Sentiment Mining Project

Spotify Tweets

Dhwani Doshi

Saunders College of Business, Rochester Institute of Technology

MGIS 650: Introduction to Data Analytics & Business Intelligence

May 2, 2024

Table of Contents

Introduction:

In today's digital landscape, social media platforms have become invaluable sources of customer feedback and sentiment. This project aimed to leverage the power of Twitter data to gain insights into user sentiment towards Spotify, the leading music streaming service. The analysis was conducted using various techniques and tools, including R and Python programming languages, as well as Tableau for data visualisation. The primary objective was to gain insights into the sentiment expressed by users towards Spotify through their tweets, identify patterns, and explore the potential implications for the company and the industry.

Methodology:

The project employed a robust methodology, combining advanced data analysis techniques with cutting-edge visualisation tools. It was divided into three distinct phases:

Sentiment Analysis using R:

1) Performed sentiment analysis on tweet text data using multiple methods (syuzhet, afinn, and nrc)
2) Cleaned and preprocessed data to ensure accurate analysis
3) Generated word clouds to visualise frequently occurring terms (Figure 1)

External Exploration using Python:

1) Conducted time series analysis to identify sentiment trends over time (Figure 3)
2) Performed correlation analysis to examine relationships between variables (Figure 4)
3) Analysed sentiment distribution (Figure 9)
4) Analysed sources of tweets (Figure 6)
5) Investigated the impact of verified accounts (Figure 7)
6) Analysed emoticon usage (Figure 8)
7) Compared sentiment classifications across different methods (Figure 10)
8) Evaluated user engagement through retweets and favourite counts (Figure 11, Figure 12)
9) Attempted statistical modelling to predict retweet counts based on sentiment

Data Discovery using Tableau:

1) Utilised Tableau's powerful visualisation capabilities
2) Created interactive dashboards and visualisations for data exploration (Figure 13)

Key Findings:

1) Sentiment Analysis

The analysis (Figure 10) revealed a pre-dominantly positive sentiment towards Spotify, with around 52% positive sentiment with the Syuzhet method, 34% positive sentiment with the Afinn method and 32% positive sentiment with the NRC method. However, a substantial proportion of negative sentiment approximately 32-35% negative sentiment and approx 10% neutral sentiment were observed, indicating room for improvement.

Correlation between sentiment methods (Figure 4) shows a strong positive correlation of 0.8 indicating a high level of agreement between sentiment scores obtained from all three methods.

2) Time Series Patterns

Time series analysis (Figure 3) showed a huge rise in the number of tweets towards the last week and quarter, this is likely due to the collection method of data. This indicated that recent tweets exist in more numbers in the dataset.

3) User Engagement

Positive sentiment scores were found to be moderately correlated with higher levels of user engagement measured by metrics such as favourite count and retweet count. The metrics retweet count and favourite count seem highly correlated (Figure 4), indicating that tweets with more likes are more likely to be reposted by others and vice versa. (Figure 11) Another analysis shows that maximum engagement of likes and retweets is shown on Sunday which indicates that people are more active on weekends. (Figure 12)

4) Influential Voices

The absence of verified accounts in the dataset suggests that either the dataset primarily focuses on tweets from non-verified accounts or there is a collection bias. Further investigation is necessary in this area. (Figure 7)

5) Platform Optimizations

Upon source analysis of tweets, sources like Twitter for iPhone, Twitter for Android, Twitter for Web and sprinklr Care are the most widely used platforms (Figure 6), twitter for iPad is the only Twitter first-party platform with less engagement, indicating room for improvement for application on iPad and tablet size devices.

6) Emotion Analysis

Emotion analysis shows a high proportion of Joy, Anticipation and Trust (Figure 8). This can indicate the platform is enjoyed and trusted by users. Anticipation can indicate that it is popular when some new albums or songs are to be released.

7) Distribution Analysis

The histogram (Figure 9) shows the possibility of one or more outlier data points with extremely high values. It seems that the favourite and retweet count of some outliers are significantly higher than the average number. This can indicate that some tweets have gone viral or gained significant popularity. These outliers could be influential tweets or contain text about popular topics or events.

8) Statistical Modeling

Using the linear regression model to predict retweets, the mean squared error was 7016059.853007733. This high value indicates the model is not accurately capturing the variation in retweet counts. The R squared value was negative which shows the model does not fit well and may be overfitting or underfitting. The predicted retweet count of negative does not make sense and suggests that the model is unreliable in making predictions.

Recommendations :

1) **Sentiment Monitoring**

   Implementing a robust monitoring system for sentiments on tweets and addressing the tweets with negative sentiment. This improves brand engagement and loyalty.

2) **Influencer Engagement**

   The absence of verified accounts engagement can be modified. Identify and collaborate with influential users and industry-leading artists to improve and encourage engagement and overall positive sentiment.

3) **Platform Optimisation**

   A lower activity on Twitter for the iPad platform is noticed. Perhaps a better User interface for applications on iPad or tablet-sized devices with better functionalities should be implemented to increase engagement.

4) **Targeted campaigns**

   The absence of categories in the dataset suggests the need for more targeted campaigns and collaboration to encourage healthy dialogue and conversation and thus increasing advertisements and sales.

5) **Statistical modeling**

   Missing and incorrect data does not fit well in a model, enhancement of the data collection process and using statistical methods to predict and identify factors that are more influential in generating more positive sentiment.

Conclusion:

The Twitter data for Spotify provided invaluable insights into user engagement patterns and shows potential areas of improvement. By leveraging advanced analytics using R and Python programming it shed a light on the overall positive sentiment towards Spotify. The insights gained from this project also emphasise the need for continuous product optimisation and user experience enhancements. By analysing sentiment trends and user feedback, Spotify can identify areas for improvement, develop new features, and align its offerings more closely with user preferences, thereby solidifying its position as a market leader.

Appendix ;

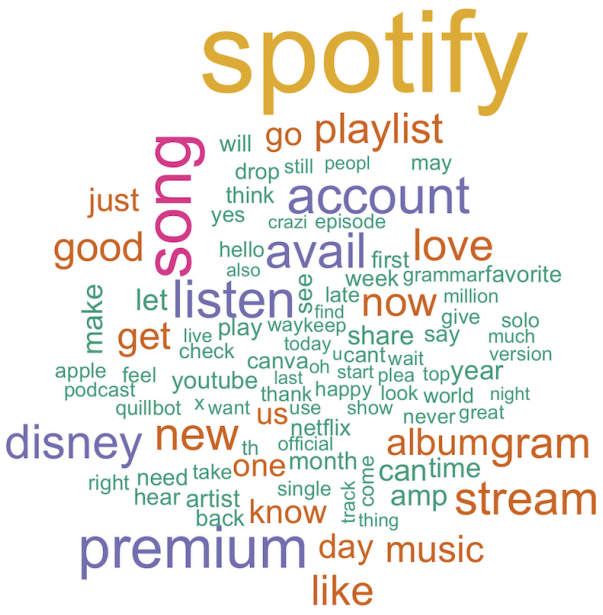

Figure 1 - Wordcloud

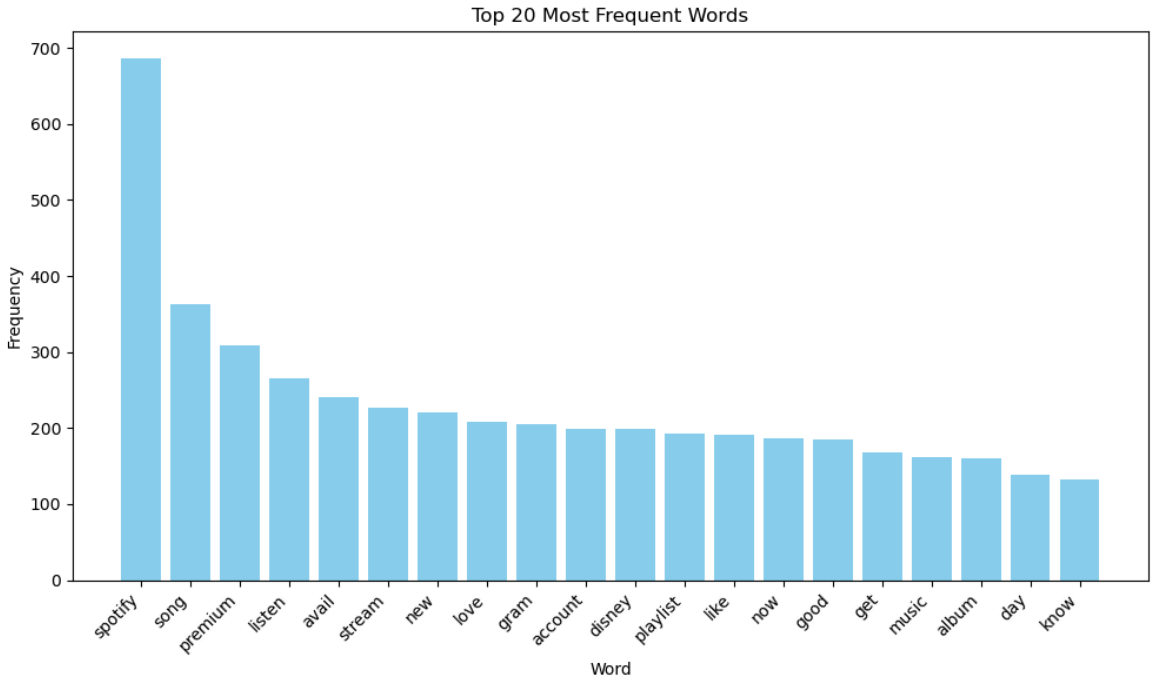

Figure 2 - Most frequent words

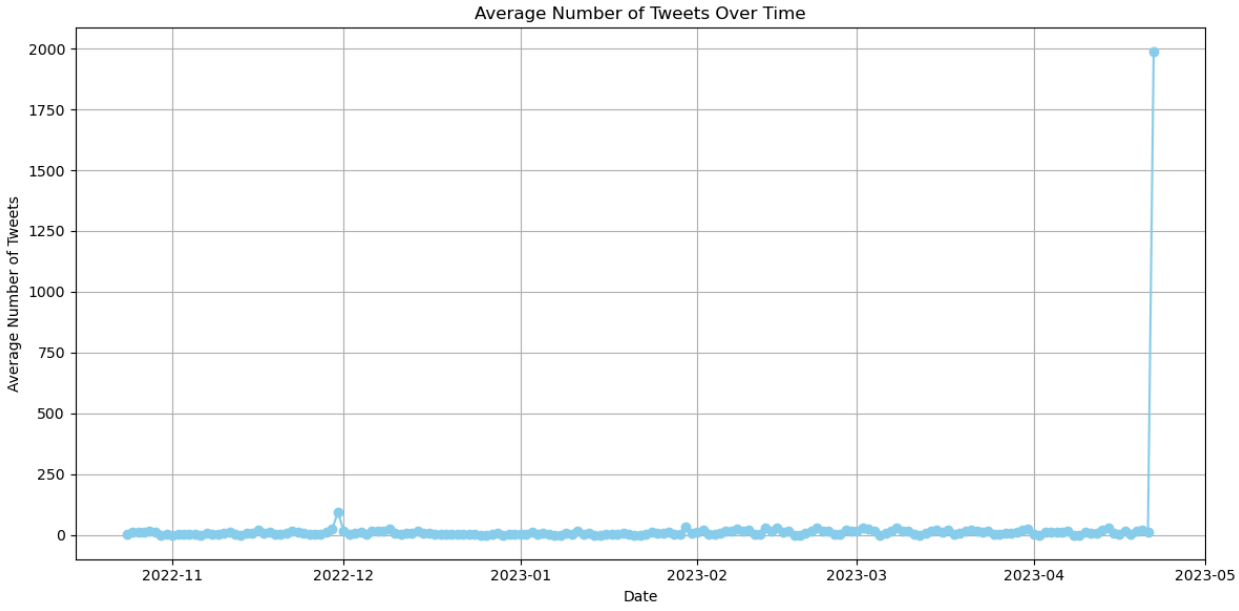

Figure 3 - Average number of tweets over time

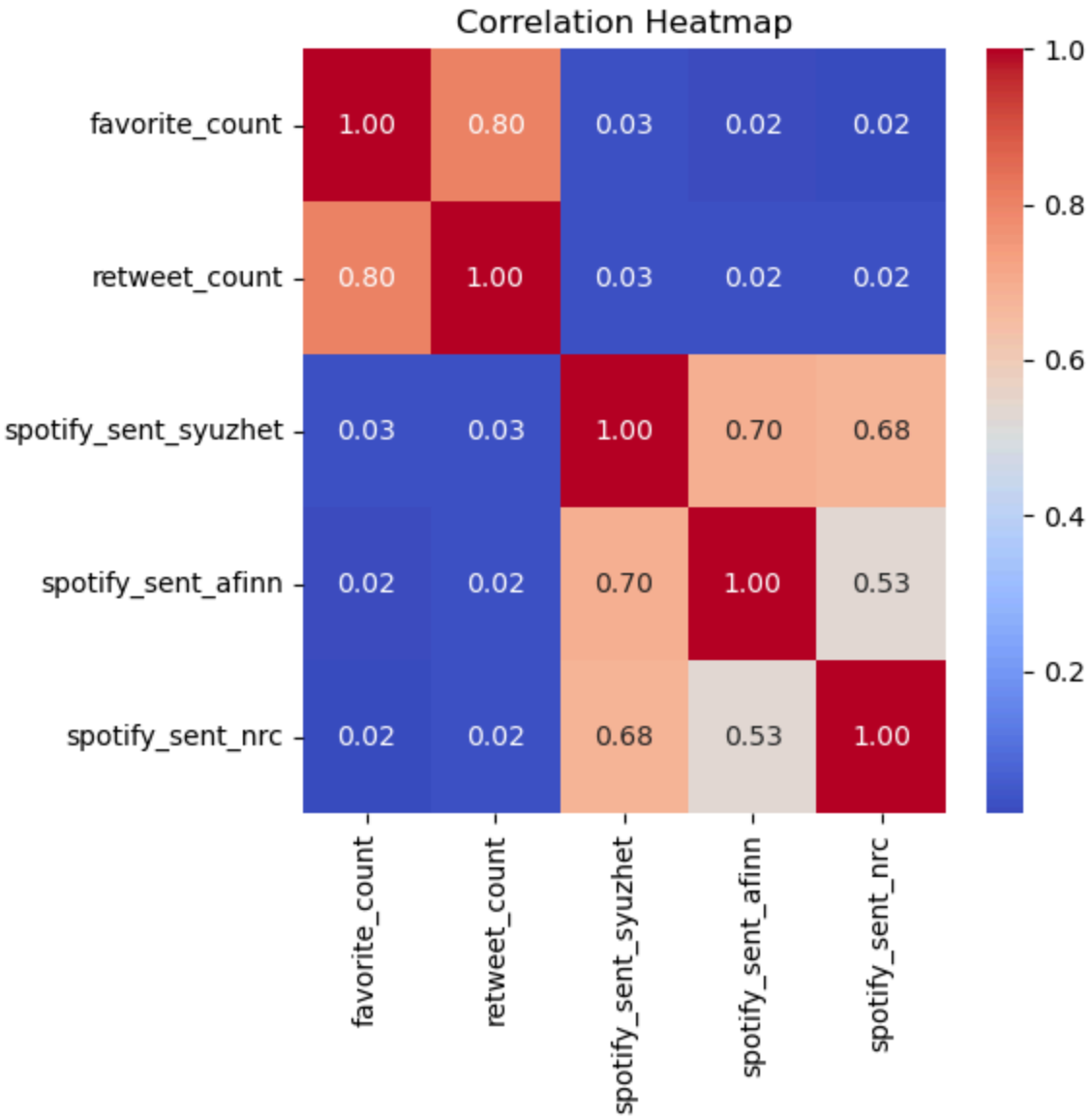

Figure 4 - correlation between sentiment scores and metrics

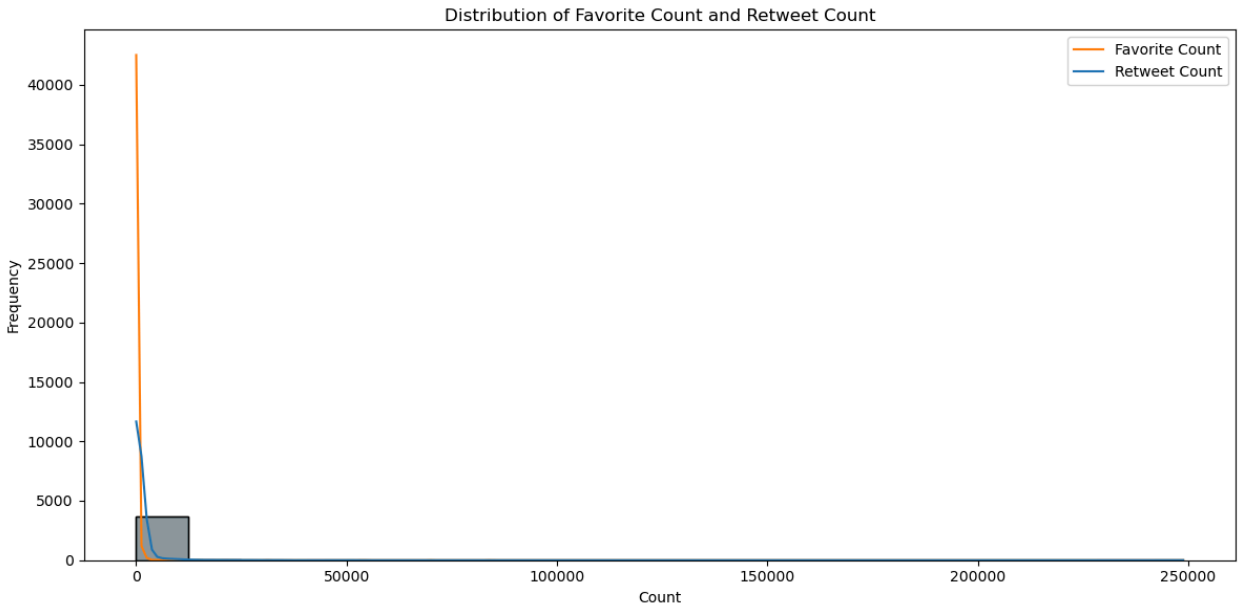

Figure 5 - Distribution of favourite and retweet count

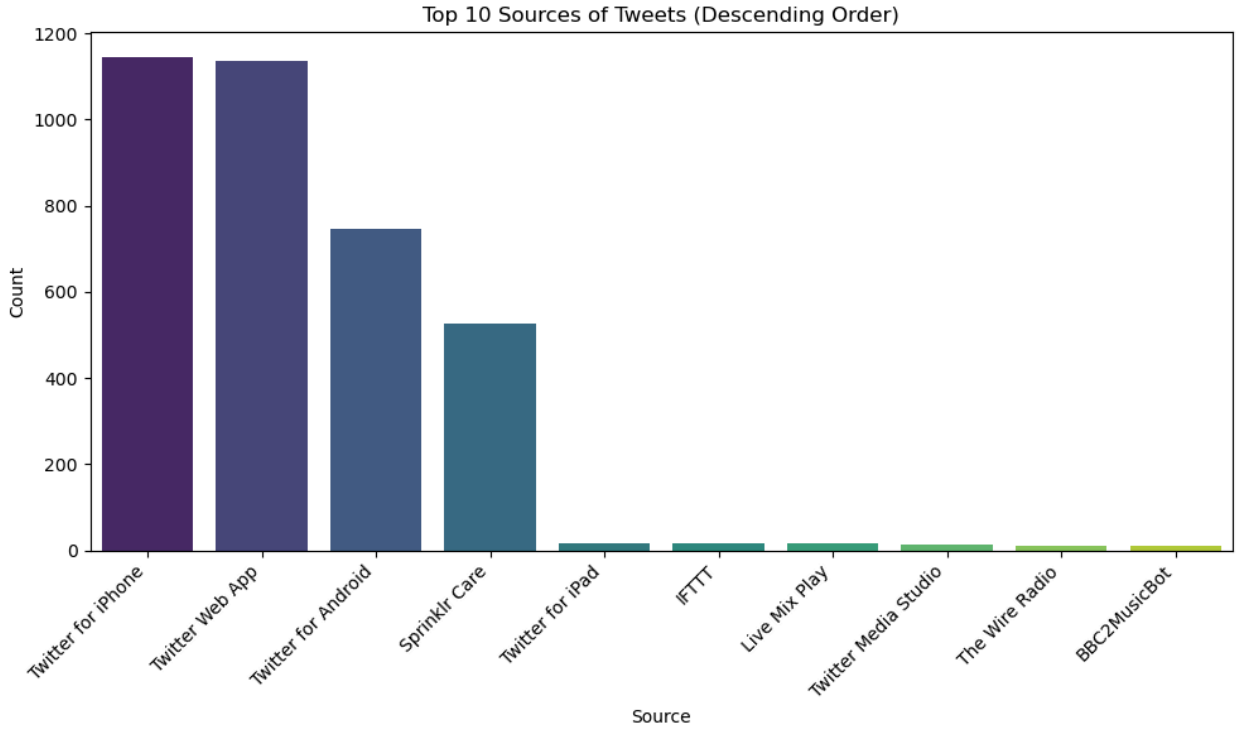

Figure 6 - Top 10 sources of tweets

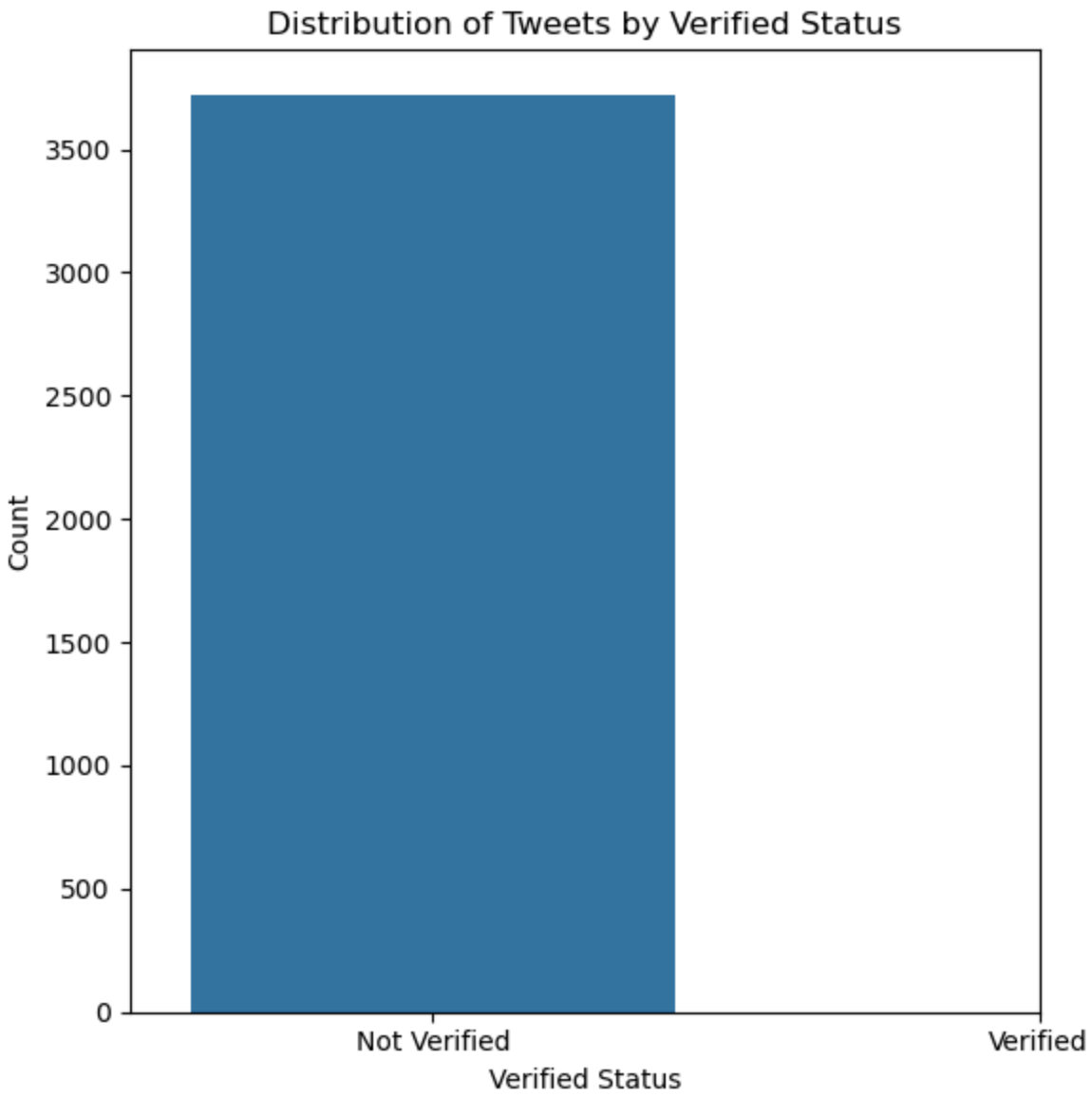

Figure 7 - Tweets by verified status

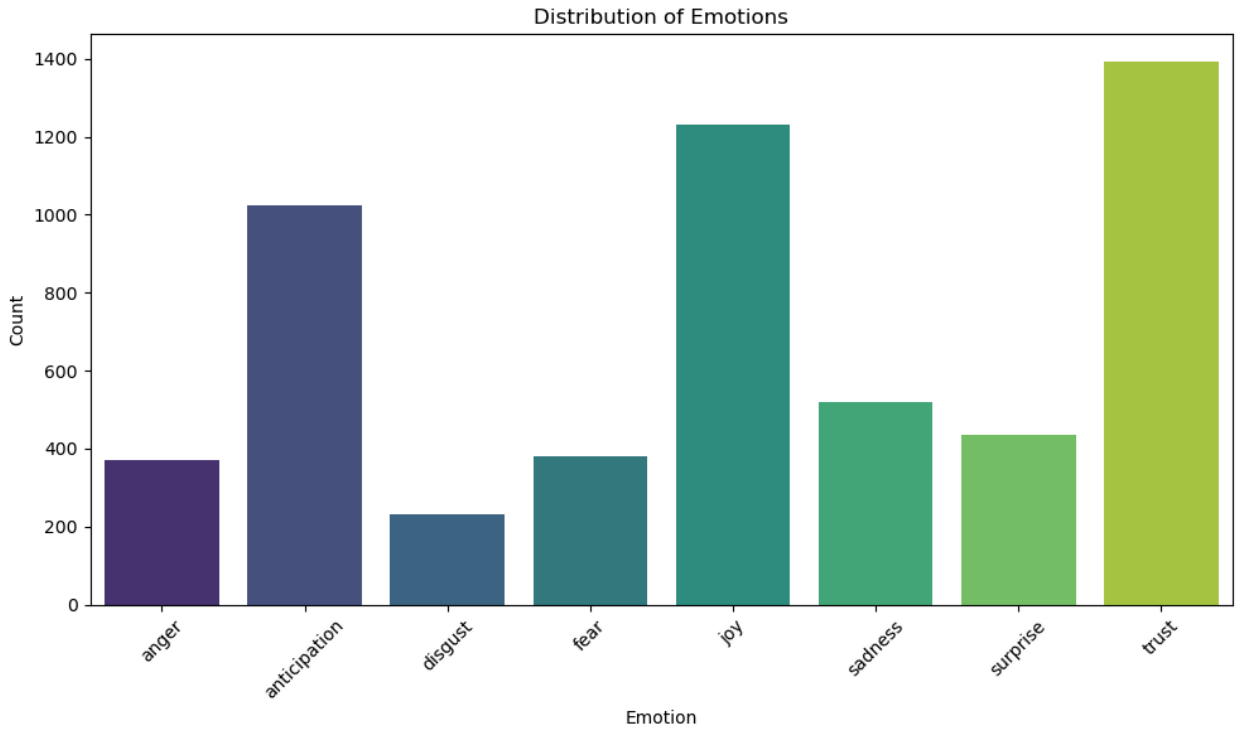

Figure 8 - Distribution of emotions

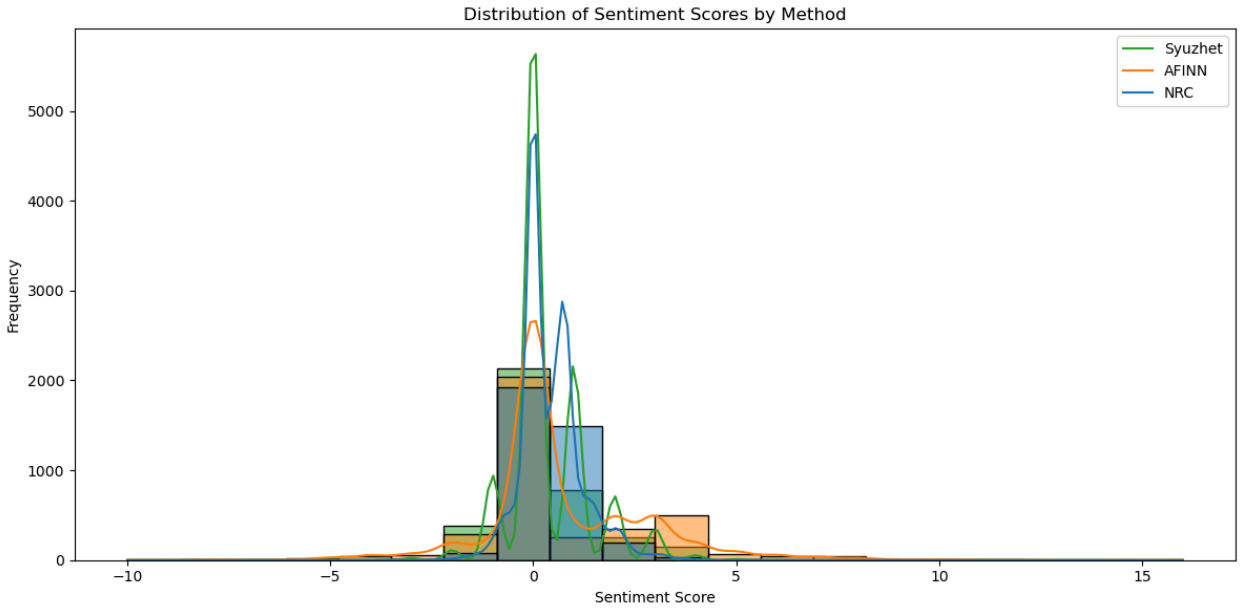

Figure 9 - Distribution of sentiment scores

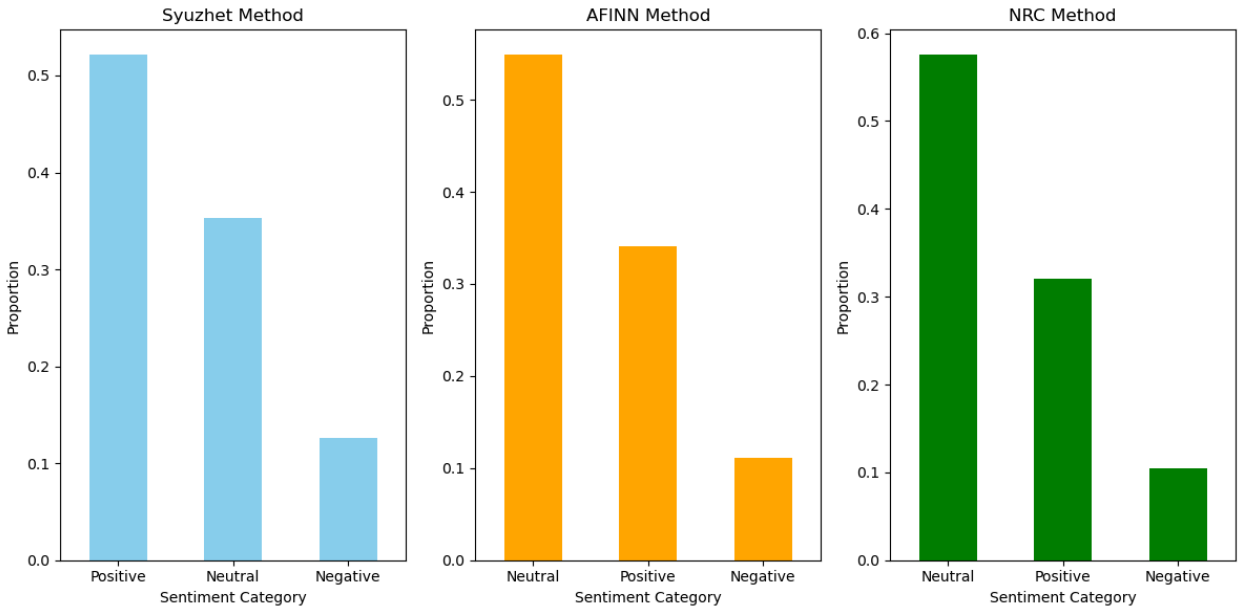

Figure 10 - Comparison of sentiment methods

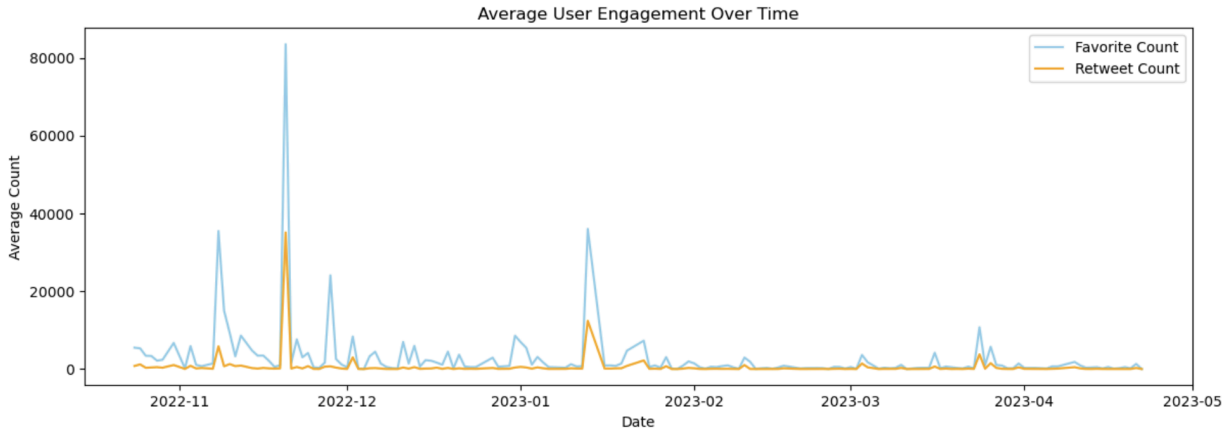

Figure 11 - Average user engagement over time

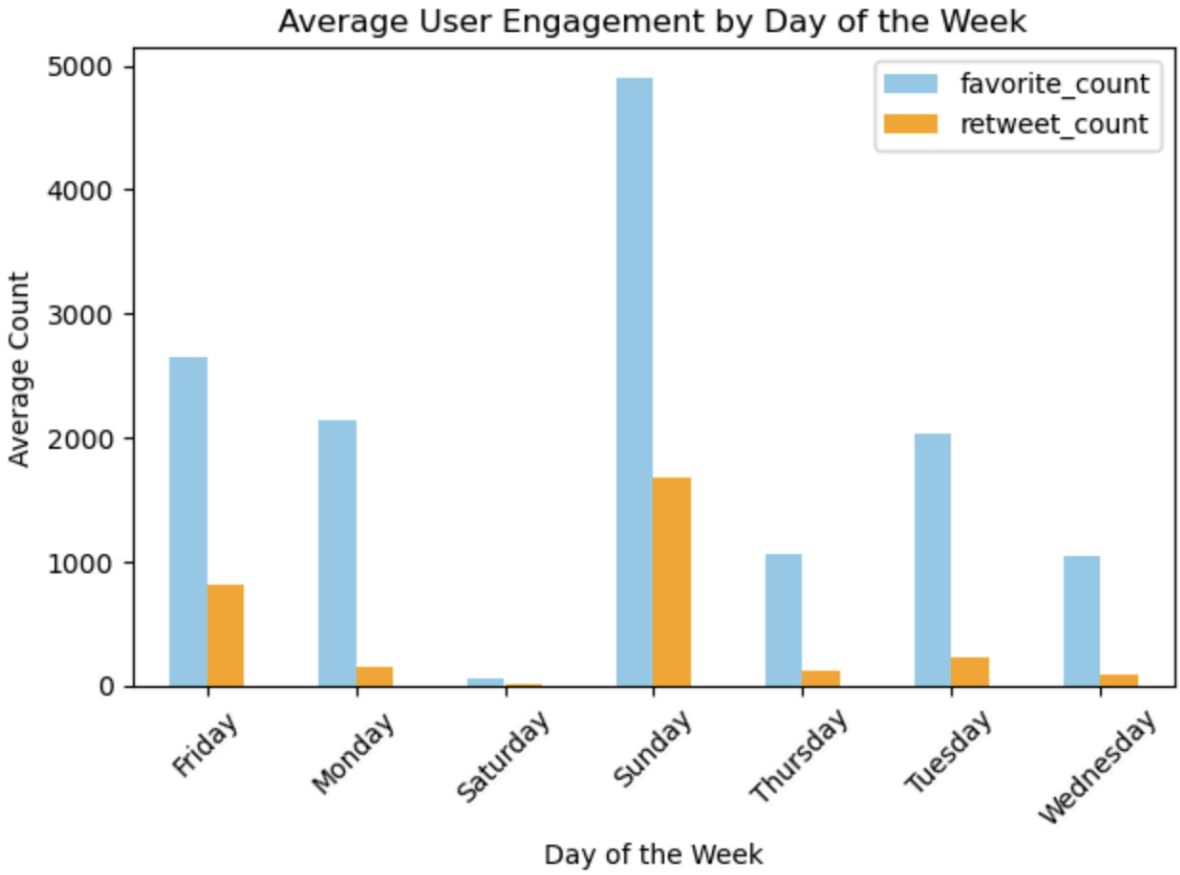

Figure 12 - Average user engagement by day

Figure 13 - Tableau visualisation