**Twitter Dataset Analysis and Topic Modeling**

# **1. Introduction**

Social media platforms such as Twitter generate vast amounts of textual data daily. Analyzing this data provides insights into trending topics, user sentiments, and social discourse. This report details the steps taken to preprocess, analyze, and extract meaningful topics from a Twitter dataset using BERT-based topic modeling.

# **2. Data Sources**

The dataset used in this study was manually downloaded from Kaggle and contains various metadata fields. However, for this analysis, only the "text" column was considered. A subset of 4,000 tweets was randomly selected for processing and analysis.

1. Why only keep the 'text' column?

Since our analysis focuses on **BERT-based topic modeling**, the key feature is the **tweet content** itself. Other columns, like timestamps, user information, or metadata, are **not directly useful** for topic modeling.

# **3. Preprocessing for BERT Topic Modeling**

Since BERT is a contextual model, proper text preprocessing is crucial. The following steps were performed:

* **Removal of URLs**
* **Handling of Mentions and Hashtags**: User mentions (@user) and hashtags (#topic) were retained as they provide context.
* **Emoji and Contraction Expansion**: Emojis were converted to text descriptions, and contractions (e.g., "can't" → "cannot").
* **Cleaning Special Characters**: Non-alphanumeric characters were removed.
* **Whitespace Normalization**

# **4. Exploratory Data Analysis (EDA)**

1. **Tweet Length Distribution and words cloud**: A histogram revealed that most tweets were between 50–160 characters long with a mean of 100 characters per tweet and std of 26.31, and Common words were visualized using a word cloud, and popular hashtags and mentions were identified.

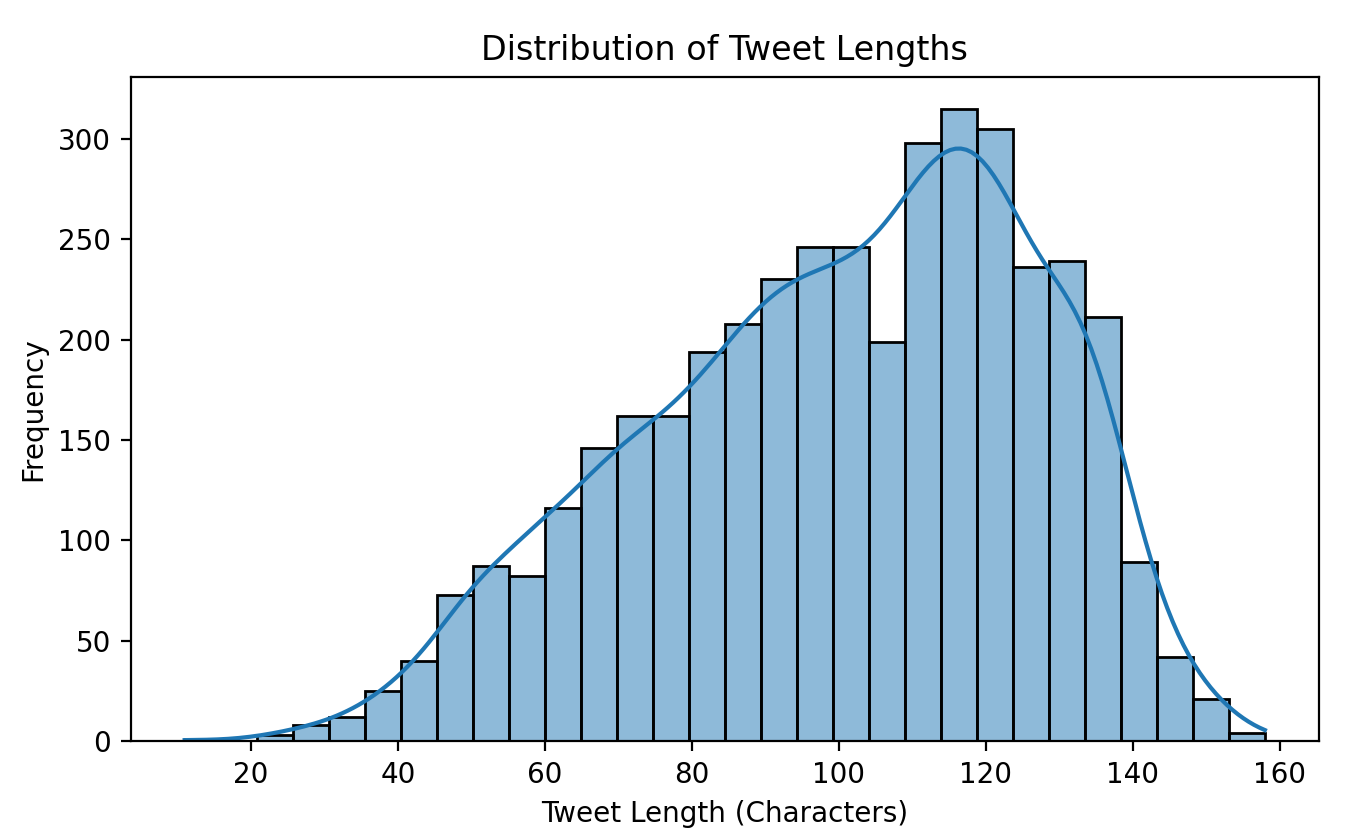


Figure 1: Tweet Lengths Bar and Distribution Shape Plot

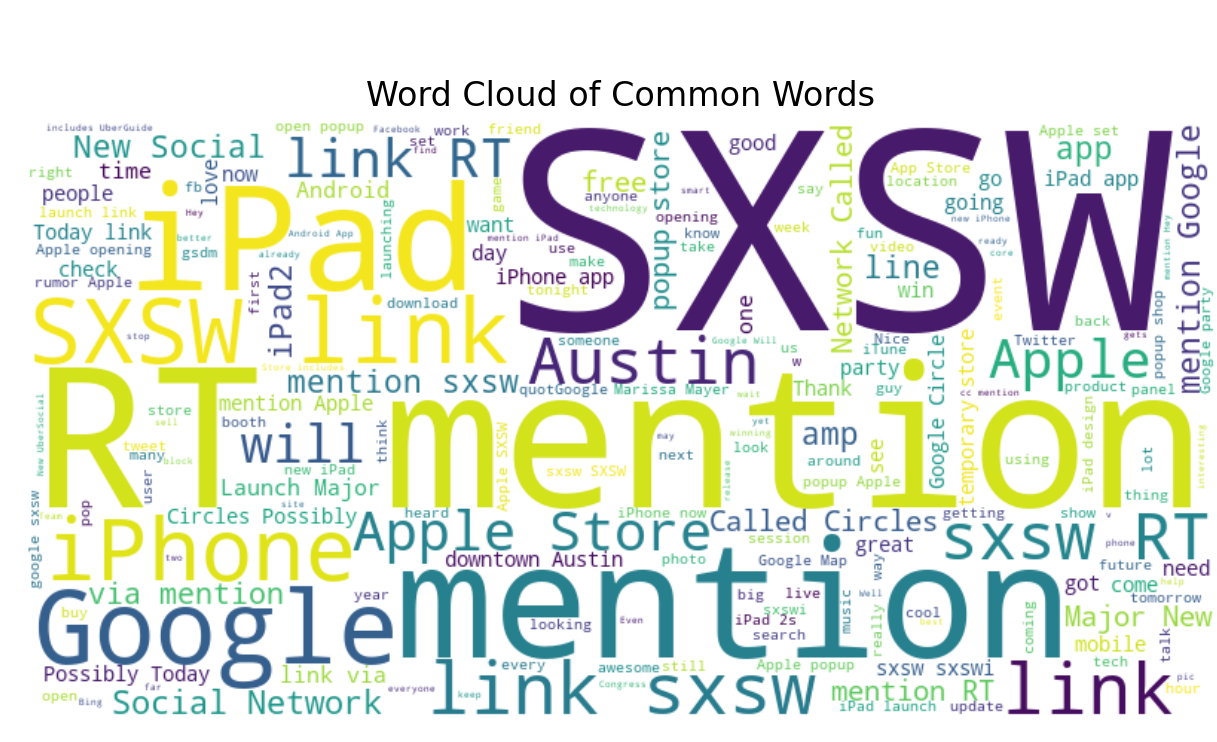


Figure 2: Common Words Cloud Graph

1. **Hashtag and Mention Analysis**: The most common hashtags and mentions were extracted, highlighting popular discussion topics and influential users.

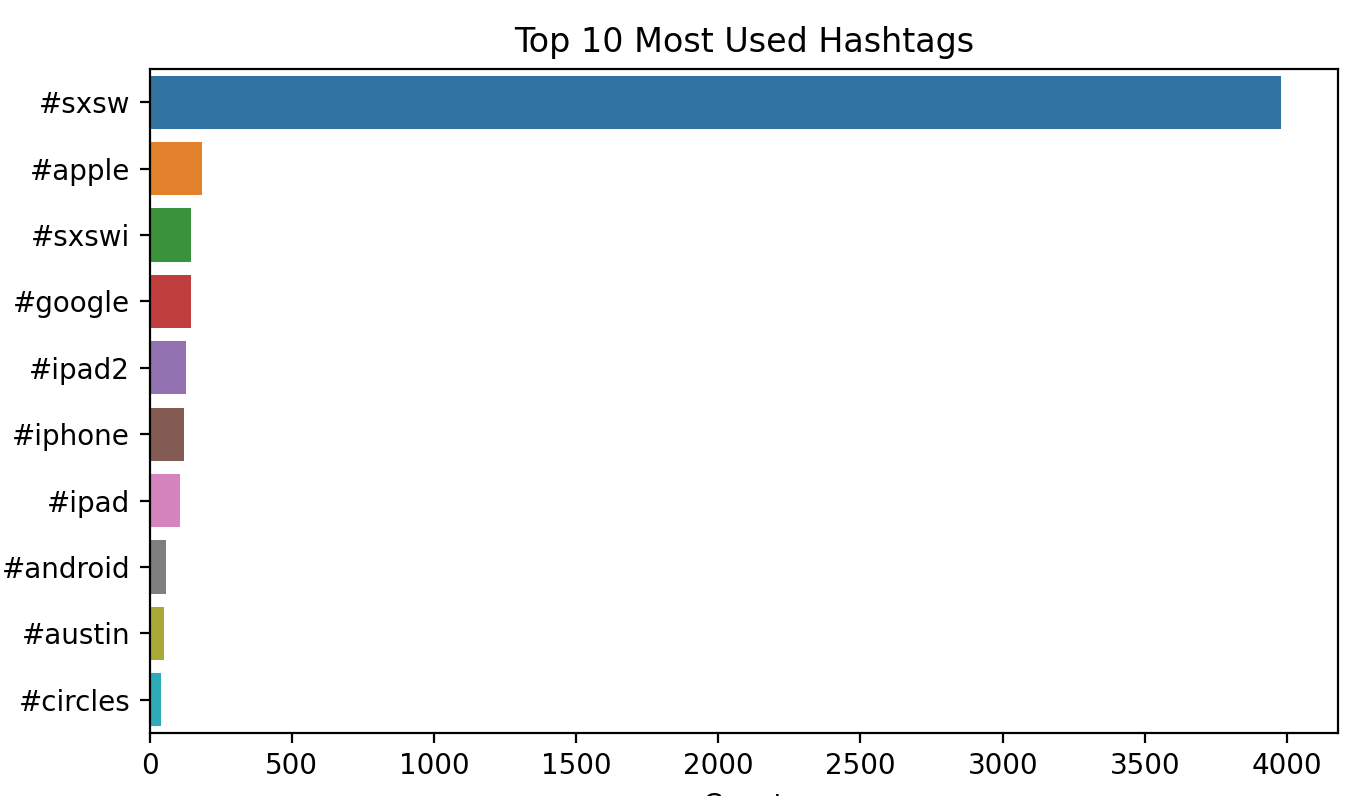


Figure 3: Most Used Hashtags Bar Plot

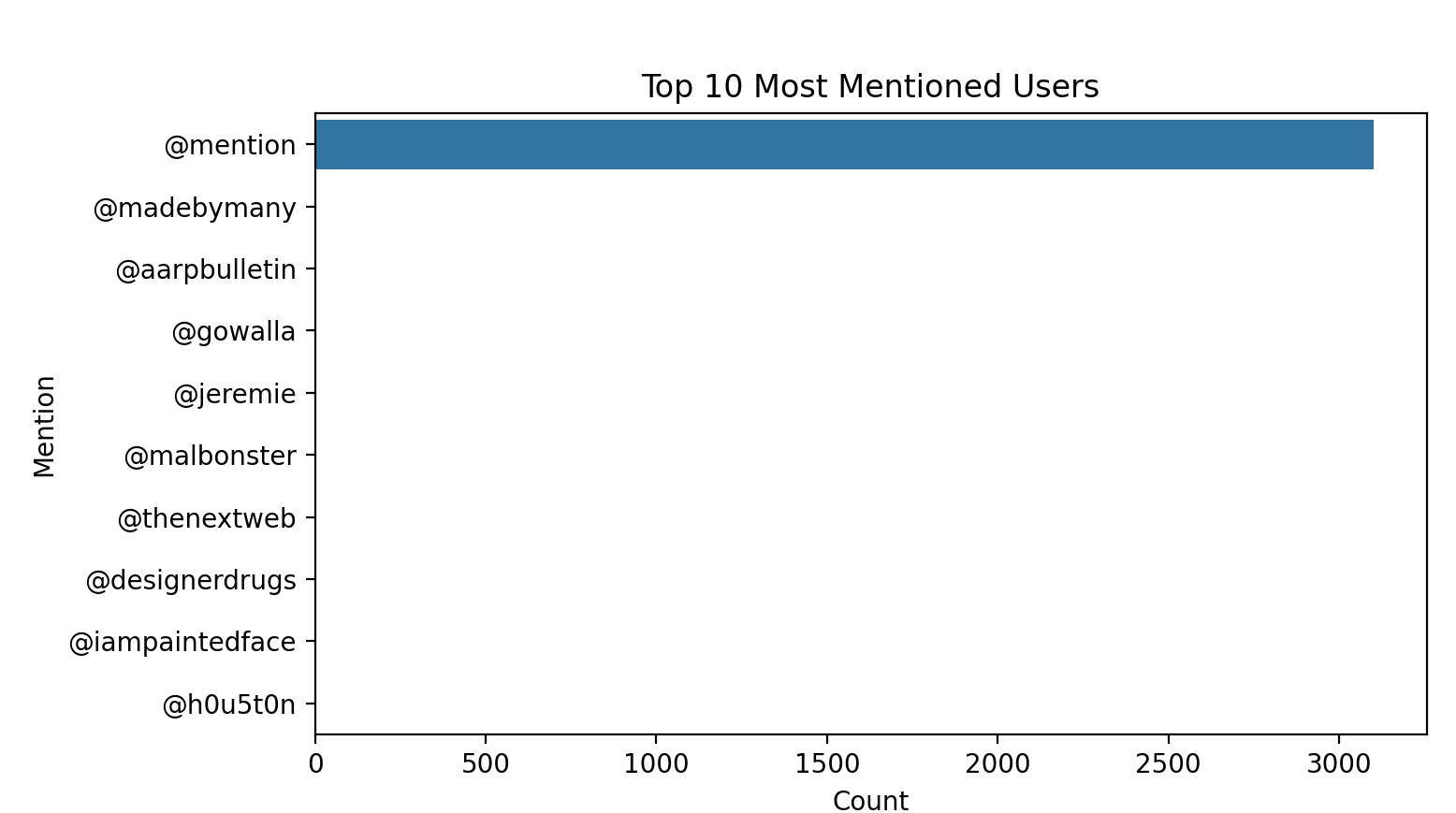


Figure 4: Most Mentioned Users Bar Plot

# **Feature Transformation & Topic Modeling**

In this step, the goal is to transform textual data into numerical features and apply clustering and topic modeling techniques to extract meaningful patterns from the data.

## **TF-IDF Transformation**

Textual data, in its raw form, cannot be directly used for machine learning models since these models require numerical data.

* **How it works**:
  + **Term Frequency (TF)**: word frequency in a document.
  + **Inverse Document Frequency (IDF)**: the word frequency by how commonly the word appears across all documents.

The final TF-IDF score for each word in each document is the product of these two measures.

* **Advantages**:
  + TF-IDF emphasizes words that are unique to specific documents, and reduces the impact of less informative words.

## **K-Means Clustering**

K-Means is an unsupervised learning technique that is used to partition the data into a predefined number of clusters.

* **How It Works**:
  + K-Means works by assigning each document to the nearest cluster center (centroid).
  + The algorithm continues to adjust the centroids until the assignments no longer change, or the changes are minimal.
* **Advantages**
  + K-Means is efficient and scalable, making it one of the most widely used clustering algorithms for large datasets.
  + It works well when the clusters are spherical and have a similar size and density.

### **5.3 Latent Dirichlet Allocation (LDA)**

LDA is a probabilistic model used for **topic modeling**, which helps identify the underlying topics within a collection of documents. Each document is modeled as a mixture of topics, and each topic is characterized by a distribution over words.

* **How it works**:
  + LDA assumes that there are a fixed number of topics in the corpus, and each document is a probabilistic mixture of these topics.
  + LDA works by iteratively assigning words in documents to topics based on the word distributions within each topic, ultimately generating the most likely topics for each document.
* **Advantages**:
  + LDA can provide a human-interpretable summary of large document collections by identifying the topics that are prevalent across the documents.

## **Results:**

**Cluster Distribution**

The visual representation of the clusters, helps to understand the distribution of tweets across the different clusters.

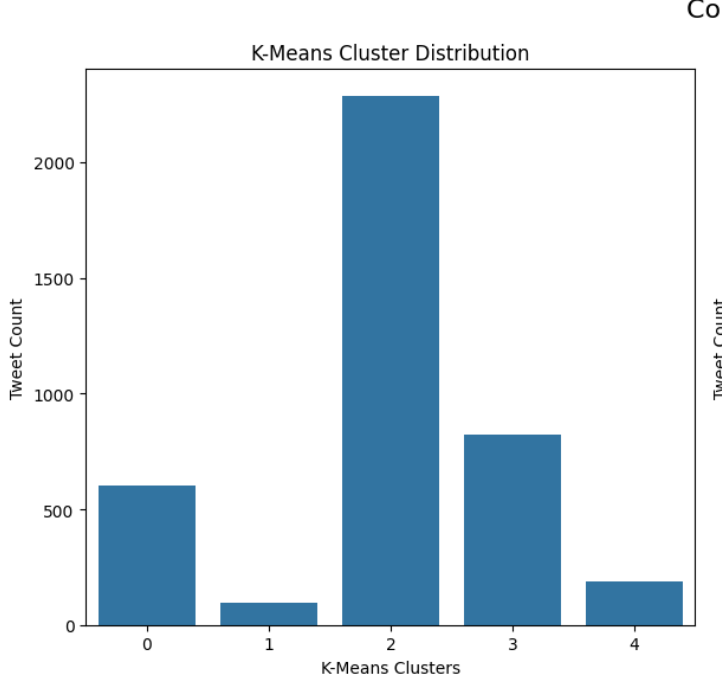


Figure 5: Cluster Distribution of Tweets (K-means)

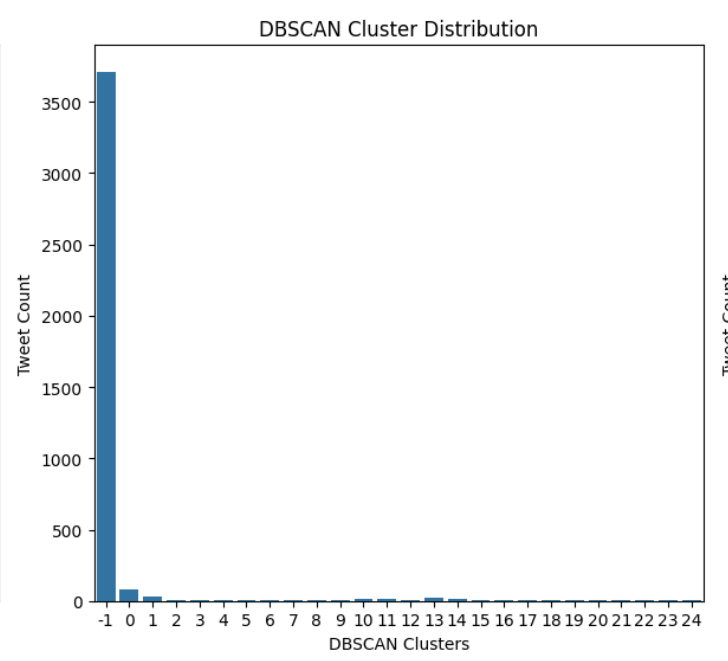


Figure 6: DBSCAN Cluster Distribution of tweets

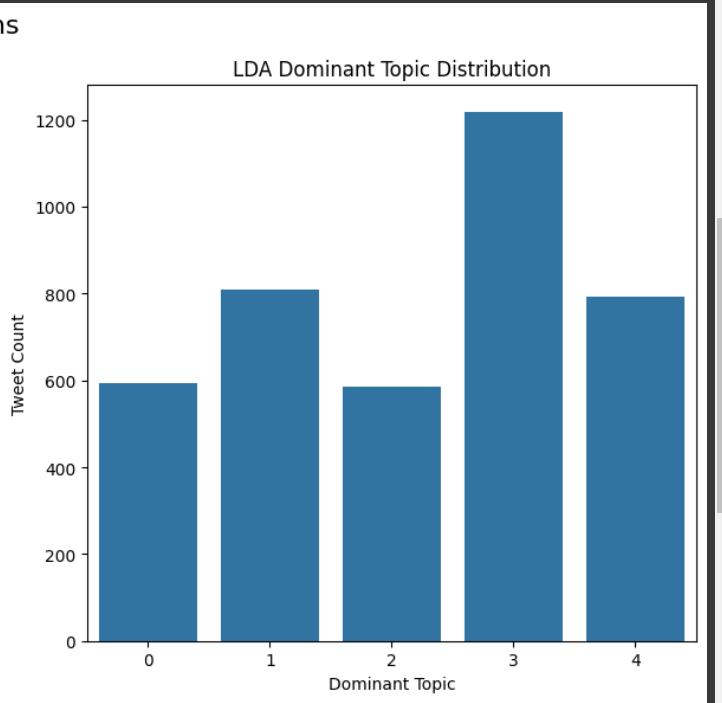


Figure 7: LDA Dominant Topic Distribution in tweets

**Interpretations**:

* **K-Means Cluster Distribution**: In this type of topic modeling, a tweet is only permitted to belong to one cluster, which might not be the balanced approach we need.
* **DBSCAN Cluster Distribution**: DBSCAN clustering shows a very different result, where a majority of tweets fall under Cluster -1, which is typically classified as noise or outliers in DBSCAN. This suggests that DBSCAN is finding many points that do not fit well into any specific cluster, leading to a larger number of noise points.
* **LDA Dominant Topic Distribution**: In contrast, LDA has a much more balanced distribution across its topics, since it allows tweets to belong to multiple clusters with different percentages.

# **BERT-Based Feature Transformation & Topic Modeling**

## **BERT Embeddings:**

Converted text into 768-dimensional vector representations using the pre-trained BERT model to capture deep contextual information, to leverage advanced language understanding for accurate text representation.

## **K-Means Clustering:**

Applied to group similar tweets based on their BERT embeddings, identifying underlying topics, to uncover natural groupings within the data, revealing common themes or topics.

## **PCA (Principal Component Analysis):**

Reduced the high-dimensional embeddings to two dimensions for visualization, to enable the visualization of complex, high-dimensional data in a 2D space for easier interpretation.

## **PCA-reduced clusters scatter plot:**

The scatter plot of PCA-reduced clusters provides an intuitive overview of how tweets are grouped.

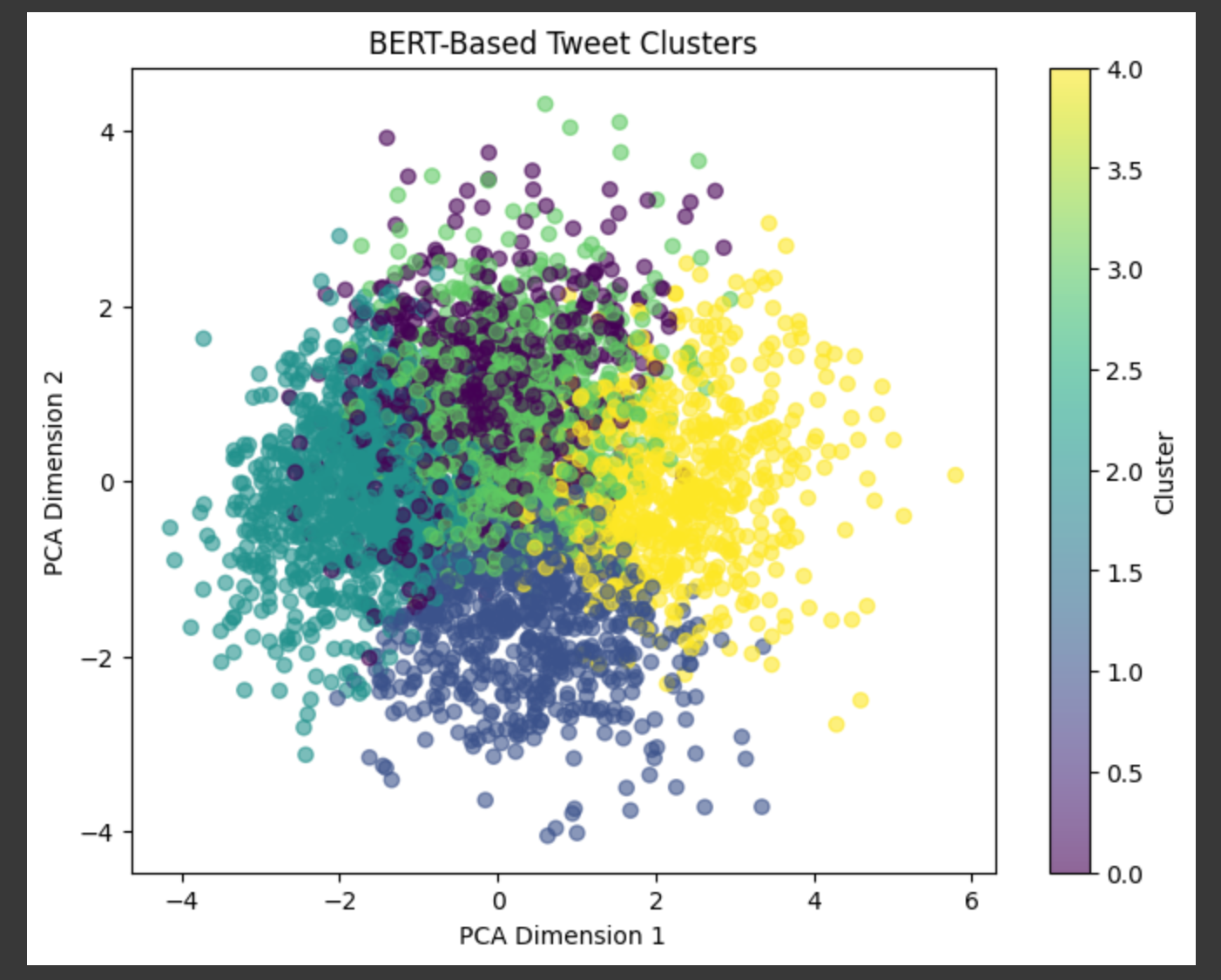


Figure 8: BERT-Based Tweet Clusters

The scatter plot shows the distribution of tweets in the reduced 2D space. Points that are closer together belong to the same cluster, indicating that they share similar content or themes.

# **Focus on iPhone Brand**

* 1. **Keyword Filtering**: Filtered tweets mentioning iPhone or related hashtags to focus on brand-specific content.
  2. **Topic Modeling (LDA)**: Extracted topics from iPhone-related tweets to uncover common themes.
  3. **Sentiment Analysis**: Analyzed the sentiment of iPhone-related tweets to gauge public opinion using a library called **textblob.**

**Results**:

* 1. **Cluster Distribution of iPhone-related Tweets (K-Means)**:

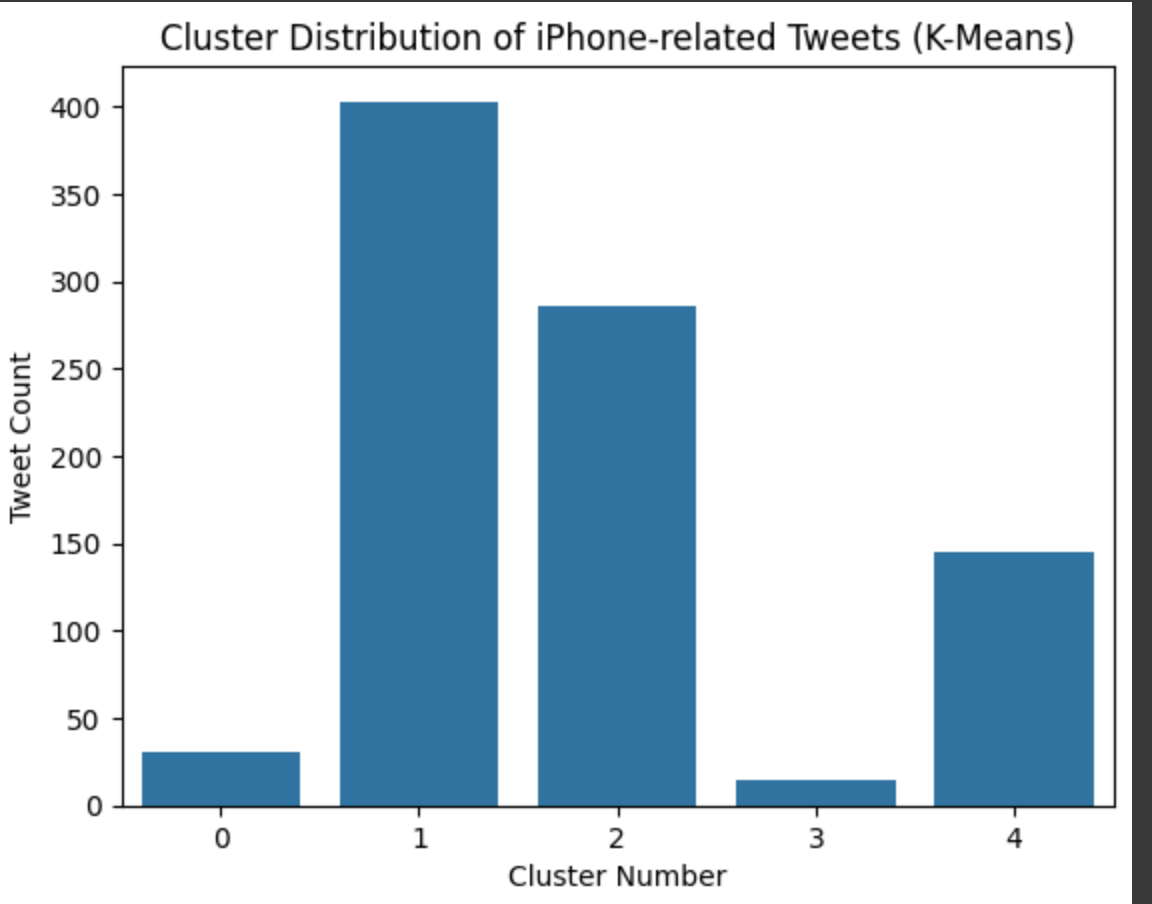


Figure 9: Distribution of iPhone related tweets

* 1. **Sentiment Distribution of iPhone-related Tweets**:

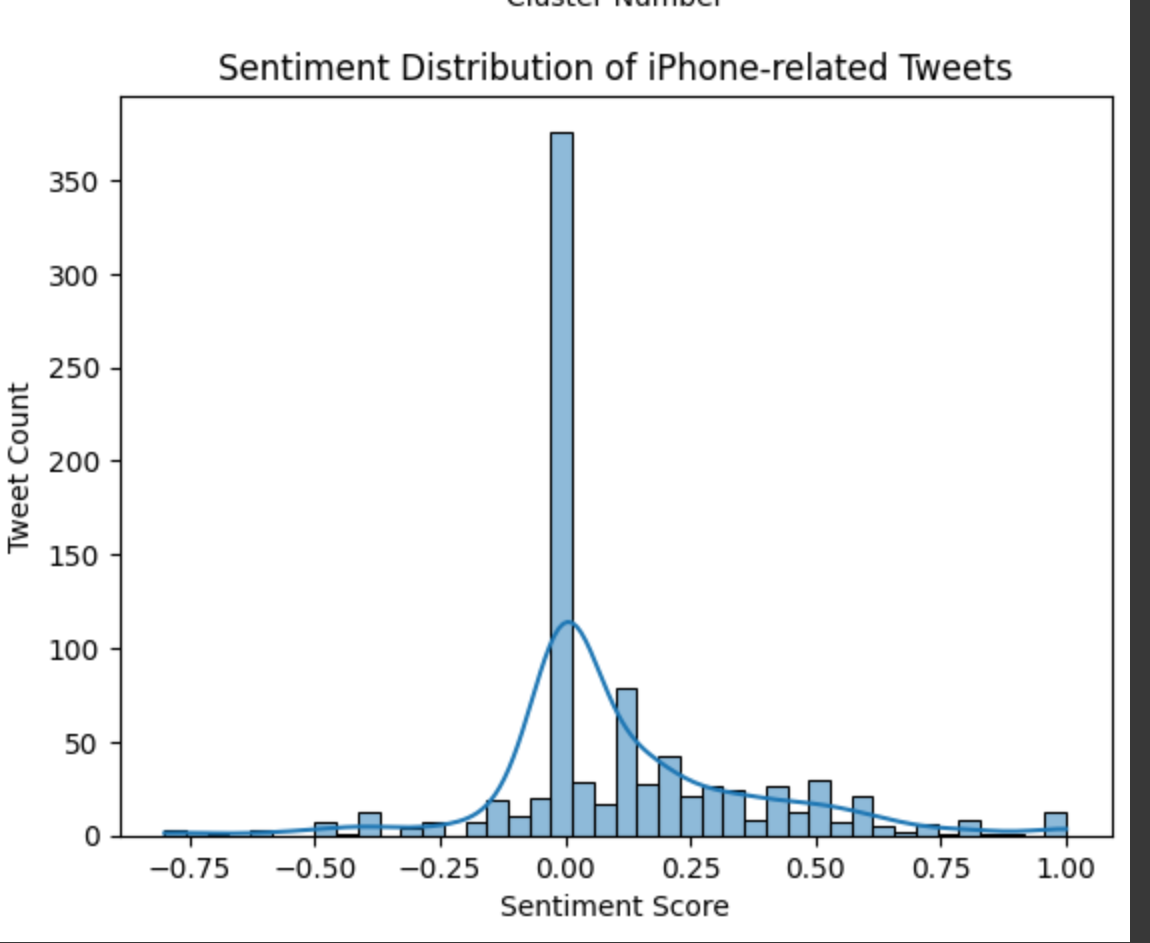


Figure 10: Sentiment Distribution of iPhone related tweets.

**Interpretations**:

* **Most tweets are neutral**, centered around a sentiment score of **0.00**.
* **Slightly more positive tweets** than negative, with a right-skewed distribution.
* **Few strongly negative tweets**, but they exist up to around **-0.75**.
* **Positive sentiment extends up to 1.00**, but extreme positivity is rare.
* **Kernel Density Estimate (KDE) curve confirms** the majority are **neutral to slightly positive**.

# **Conclusion**

This study explored topic modeling on Twitter data using both traditional NLP techniques (TF-IDF, LDA, and K-Means) and deep learning models (BERT embeddings).

* **Preprocessing Enhancements**: Cleaning and structuring tweet text improved the quality of extracted topics.
* **Clustering Insights**: K-Means and DBSCAN produced different results, with DBSCAN identifying more noise, while LDA provided a more balanced topic distribution.
* **BERT-based Analysis**: Deep learning-based clustering revealed more nuanced topic groupings, demonstrating the power of contextual embeddings.
* **iPhone-Specific Analysis**: Focusing on brand-related tweets uncovered key themes and sentiment trends, with most tweets being neutral or slightly positive.

This analysis highlights the effectiveness of combining classical and modern NLP techniques for social media insights, offering a scalable approach to understanding trends, sentiments, and brand perception on Twitter.

# Bibliography

DhanushKumar. (2024, 07 1). *Topic Modelling with BERTopic.* Récupéré sur Medium: https://medium.com/%40danushidk507/topic-modelling-with-bertopic-249095144555

Hebbar, N. (2020, 12 29). *Twitter Sentiment Analysis Using Python for Complete Beginners.* Récupéré sur Medium: https://medium.com/swlh/tweet-sentiment-analysis-using-python-for-complete-beginners-4aeb4456040

Jayasree Ravi, S. K. (2023, 02 7). *Text embedding techniques for efficient clustering of twitter data.* Récupéré sur PMC – National Center for Biotechnology Information (NCBI): https://pmc.ncbi.nlm.nih.gov/articles/PMC9904526/

Kleppen, E. (2023, 02 12). *Topic Modeling For Beginners Using BERTopic and Python.* Récupéré sur Medium: https://python.plainenglish.io/topic-modeling-for-beginners-using-bertopic-and-python-aaf1b421afeb

Nils Constantin Hellwig, J. F. (2024, 10 29). *Exploring Twitter Discourse with BERTopic: Topic Modeling of Tweets during the 2021 German Federal Election.* Récupéré sur Springer : https://link.springer.com/article/10.1007/s10772-024-10142-4

*understanding-tf-idf-term-frequency-inverse-document-frequency.* (2025, 02 07). Récupéré sur geeksforgeeks: https://www.geeksforgeeks.org/understanding-tf-idf-term-frequency-inverse-document-frequency/