Analyzing Twitter Engagement on Apple Vision Pro Using K-Means Clustering

Andrew Angulo

Denison University
September 2024

Abstract

The Apple Vision Pro has generated significant discussion on social media platforms, particularly Twitter. This study employs the K-Means clustering algorithm to analyze user engagement with tweets related to the Apple Vision Pro. By examining metrics such as reply count, quote count, retweet count, like count, views, and bookmark count, the research identifies distinct patterns and clusters of user interaction. The findings reveal three primary engagement levels—low, moderate, and high—which offer insights into public perception and the influence of key users. These results can inform targeted marketing strategies and enhance understanding of consumer sentiment towards the Apple Vision Pro.

1 Introduction

The launch of innovative technology products often sparks conversations and debates among consumers, influencers, and industry experts. The Apple Vision Pro, Apple's latest venture into virtual reality, has not only captivated tech enthusiasts but also become a trending topic on social media platforms like Twitter. Understanding the dynamics of these discussions is crucial in assessing the public perception, market reception, and informing strategic decision making for both Apple and marketers.

Social media serves as a vital channel for real-time feedback and engagement, providing a wealth of data that can be analyzed to extract meaningful insights. Twitter, in particular, offers a platform where users express their opinions, share experiences, and interact with brands and each other, making it an invaluable resource for sentiment analysis and engagement studies.

Despite the abundance of data, extracting actionable insights requires sophisticated analytical techniques. Clustering algorithms, such as K-Means, are instrumental in categorizing data into meaningful groups based on similarity, enabling researchers to identify patterns and trends that may not be immediately apparent.

This study aims to apply the K-Means clustering algorithm to Twitter engagement metrics related to the Apple Vision Pro. By analyzing features such as reply count, quote count, retweet count, like count, views, and bookmark count, this research paper seeks to show the distinct patterns of user interaction. The primary objectives of this study are:

- To identify and categorize different levels of user engagement with tweets about the Apple Vision Pro.
- To analyze the characteristics of each engagement cluster to understand user behavior and sentiment.
- To provide actionable insights that can inform marketing strategies and enhance consumer engagement.

By achieving these objectives, this will help us to achieve a deeper understanding of social media engagement dynamics, offering valuable insight for marketing professionals and brand strategists aiming to optimize their social media presence and consumer interactions.

2 Data Collection and Preprocessing

The dataset used in this analysis was obtained from Kaggle and consists of tweets about the Apple Vision Pro. The data includes various metrics which are used for understanding user interaction:

- **Reply Count**: The number of replies a tweet has received.
- **Quote Count**: The number of times a tweet has been quoted.
- **Retweet Count**: The number of times a tweet has been retweeted.
- **Like Count**: The number of likes a tweet has received.
- **Views**: The number of times a tweet has been viewed.
- **Bookmark Count**: The number of times a tweet has been bookmarked.

To ensure the integrity of the data, any rows containing missing values in these features were removed. The data was then loaded into a pandas DataFrame to facilitate efficient manipulation and analysis.

Table 1: Sample of Twitter Engagement Data on Apple Vision Pro

(a) Engagement Metrics

Tweet ID	Author	Reply Count	Quote Count	Retweet Count
1769458624638619691	Harndefty	0	0	0
1769456825731346925	Leigh	0	0	0
1769454738704302260	Techlistics	0	0	0
1769438420114317638	Deals_Store	0	0	0
1769434792159039744	AITnews	0	0	0

(b) Additional Metrics

Like Count	Views	Bookmark Count
0	26	0
0	6	0
1	52	0
1	281	0
0	342	0

Table 1 provides a snapshot of the dataset, showing the structure and types of engagement metrics collected for each tweet.

2.1 Feature Selection and Normalization

Given that the engagement metrics are on different scales (for example, views can be in the thousands while bookmark counts may be much lower), normalization is necessary to prevent features with larger magnitudes from dominating the clustering process. We applied min-max normalization to scale each feature to a range between 0 and 1, using the formula:

$$X_{\text{normalized}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{1}$$

This transformation ensures that each feature contributes equally to the distance calculations in the K-Means algorithm.

2.2 K-Means Clustering Algorithm

K-Means clustering is an unsupervised machine learning algorithm used to partition data into K distinct clusters based on feature similarity. The algorithm aims to minimize the within cluster sum of squares, which effectively groups similar data points together.

The primary objective of K-Means is to divide the dataset into K clusters where each data point belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

The steps of the K-Means algorithm are as follows:

1. **Initialization**: Randomly select K initial centroids from the dataset.

2. **Assignment Step**: Assign each data point to the nearest centroid based on Euclidean distance.

3. **Update Step**: Recalculate the centroids as the mean of all data points assigned to each cluster.

4. **Iteration**: Repeat the assignment and update steps until convergence (when centroids no longer change significantly) or a maximum number of iterations is reached.

In this study, we set $K = 3$, were saying that the engagement data might naturally form three clusters representing low, medium, and high engagement levels.

To evaluate the clustering performance and monitor convergence, we calculated the Mean Squared Error at each iteration:

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{x}_i - \mathbf{c}_{k_i}\|^2 \tag{2}$$

where:

- N is the number of data points.
- $\mathbf{x}_i$ is the i-th data point.
- $\mathbf{c}_{k_i}$ is the centroid of the cluster to which $\mathbf{x}_i$ is assigned.

This metric quantifies the average squared distance between each data point and its assigned centroid, providing a measure of clustering performance.

2.3 Dimensionality Reduction for Visualization

The dataset consists of six engagement features, making direct visualization challenging. To visualize the clustering results, we selected specific pairs of normalized features to create two-dimensional plots. This approach allows us to examine the relationships between features and how data points are clustered based on these features.

We chose the following feature pairs for visualization:

- **Like Count vs. Retweet Count**: To understand how likes correlate with retweets.
- **Reply Count vs. Quote Count**: To examine the interaction between replies and quotes.
- **Views vs. Bookmark Count**: To analyze the relationship between the number of views and bookmarks.

Each pair provides a different perspective on user engagement, telling us various aspects of tweet interactions.

3 Results

3.1 Convergence of the K-Means Algorithm

The K-Means algorithm was executed with a maximum of 100 iterations. The Mean Squared Error was calculated at each iteration to monitor the convergence. The algorithm converged after a certain number of iterations, indicating that the centroids stabilized and data point assignments no longer changed significantly.

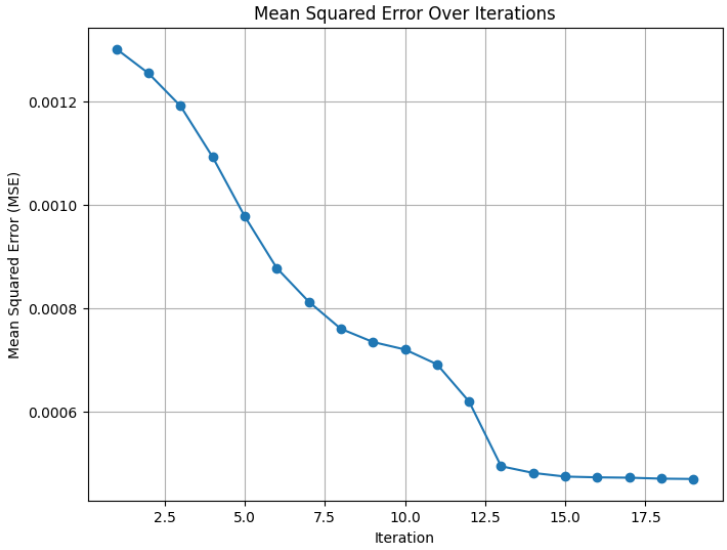

Figure 1: Mean Squared Error over iterations of the K-Means algorithm.

Figure 1 shows the MSE decreasing over iterations, showing the convergence of the algorithm.

3.2 Clustering Outcomes

After convergence, the dataset was split into three clusters. To visualize the clusters, we plotted data points using selected pairs of normalized features.

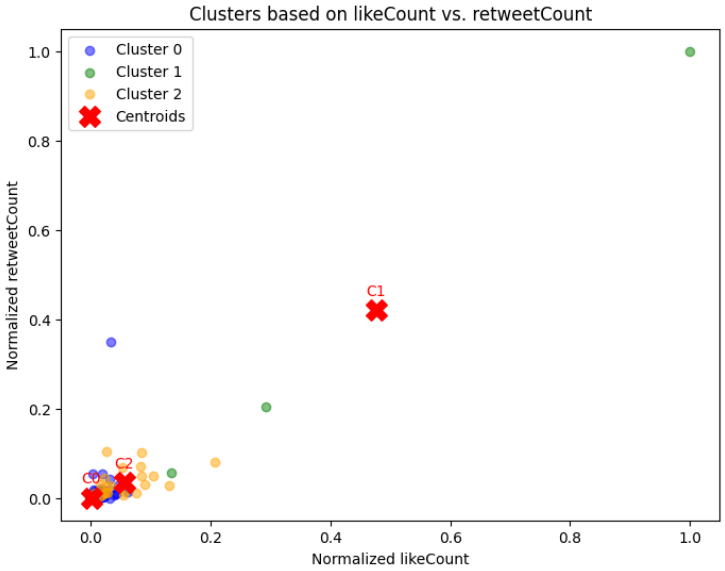

Figure 2: Clusters based on normalized Like Count vs. Retweet Count.

Figure 2 shows the clustering of tweets based on their like counts and retweet counts. Each point represents a tweet, colored according to its assigned cluster. The centroids are indicated by red crosses. The x-axis represents the normalized Like Count, and the y-axis represents the normalized Retweet Count.

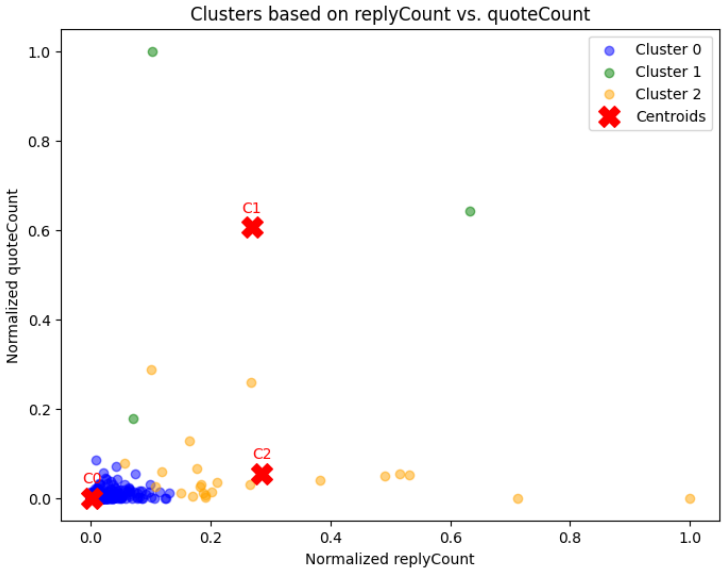

Figure 3: Clusters based on normalized Reply Count vs. Quote Count.

Figure 3 illustrates the clustering of tweets based on their reply counts and quote counts. The x-axis represents the normalized Reply Count, and the y-axis represents the normalized Quote Count.

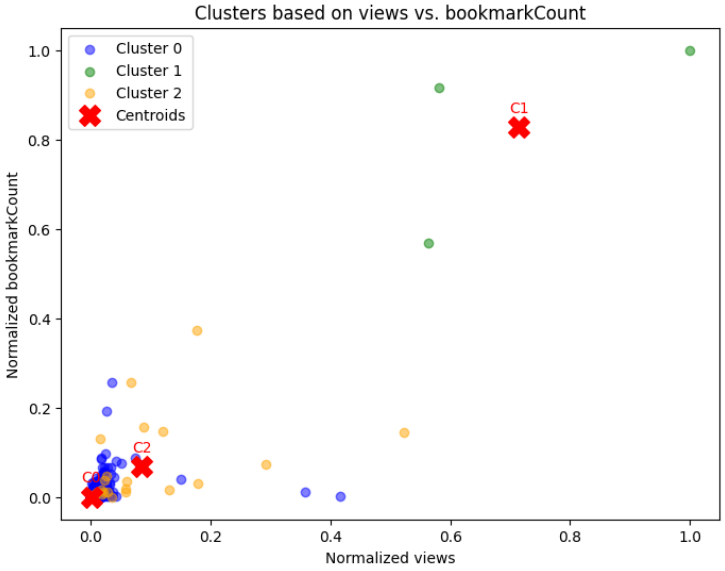

Figure 4: Clusters based on normalized Views vs. Bookmark Count.

Figure 4 depicts the clustering of tweets based on their views and bookmark counts. The x-axis represents the normalized Views, and the y-axis represents the normalized Bookmark Count.

These plots provide a visual representation of how tweets are grouped based on different engagement metrics, highlighting the distinct clusters formed by the K-Means algorithm.

3.3 Cluster Interpretation

Upon analysis, the clusters can be interpreted as follows:

- **Cluster 0 (Low Engagement)**: Tweets with minimal engagement across all metrics. These tweets have low counts of replies, quotes, retweets, likes, views, and bookmarks. This cluster likely represents the majority of tweets with limited interaction.
- **Cluster 1 (Moderate Engagement)**: Tweets with moderate levels of engagement. These tweets have higher engagement metrics than Cluster 0 but do not reach the peaks observed in Cluster 2. This cluster may represent general interest or discussions among typical users.
- **Cluster 2 (High Engagement)**: Tweets with significantly high engagement metrics. These are likely tweets from influential users or accounts with a large following, generating substantial interaction in the form of replies, retweets, likes, and views.

3.4 Feature Analysis

To further understand the characteristics of each cluster, we calculated the mean values of the engagement metrics within each cluster. Table 2 summarizes these findings.

Table 2: Average Engagement Metrics by Cluster

Cluster	**Reply Count**	**Quote Count**	**Retweet Count**	**Like Count**	**Views**	**Bookmark Count**
0	0.86	0.16	1.51	8.02	1.40×10^3	0.63
1	142.00	250.00	2,443.00	9,984.67	1.72×10^6	1,018.00
2	150.70	22.70	206.48	1,165.22	2.04×10^5	82.61

Table 2 provides a basis for interpreting the clustering results. The table shows the distinct engagement levels across the three clusters, showing us the variability in user interactions with tweets about the Apple Vision Pro.

4 Conclusion

The usage of K-Means clustering algorithm for Twitter engagement data for the Apple Vision Pro has revealed interesting patterns in user interaction. The information of clusters corresponding to low, moderate, and high engagement levels shows us the influence of different users in the conversation about the product.

These findings show the importance of influential users in shaping public discourse on social media. For companies like Apple, using these insights can enhance marketing strategies and improve engagement with their target audience. By focusing on influencer partnerships, content optimization, and community engagement, companies can effectively navigate the dynamics of social media interactions to foster a more engaged and responsive user base.