Assignment 4

Name: Mohamed-Obay Alshaer
Student Number: 300170489

Submission Date: March 21, 2025

Professor: Daniel Shpairo

SEG4300 - Applied Machine Learning

1 Clustering and Supervised Prediction on AG News Dataset

1.1 Dataset Selection

The AG News dataset was selected for this clustering analysis. This dataset contains 120,000 training and 7,600 test samples of news articles categorized into four classes: World, Sports, Business, and Science/Technology. Each sample includes a news title, content, and class label. This dataset is ideal for clustering due to its natural topic structure, text data complexity, adequate size, real-world relevance, and built-in validation potential through predefined categories.

1.2 Data Preprocessing

Text data preprocessing involved:

- Text cleaning: removing special characters and converting to lowercase
- Tokenization: splitting text into individual words
- Stopword removal: eliminating common words with limited semantic value
- Vectorization: converting text to numerical features using TF-IDF
- Dimensionality reduction: applying techniques for visualization and more efficient clustering

1.3 Clustering Analysis

The Elbow Method was used to determine the optimal number of clusters, which indicated 4 clusters would be appropriate. K-means clustering was then applied to group the articles. The resulting clusters aligned remarkably well with natural news categories:

- **Cluster 0 - Business/Economy**: Characterized by terms like "market," "company," "business," and "shares"
- **Cluster 1 - Sports**: Defined by keywords such as "team," "game," "player," and "season"
- **Cluster 2 - Technology/Science**: Identified by terms including "technology," "internet," "computer," and "software"
- **Cluster 3 - World News/Politics**: Marked by words like "government," "president," "country," and "war"

This unsupervised discovery of patterns closely matching the original four categories demonstrates the effectiveness of K-means in identifying natural groupings in text data.

1.4 Supervised Prediction of Clusters

A Logistic Regression model was trained to predict cluster membership based on the same features used for clustering. Performance metrics included:

- **Accuracy**: Approximately 90% correct cluster predictions

- **Confusion Matrix**: Revealed some confusion between certain clusters, likely due to topic overlap
- **Classification Report**: Showed balanced performance across clusters

The high accuracy indicates well-defined, separable clusters in the feature space.

1.5 Findings and Applications

Key Insights:

- Unsupervised clustering effectively discovered meaningful categories without prior knowledge
- The developed pipeline (preprocessing → vectorization → clustering → prediction) provides a robust approach for analyzing text data
- High alignment between unsupervised clusters and predefined categories validates both our approach and the original dataset categorization

Potential Applications:

- Content organization: Automatically categorizing news articles and documents
- Recommendation systems: Suggesting similar content based on cluster membership
- Topic discovery: Identifying emerging trends in large text corpora
- Document retrieval: Enhancing search functionality through content grouping

1.6 Limitations and Future Improvements

While the current approach was successful, several enhancements could be explored:

- Advanced feature engineering using modern embedding techniques (e.g., BERT)
- Alternative clustering methods such as density-based or hierarchical clustering
- Dynamic topic modeling to handle evolving topics over time
- Improved cluster interpretation methods for automatic labeling

1.7 Conclusion

This project demonstrates the power of combining unsupervised and supervised learning techniques to extract meaningful insights from unstructured text data. The successful discovery of natural news categories through clustering highlights the value of unsupervised learning in revealing intrinsic data patterns, while the high prediction accuracy of the supervised model confirms the robustness of these discovered patterns.