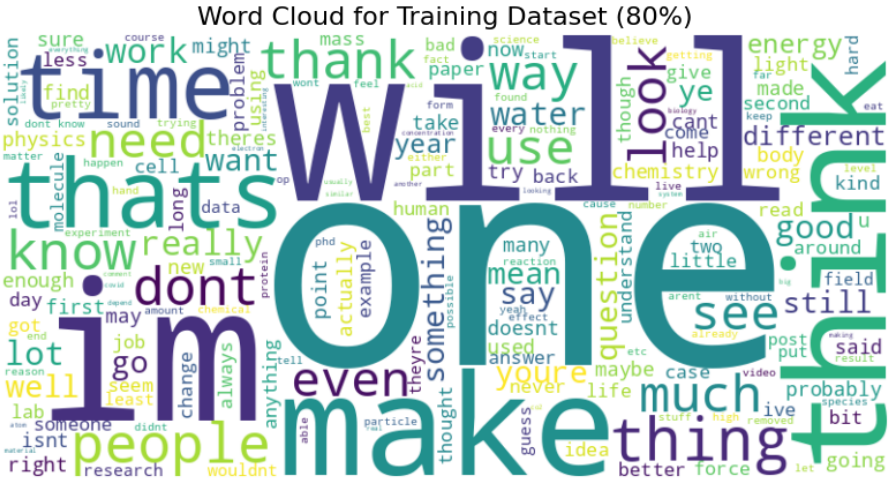

Figure 5: Word Cloud for Training Dataset

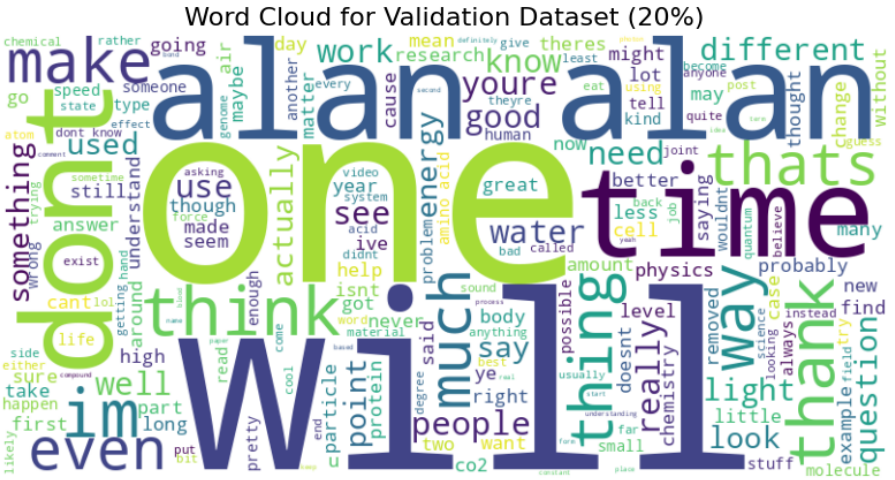

Figure 6: Word Cloud for Validation Dataset

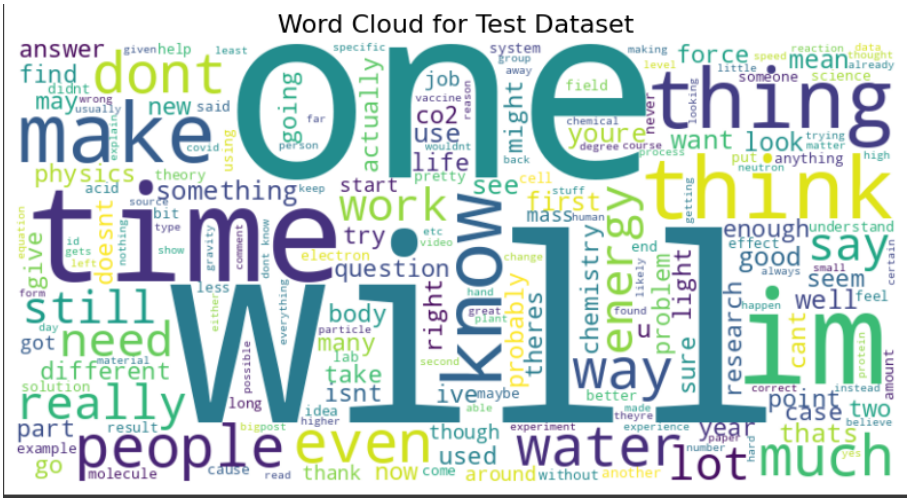

Figure 7: Word Cloud for Test Dataset

3 Methodology

3.1 Preprocessing

Each comment in the dataset undergoes preprocessing, which includes:

- Converting text to lowercase.
- Removing punctuation and special characters.
- Tokenizing and vectorizing the text using TF-IDF, Word2Vec or CountVectorizer.

3.2 Clustering and Classification Algorithms

This project utilizes multiple approaches to cluster and classify comments into predefined topics.

- **K-Medoids:** The K-Medoids algorithm is employed due to its robustness to noise and outliers. It minimizes the sum of dissimilarities between points and their respective cluster medoids. A grid search is conducted to determine the optimal number of clusters and feature extraction parameters, such as the number of features in the TF-IDF vectorizer.
- **Latent Dirichlet Allocation (LDA):** LDA is used for clustering through topic modeling. It identifies latent topics in the comments and groups them accordingly. The input data for LDA is prepared using the **CountVectorizer**, which converts the text into a bag-of-words representation. Similar to K-Medoids, a grid search is performed to tune hyperparameters such as the number of components (topics), learning decay, and the number of iterations.
- **Supervised Classification Baseline:** A Random Forest classifier is used as the supervised baseline to predict the predefined topics directly. This approach provides a comparison point for the unsupervised methods.
- **Random Baseline:** A random classifier is included as a baseline to evaluate the clustering and classification results against chance performance.

4 Evaluation

The clustering results are evaluated using the following metrics:

- **Silhouette Score**: Measures the quality of clusters by comparing intra-cluster cohesion and inter-cluster separation.
- **Adjusted Rand Index (ARI)**: Measures the similarity between the predicted clusters and ground truth labels.
- **F1 Score and Accuracy**: Evaluated by comparing clusters with ground truth labels on validation and test datasets.

5 Results

The next sections will present detailed results from the clustering experiments and visualizations for training, validation, and test datasets.

5.1 K-Medoids Clustering

The clustering process involves several key steps, outlined as follows:

1. **Dataset Splitting**: The training dataset is split into training and validation subsets using an 80-20 ratio. The test dataset remains separate for final evaluation.
2. **Text Vectorization**: The comments are transformed into numerical representations using TF-IDF vectorization. A grid search is performed to find the optimal maximum feature size for the vectorizer.

3. **Dimensionality Reduction**: PCA is applied to reduce the dimensionality of the vectorized data to 2D for visualization purposes.

4. **Clustering with K-Medoids**: The K-Medoids algorithm is applied to cluster the data. The number of clusters is tuned using grid search, and the optimal configuration is determined based on the Silhouette Score.

5. **Evaluation**:

   - Validation and test datasets are used to evaluate the clustering model.
   - Metrics such as **Silhouette Score** and **Adjusted Rand Index (ARI)** are computed to assess cluster quality and alignment with ground truth labels.

6. **Visualization**: Clusters are visualized for training, validation, and test datasets using 2D PCA projections to provide an intuitive understanding of the clustering process.

7. **Results Saving**: Cluster assignments are saved for training, validation, and test datasets in separate CSV files for further analysis.

5.2 Grid Search Results

The results of the grid search conducted to determine the optimal number of clusters and the maximum number of features for the TF-IDF vectorizer are visualized in the heatmap below. The heatmap represents the Silhouette Scores for different combinations of parameters, with darker shades indicating higher scores.

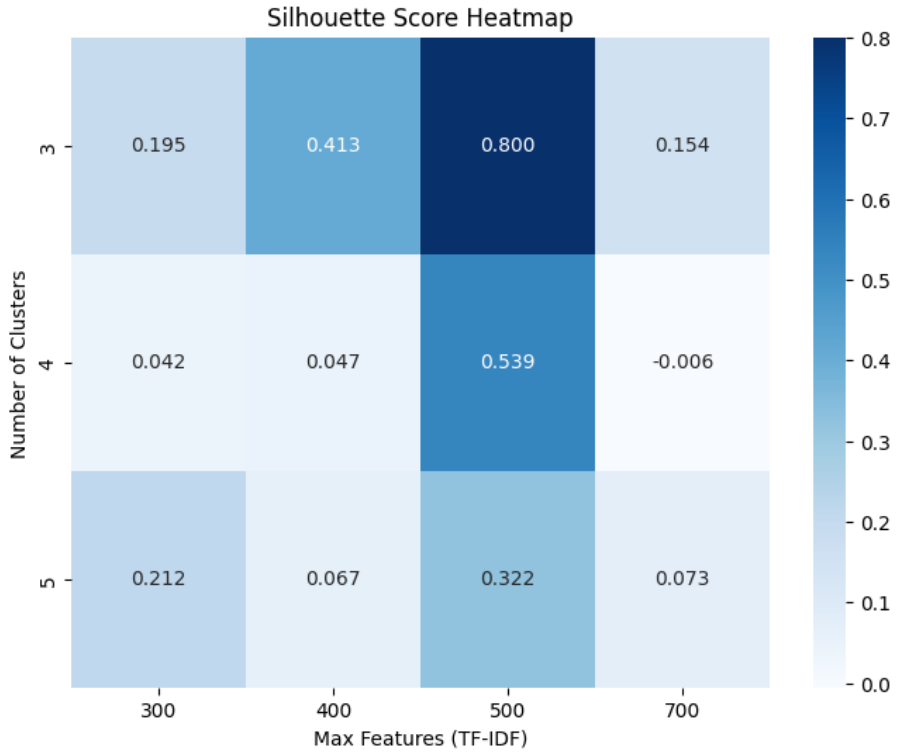

Figure 8: Silhouette Score Heatmap for Grid Search

From the heatmap, we can observe that the best parameters obtained from the grid search are:

- **Number of Clusters: 3**
- **Max Features: 500**
- **Best Silhouette Score: 0.8002**

5.3 Validation and Test Results

Using the best parameters, the clustering performance was evaluated on the validation and test datasets. The following table compares the results of K-Medoids with TF-IDF and Word2Vec vectorization methods:

Metric	TF-IDF	Word2Vec
Validation Silhouette Score	0.7904	0.4909
Validation ARI	-0.0083	0.0037
Test Silhouette Score	0.7286	0.4621
Test ARI	-0.0008	-0.0027

Table 1: Comparison of K-Medoids Clustering Performance with TF-IDF and Word2Vec

The Silhouette Score indicates how well-separated the clusters are, with higher values being better. However, the negative Adjusted Rand Index (ARI) for both validation and test datasets suggests limited alignment between the predicted clusters and the ground truth labels. This may reflect the unsupervised nature of clustering, which does not aim to align perfectly with the original labeled topics.

The TF-IDF method showed significant increase in the silhouette score, so the next plots and results were calculated only for the first model.

5.4 Clustering Visualization

To further analyze the clustering performance, we projected the high-dimensional TF-IDF features into a two-dimensional space using Principal Component Analysis (PCA). This allowed us to visualize the clustering results for the training, validation, and test datasets.

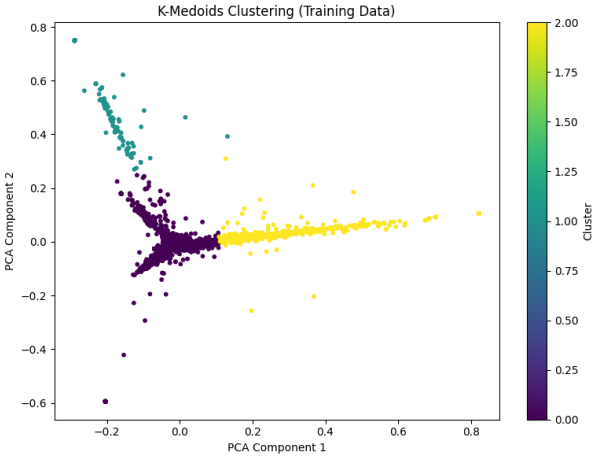

Figure 9: K-Medoids Clustering Results on Training Data

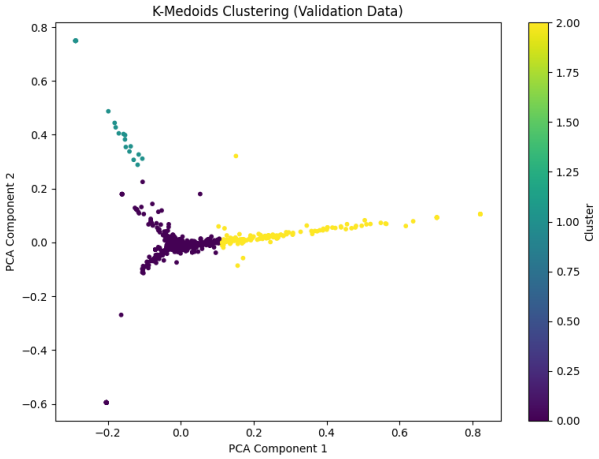

Figure 10: K-Medoids Clustering Results on Validation Data

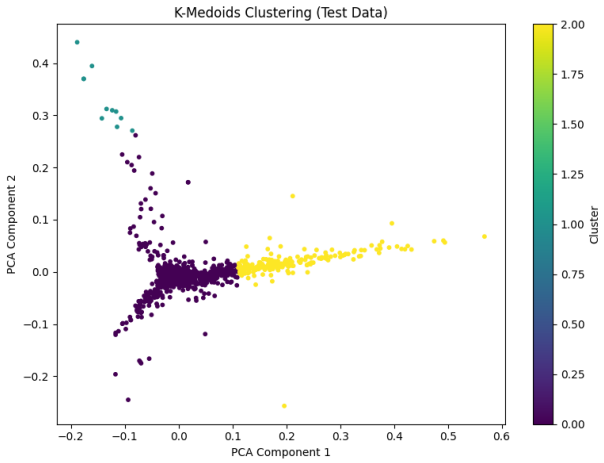

Figure 11: K-Medoids Clustering Results on Test Data

In these visualizations:

- **Training Data:** The clusters formed show the initial separations achieved during the model training phase. The visualization indicates that the training data points are distinctly grouped into three main clusters corresponding to the number of clusters chosen during the grid search.
- **Validation Data:** This plot highlights the clustering of unseen validation data, demonstrating the model's ability to generalize its learned patterns. Although the clusters are somewhat well-separated, minor overlaps are present, suggesting potential for further optimization.
- **Test Data:** The final clustering visualization on the test dataset validates the model's performance on completely unseen data. The clusters remain consistent with the patterns observed in the training and validation phases.

These plots visually confirm the clustering behavior across different data splits and emphasize the K-Medoids model's consistency and robustness in separating comments into predefined clusters.

5.5 Evaluation Metrics: F1 Score and Accuracy

To further evaluate the clustering results, we calculated the following metrics on both the validation and test datasets:

- **F1 Score (Weighted)**: This metric accounts for both precision and recall, providing a harmonic mean of the two. The weighted F1 score considers the imbalance in class distribution, ensuring that all topics (clusters) are represented proportionally.
- **Accuracy**: This metric measures the overall alignment between the predicted clusters and the ground truth labels. It provides a straightforward interpretation of how well the clustering matches the labeled topics.

Implementation: To ensure the compatibility of the labels for calculation, a preprocessing step was performed to convert string labels into numeric representations. This ensures that the metrics can be computed without errors, especially when the ground truth labels are categorical strings.

Validation Dataset Results:

- F1 Score (Validation): *0.2581*
- Accuracy (Validation): *38.41%*

Test Dataset Results:

- F1 Score (Test): *0.2748*
- Accuracy (Test): *38.02%*

Comparison with Supervised Models: These metrics will serve as a benchmark for comparing the unsupervised clustering approach (K-Medoids) with a supervised model. The goal is to assess how well the clustering aligns with labeled topics, providing insights into the trade-offs between supervised and unsupervised methods in this task.

5.6 Latent Dirichlet Allocation (LDA) Clustering

The second clustering approach utilized in this project is **Latent Dirichlet Allocation (LDA)**. LDA is a topic modeling algorithm that assumes each document is a mixture of topics, and each topic is a mixture of words. It is particularly well-suited for unsupervised text classification.

Implementation Details:

- Two LDA models were implemented, differing in the feature extraction method:
  - **Model 1: Using CountVectorizer:** The text data was vectorized using **CountVectorizer** with a maximum of 500 features. The performance metrics are as follows:
    - **Validation F1 Score:** 15.41%
    - **Validation Accuracy:** 12.13%
    - **Test F1 Score:** 54.87%
    - **Test Accuracy:** 54.54%
  - **Model 2: Using TF-IDF Vectorizer:** The text data was vectorized using **TF-IDF** with a maximum of 500 features. The performance metrics are as follows:
    - **Validation F1 Score:** 28.68%
    - **Validation Accuracy:** 29.84%
    - **Test F1 Score:** 30.24%
    - **Test Accuracy:** 28.31%

Max Features	Number of Topics	Learning Decay	Max Iterations	Silhouette Score	ARI
300	3	0.3	10	0.5220	0.0123
300	3	0.3	20	0.5113	0.0116
300	3	0.3	30	0.5018	0.0124
300	5	0.3	10	0.4899	0.0079
300	5	0.3	20	0.4738	0.0073
300	5	0.3	30	0.4714	0.0055
300	7	0.3	10	0.4546	0.0083
300	7	0.3	20	0.4424	0.0072
300	7	0.3	30	0.4345	0.0077
500	3	0.3	10	0.5319	0.0498
500	**3**	**0.3**	**20**	**0.5275**	**0.0499**
500	3	0.3	30	0.5216	0.0465
500	5	0.3	10	0.4801	0.0249
500	5	0.3	20	0.4507	0.0346
500	5	0.3	30	0.4480	0.0406
500	7	0.3	10	0.4529	0.0218
500	7	0.3	20	0.4327	0.0272
500	7	0.3	30	0.4184	0.0257

Table 2: Results of Custom Grid Search for LDA Clustering on Validation Dataset

The best parameters selected from this grid search were:

- **Max Features: 500**
- **Number of Topics: 3**
- **Learning Decay: 0.3**
- **Max Iterations: 20**

Top Words for Topics: The top words for each topic learned by the LDA model are shown below:

- **Topic 0:** *like, people, chemistry, good, things, try, really, just, way, theres, ive, need, want, physics, chemical, lot, thing, wouldnt, liquid, youre*
- **Topic 1:** *think, dont, water, distance, thats, light, means, know, just, use, acid, speed, used, thought, sodium, gas, right, concentration, answer, said*
- **Topic 2:** *time, im, make, thank, makes, different, thanks, question, look, say, mean, isnt, species, job, energy, dont, reaction, cells, got, sure*

5.7 LDA Cluster Visualization

The clustering results of the LDA model on the validation and test datasets are visualized using PCA (Principal Component Analysis). The two-dimensional scatter plots below represent the reduced dimensions of the topic distributions, with each point indicating a comment assigned to a cluster. Different colors represent different clusters.

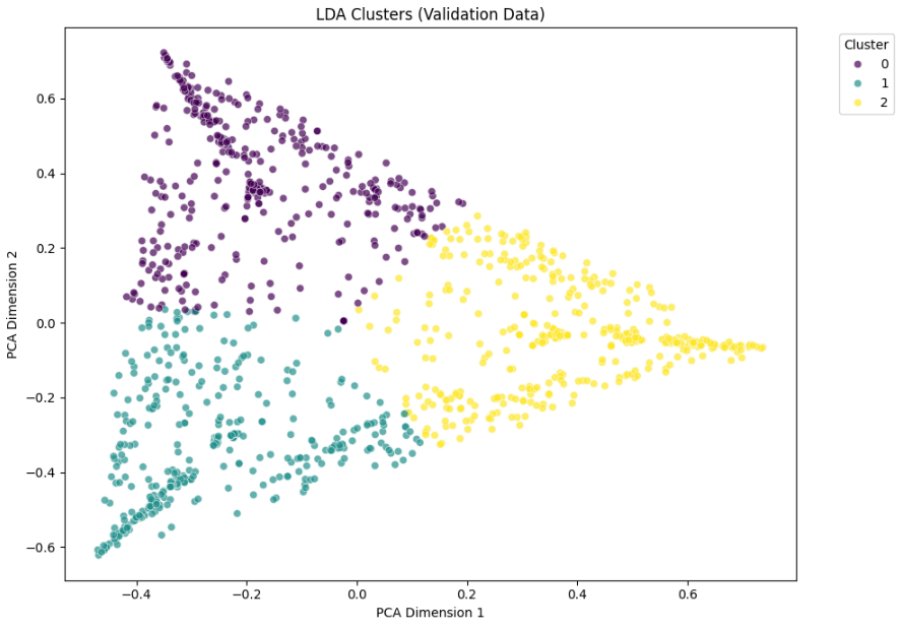

Figure 12: LDA Clusters on Validation Data

In the validation dataset (Figure 12), the clusters are relatively distinct, suggesting that the LDA model effectively identifies topics. However, there is some overlap between clusters, indicating potential ambiguities in topic separation.

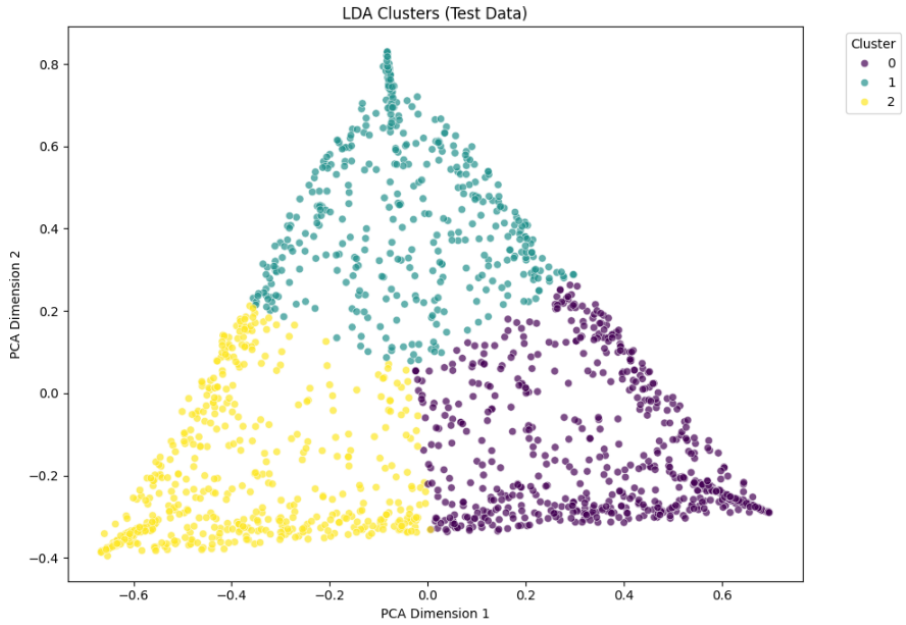

Figure 13: LDA Clusters on Test Data

In the test dataset (Figure 13), the clusters are similarly distributed, reflecting the consistency of the LDA model across unseen data. The triangular structure highlights how the three topics are separated, though some points remain close to cluster boundaries, showing potential for refinement.

5.8 Interpretation of the Results

The clusters generated by the LDA model represent groups of comments that share similar topics or semantic contexts. Each cluster corresponds to a latent topic derived from the model. Below is an interpretation of the results:

- **What the Clusters Represent:**
  - **Cluster 0:** Based on the most frequent terms in this cluster (e.g., "like," "people," "chemistry," "liquid"), this group likely represents general discussions or exploratory statements related to Chemistry.
  - **Cluster 1:** Common terms such as "think," "distance," "water," and "speed" suggest that this cluster focuses on physics-related comments, including conceptual or theoretical discussions.
  - **Cluster 2:** With terms like "time," "energy," "reaction," and "cells," this cluster likely represents topics related to Biology and chemical reactions, emphasizing biological processes and energy-related discussions.
- **Silhouette and ARI Scores:**
  - The validation and test Silhouette Scores (0.522 and 0.500, respectively, for advanced preprocessing) indicate moderate cohesion and separation of clusters.
  - Adjusted Rand Index (ARI) scores (0.0123 for validation and 0.0522 for test) suggest limited agreement with the ground truth labels, highlighting that the

model clusters based on latent topics that might not align perfectly with predefined labels.

- **Challenges and Overlaps:**
  - Some overlap is observed between clusters, particularly between topics with shared terminology or ambiguous contexts (e.g., physical chemistry or energy-related concepts).
  - Ambiguity in natural language, especially in multi-disciplinary datasets, contributes to lower ARI scores, as comments may not strictly adhere to a single topic.

Conclusion: The LDA model provides meaningful clusters representing latent topics in the dataset. However, due to overlaps and the nature of unsupervised clustering, further refinement or comparison with supervised models may be required to achieve higher alignment with ground truth labels.

5.9 Baseline Comparisons: Random and Supervised Models

To contextualize the clustering performance of the unsupervised models, we compare them to two baselines: a random baseline and a supervised model baseline. These baselines help in assessing the added value of clustering techniques relative to standard classification approaches.

5.9.1 Random Baseline

The random baseline serves as a lower bound for the evaluation metrics. A dummy classifier is trained to predict random classes based on a uniform distribution. The TF-IDF-transformed features are used to ensure consistency with other models.

- **Validation Accuracy:** 33.18%
- **Test Accuracy:** 33.35%
- **Validation F1 Score:** 33.46%
- **Test F1 Score:** 33.45%

The performance metrics are close to the expected value for random guessing in a dataset with three equally likely classes (accuracy of approximately $\frac{1}{3}$).

5.9.2 Supervised Baseline

The supervised baseline uses a Random Forest Classifier trained on the TF-IDF-transformed training dataset. The model predictions on the validation and test sets were evaluated using accuracy and F1 scores.

- **Validation Accuracy:** 60.61%
- **Test Accuracy:** 69.74%
- **Validation F1 Score:** 59.89%

- **Test F1 Score:** 69.74%

The supervised baseline significantly outperforms the random baseline, as expected, achieving high accuracy and F1 scores. This demonstrates the potential of leveraging labeled data in classification tasks, where the goal is to maximize alignment with ground truth labels.

Metric	**Random**	**Supervised**	**K-Medoids**	**LDA**
Test Accuracy	33.35%	69.74%	35.30%	54.54%
Test F1 Score	33.45%	69.74%	33.09%	54.87%

Table 3: Performance Comparison of All Models on Test Dataset

6 Conclusion

The random baseline provides a minimal level of performance, with both accuracy and F1 scores close to 33%, which aligns with the random chance for three classes. In contrast, the supervised baseline significantly outperforms both the random baseline and unsupervised clustering methods (K-Medoids and LDA). This demonstrates the strength of leveraging labeled data in achieving high alignment with ground truth labels.

While the supervised model shows superior performance, it relies on labeled data, which may not always be available. In contrast, unsupervised methods like K-Medoids and LDA can extract meaningful insights in the absence of labeled data, making them valuable for exploratory analysis or when labeled datasets are scarce.