Sentiment Analysis of Tweet Texts using Clustering Algorithms

1. Introduction

In this section I perform Sentiment Analysis on tweet texts using unsupervised clustering algorithms. The main objective is to group similar sentiments without predefined labels by using three clustering algorithms: K-Means, Agglomerative Clustering, and DBSCAN. Two tokenization methods, Bag of Words (BoW) and TF-IDF, are used for feature extraction.

2. Basic Definitions

Bag of Words (BoW)

A simple method of converting text into numerical features by counting the occurrence of each word.

TF-IDF (Term Frequency - Inverse Document Frequency)

A statistical measure to evaluate how important a word is to a document relative to a collection of documents.

K-Means Clustering

An unsupervised algorithm that groups data into K clusters based on feature similarity.

Agglomerative Clustering

A type of hierarchical clustering that builds clusters by merging pairs of data points.

DBSCAN

A clustering algorithm that groups data points based on their density, identifying outliers as noise.

PCA (Principal Component Analysis)

A dimensionality reduction technique used to visualize high-dimensional data in 2D or 3D space.

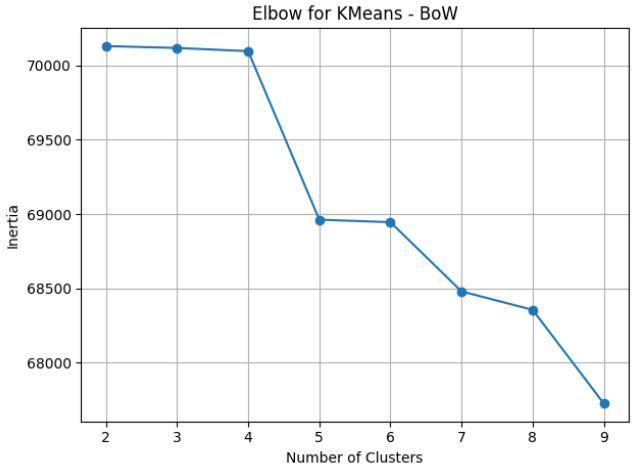
3. Methodology

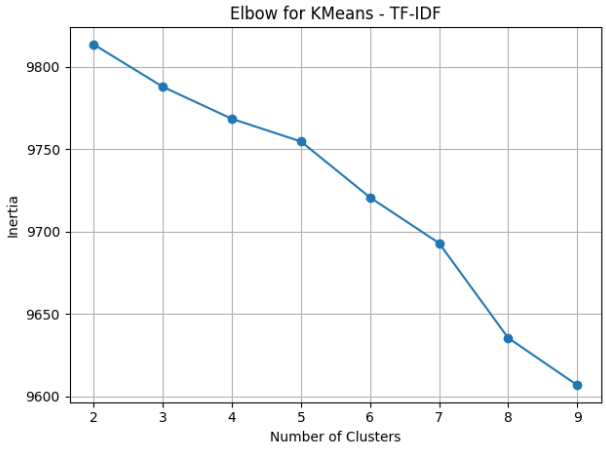
The tweet dataset was preprocessed and vectorized using two tokenization techniques: BoW and TF-IDF. Three clustering algorithms (K-Means, Agglomerative Clustering, and DBSCAN) were applied. For K-Means, the optimal number of clusters was determined using the elbow method. Confusion matrices were generated for cluster evaluation and PCA was used for 2D visualization of clustered data.

4. Results and Visualizations

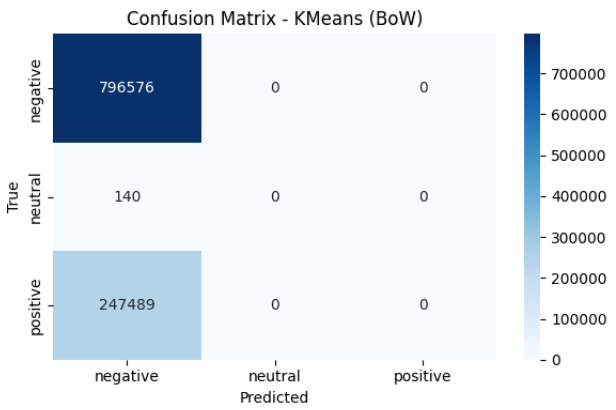
This section includes elbow curves, confusion matrices, and PCA-based scatter plots for each algorithm and tokenization method.

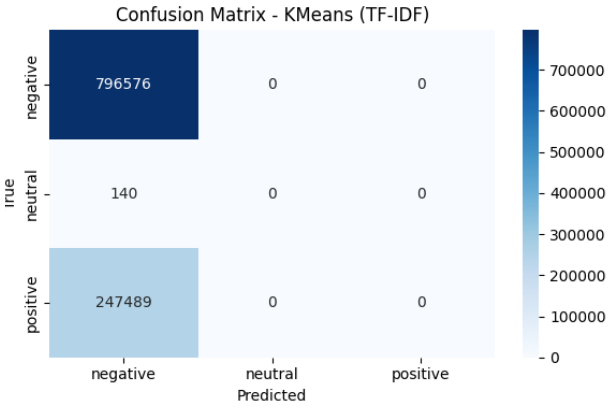
4.1 Elbow Curve for K-Means (for both BoW & TF-IDF)

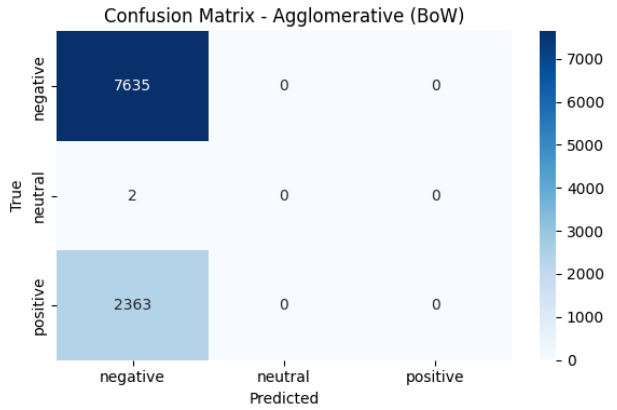


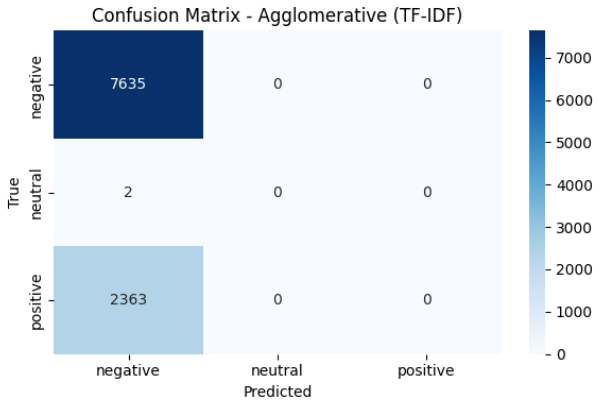


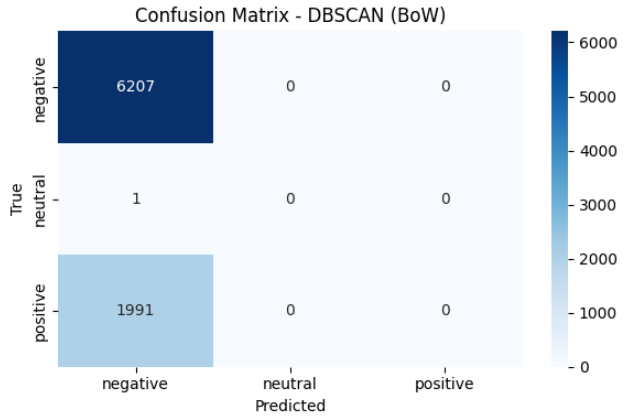
4.2 Confusion Matrices

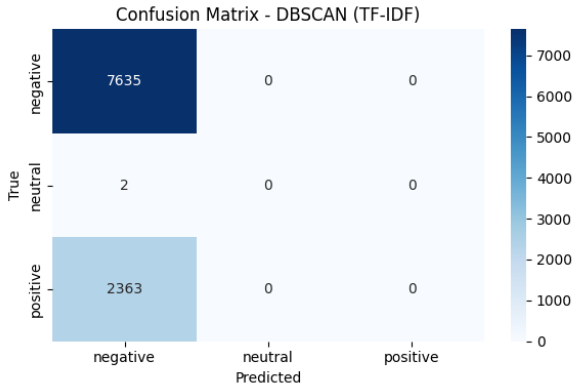




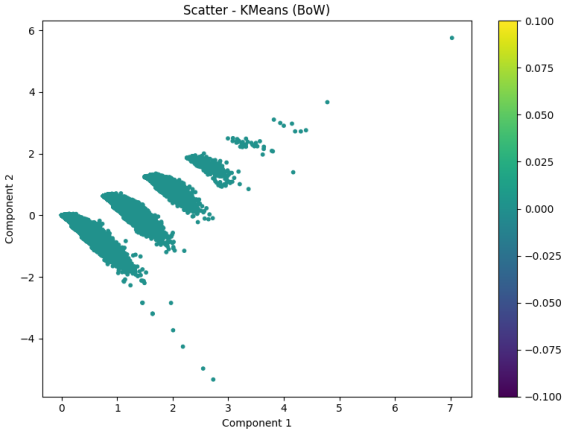


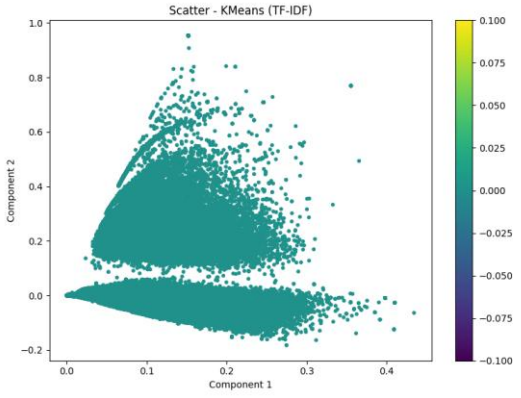


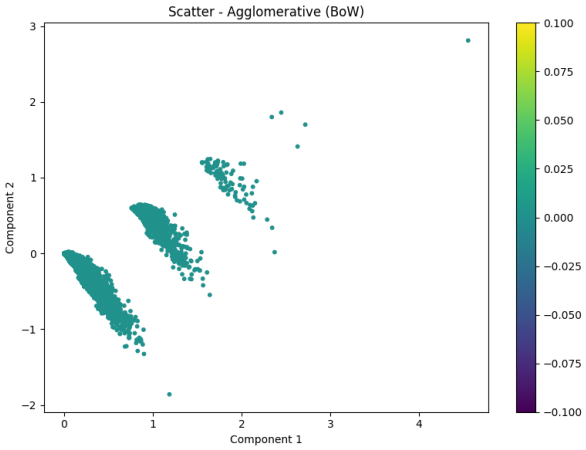


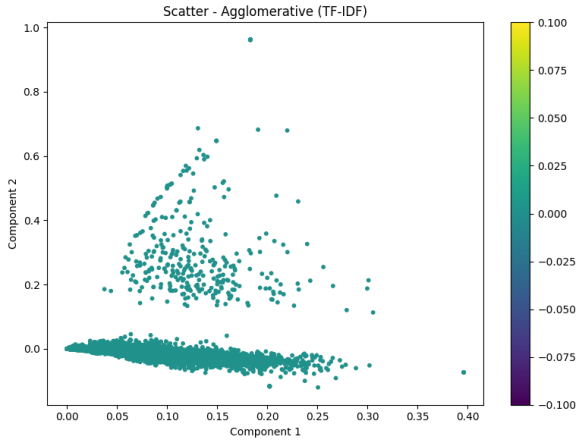


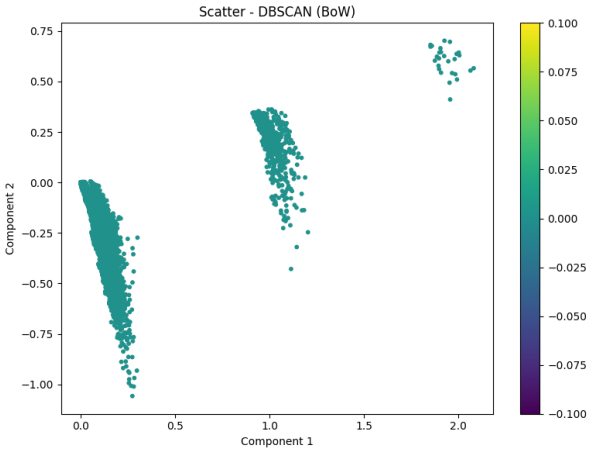
4.3 PCA-based Scatter Plots (2D)

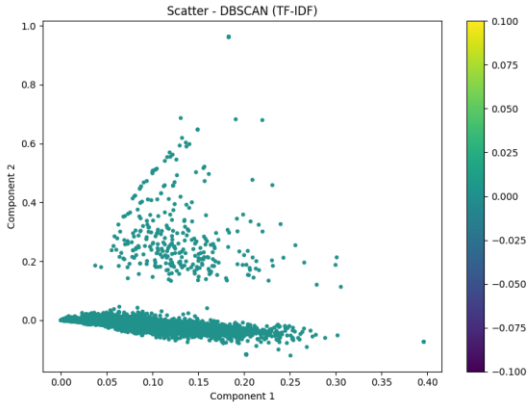












5. Inferences

5.1 For Elbow Curves:

The optimal value of **k was chosen as 2** using the **elbow method**, as the within-cluster sum of squares (WCSS) graph showed a sharp decline after **k=2**, and flattened out beyond that — indicating diminishing returns for higher k values for both the vectorization techniques.

5.2 For Confusion Matrix:

On reviewing the confusion matrices and accuracy scores:

- The **accuracy appears consistently around 76% across all models and vectorization methods**.
- However, this high accuracy is misleading because the models are heavily biased towards predicting the **'Negative'** class — likely because the dataset has a **dominant number of negative samples**.
- As a result, while the accuracy seems decent, the clustering algorithms **fail to meaningfully distinguish between different sentiment categories**.
- Most predictions are classified as 'Negative', with very few or no samples labeled as 'Neutral' or 'Positive', leading to **imbalanced clustering outcomes**

Confusion Matrix Summary

Algorithm	Vectorization	True Negative	True Neutral	True Positive	Predicted Negative	Predicted Neutral	Predicted Positive	Accuracy
K-Means	BoW	796,576	140	247,489	1,044,205	0	0	76.3%
K-Means	TF-IDF	796,576	140	247,489	1,044,205	0	0	76.3%
Agglomerative	BoW	7,635	2	2,363	10,000	0	0	76.4%
Agglomerative	TF-IDF	7,635	2	2,363	10,000	0	0	76.4%
DBSCAN	BoW	6,207	1	1,991	8,199	0	0	75.7%
DBSCAN	TF-IDF	7,635	2	2,363	10,000	0	0	76.4%

5.3 For Scatter Plots:

- DBSCAN with BoW vectorization identified three clear, well-separated clusters, suggesting natural groupings in the data.
- DBSCAN with TF-IDF vectorization produced mainly two distributions: a dense horizontal cluster near the bottom and scattered points above, showing less distinct clustering than with BoW.
- K-Means with BoW revealed multiple parallel linear structures, potentially corresponding to different language patterns in the feature space.

- K-Means with TF-IDF created a more triangular distribution with two main sections, indicating TF-IDF transformed the data differently than BoW for K-Means.
- Agglomerative clustering results resembled their DBSCAN counterparts for both vectorization methods, with BoW showing more defined clusters than TF-IDF.
- The vectorization method (BoW vs TF-IDF) had a more significant impact on clustering results than the choice of algorithm in many cases.

Conclusion

This project successfully applied unsupervised clustering algorithms for sentiment analysis on tweet texts This project explored unsupervised clustering algorithms for sentiment analysis on tweet texts using Bag-of-Words and TF-IDF techniques. Although the models showed around **76% accuracy**, the results were biased towards the majority 'Negative' class, making clustering unreliable for sentiment prediction.

References

https://thesai.org/Downloads/Volume15No3/Paper_14-Clustering_Algorithms_in_Sentiment_Analysis_Techniques.pdf#page=4.64