Assignment 4: Clustering with a HuggingFace Dataset

Dataset Selection

For this assignment, the dataset gxb912/large-twitter-tweets-sentiment from the Hugging Face Datasets Library was chosen. This dataset is a collection of tweets formatted in a tabular data structure, annotated for sentiment analysis. Each tweet is associated with a sentiment label, where 1 indicates a Positive sentiment and 0 indicates a Negative sentiment. The dataset is in English.

Dataset Structure:

- text: A string containing the tweet's content.
- sentiment: An integer where 1 indicates Positive sentiment and 0 indicates Negative sentiment.

The dataset was preprocessed to remove null values and short texts, ensuring data consistency.

Data Preprocessing

The preprocessing steps included:

- Converting text to lowercase.
- Removing punctuation and non-ASCII characters.
- Vectorizing the text data using the CountVectorizer and transforming it into a TF-IDF representation.
- Applying Truncated Singular Value Decomposition (SVD) to reduce dimensionality for visualization and efficiency.

Clustering Analysis

The K-Means clustering algorithm was used to group the tweets. The elbow method was applied to determine the optimal number of clusters, which was found to be five (5) based on the inertia plot, as shown in elbowGraphKMeans.png.

The clustered data was then visualized in a 2D space using the reduced dimensions, as depicted in clustersKMeans.png. The visualization helped in understanding the distribution of clusters and their separation. However, there was some overlap, indicating potential variations in sentiment structure within the dataset.

Interpretation of Clusters

As shown in clustersKMeanInterpratation.png, the clustering results varied in effectiveness. Some clusters formed meaningful categories, such as Cluster 4, which predominantly represented positive sentiment, and Clusters 2 and 3, which captured elements of mostly positive and some negative sentiment. However, Clusters 1 and 0 did not align well with the

sentiment labels, suggesting that the K-Means algorithm might be capturing different linguistic patterns rather than purely sentiment-based groupings. This highlights the possibility that the clustering model detected other textual features beyond just sentiment, which could be further explored through additional feature engineering.

Cluster Prediction using Supervised Learning

A supervised learning model was trained to predict the cluster labels assigned by K-Means. The text data was re-vectorized using TF-IDF and further reduced in dimensionality. A Random Forest Classifier was trained on this data. The classification report showed a high accuracy of 96.3%, with strong precision, recall, and F1-scores across all cluster labels, as presented in classificationReport.png.

Key Insights and Conclusion

- The clustering approach successfully identified natural groupings in the data, but not all clusters directly corresponded to sentiment labels.
- The elbow method was effective in selecting the optimal number of clusters.
- The supervised model performed exceptionally well in predicting cluster labels, suggesting that the clusters had distinguishable characteristics.
- Some overlap in sentiment categories within clusters indicates that further feature engineering or different clustering approaches might refine the results.
- The results suggest that K-Means may have captured patterns beyond sentiment, potentially including aspects like tweet structure or topic similarity.

Overall, the experiment demonstrated the effectiveness of K-Means clustering for text classification and the ability of supervised learning to enhance interpretability of unsupervised results.