Social Media Analysis for Starbucks

A Executive Summary

The project aimed to conduct an exploratory analysis of tweets related to **Starbucks** in order to gain insights into customer opinions and preferences for Starbucks products. To achieve this goal, the dataset was collected over a period of 2 weeks using Twitter API and pre-processed to remove any irrelevant data. The cleaned data was then analyzed using various Natural Language Processing (NLP) techniques such as sentiment analysis, topic modeling, and word cloud generation.

The sentiment analysis revealed that the overall sentiment of tweets related to Starbucks was positive. The topic modelling identified several topics, such as sharing and positivity, gift cards and promotions, coffee and drink products and labor disputes. The word cloud generated from the tweets showed the most frequent words used by users in their tweets, such as "Starbucks", "gift card" and "olive oil". Furthermore, geospatial analysis revealed that the highest number of tweets originated from India and the United States, indicating the brand's popularity in these regions. The time-series analysis showed the highest volume of positive tweets occurred during evening hours.

Additionally, another goal of the project was to identify a micro-influencer whom the company can engage with to reach out to more audiences. The potential influencer was selected based on criteria such as verified user, popularity, positive tweet sentiment score, and active user of Twitter.

B Approach Breakdown

The first step in the analysis was to understand the dataset, which contained tweets related to Starbucks. The dataset had one row for each tweet and contained various features as described in data dictionary. The dataset was then checked for inconsistencies, duplicate values, and missing values and data cleaning was performed to ensure the quality of the dataset.

The next step was to perform exploratory data analysis on the cleaned dataset to gain insights about the tweets. For numerical features, statistical analysis was performed, including descriptive statistics to summarize the key features of the data. For categorical feature, using the location of the tweets geospatial analysis was carried out. Text analytics was then performed on the tweet feature containing text to understand sentiments towards the brand, the descriptive words that people relate with the brand and to uncover popular topics being talked about related to Starbucks. As part of text analytics, text cleaning, text pre-processing was first performed followed by sentiment analysis and topic modelling.

Furthermore, a new dataset was created at the user level, with each row representing a Twitter user instead of a tweet, to analyze user-level insights related to the individuals who tweeted about Starbucks. Finally, the user-level dataset was filtered based on selection criteria to identify potential micro-influencers for the company.

C Data Collection and Preparation

The data was collected using Twitter API and was stored in the csv (comma separated value) file. The dataset contained 1601 rows and 15 columns(features). The timeline of the tweets was from 22 February 2023 to 3 March 2023 and the location of tweets collected was global. The data was collected from 1268 unique users. The features collected about a tweet are shown in Table 1, while the details are same as described in data dictionary.

Table 1: Features and their data types

Features	**Data Type**
author_id, statuses_count, followers_count, friends_count, listed_count, favourites_count, tweet_id, retweet_count, favorite_count	Integer
location, text, created_at, lang, hashtags	String
verified	Boolean

- **Data Cleaning**: The collected data was first loaded into Python as a data frame and checked for duplicate rows, missing values, and inconsistencies in the data. It was found that the data set had no duplicate values, 476 missing locations, and non-English words/characters in the "location" and "text" columns. To address this, missing values and location values containing non-English words/characters in the "location" column were replaced with "Unknown". The data type of "author_id" and "tweet_id" was changed to "string" type as IDs are not used for numerical analysis. The data-type of 'created_at' was also changed from string to date-time type. The redundant feature "lang" was dropped as all values contained the same value and was not required for analysis. Additionally, the "hashtags" feature was dropped as it contained both text and indices, requiring cleaning, and was simpler to extract just the text of the hashtags directly from the "text" later on.
- **Feature Engineering**: New features were generated to gain additional insights about the data. Two new features, "text_nwords" and "text_nchars", were generated to describe the number of words and characters in the tweet text, respectively. Other features, "Hashtags" and "User_mentions" were created by extracting the hashtags and mentions of users from the "text" (tweet) column.

D Analysis of Brand

This section utilizes exploratory analysis and text analytics to investigate the brand. The exploratory analysis examines both numerical and categorical features, yielding valuable insights into metrics such as average likes and retweets per tweet, the timeline of tweets, average tweet length, and geospatial location. Meanwhile, text analytics provides insights into brand sentiment, related topics, and words commonly associated with the brand.

D.1 Exploratory Analysis

(i) Statistical Analysis for Numerical Features

The numeric features which are explored include seven features - 'statuses_count', 'followers_count', 'listed_count', 'favourites_count', 'retweet_count', favourite_count, 'text_nwords' and 'text_chars'. To summarise and understand these features, descriptive statistics like mean, median, standard deviation, min and max values of each numeric feature was calculated using describe() function. The distribution of each feature was then visualised using histograms and box-plots in Python notebook. It is observed that all the features except text_nwords and text_chars follow a skewed distribution. This implies that median is more appropirate measure of the central tendency. The following are the key observations from the statistical summary provided in Table 2:

- The average character length and average words length of the tweet is 168 and 21 respectively.
- The highest and lowest follower's count are 4,325,940 and 0 respectively.
- The maximum 'retweet_count' is 775, however, majority of tweets have 0 retweet count. This is evident as retweet count for 75% is 0.

Table 2: Statistical Summary

	count	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
statuses_count	1601	90318	328166	1	1096	6355	34326	2929390
followers_count	1601	11221	128478	0	49	363	1478	4325940
friends_count	1601	1357	3554	0	65	355	1357	67829
listed_count	1601	95	461	0	0	4	23	9412
favourites_count	1601	12064	35988	0	34	958	7429	540285
retweet_count	1601	1	19	0	0	0	0	775
favorite_count	1601	2	28	0	0	0	1	1051
text_nchars	1601	168	67	25	113	169	213	351
text_nwords	1601	21	10	1	13	21	26	55

(ii) Geo-spatial Analysis

The "location" column in the dataset was used for Geo-spatial analysis. After reviewing the dataset, it was observed that it contained various irrelevant values, such as text in other languages, numbers, email addresses, etc. To address this issue, text pre-processing techniques were employed to remove the irrelevant values and replace them with the term "Unknown". Once the dataset was cleaned, it was saved as a CSV file and imported into Tableau. Using the location names present in the cleaned dataset, corresponding geographical coordinates were obtained and used to plot the data onto a world map. Figure 1 shows the resulting geospatial distribution of tweet counts, with the size of the data points reflecting the frequency of tweets originating from each location. The analysis revealed that the highest number of tweets originated from India and the United States.

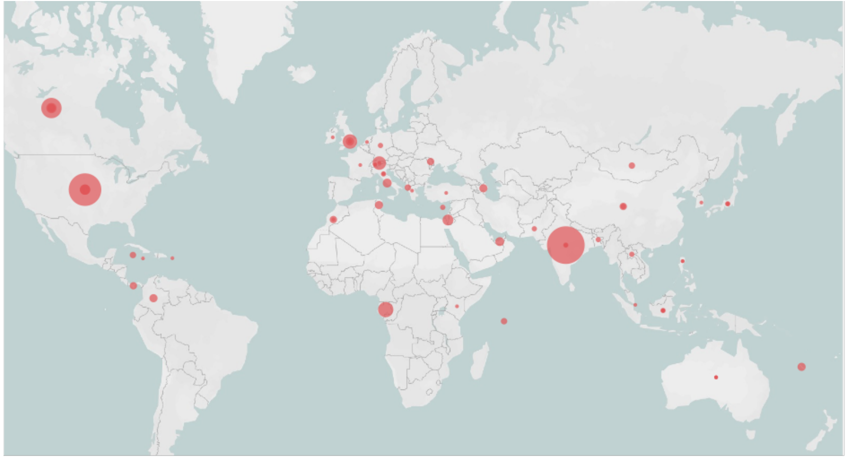

Figure 1: Geospatial Analysis of Tweets based on country location

(iii) Time-Series Analysis

The ‘created_at‘ column was used to perform time-series analysis of the tweets. To facilitate this analysis, the column was split into two new columns - ‘hour’ and ‘date’ - which reflect the time and date of the tweets, respectively. The tweet count was then plotted on an hourly basis as shown in Figure 2 and a daily basis, providing insights into the frequency of tweets over time.

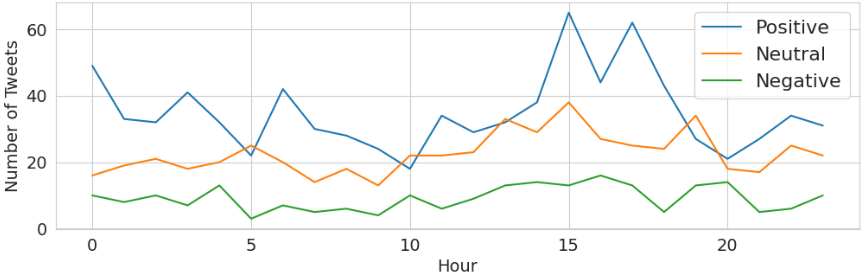

Figure 2: Time Series Analysis with sentiments

D.2 Text Analysis

(i) Text Pre-processing

To analyze the tweets, the following steps were taken as part of text cleaning and pre-processing:

- Language Detection - Removing non-English words from the text.
- Stripping - Removing whitespace and unusable characters from the text.
- Case Folding - Converting all letters to lower case
- Stopping - Removing stop words, which are common words that do not carry any semantic meaning.
- Lemmatization - Reducing words to their base or root form.
- Tokenization - Breaking down text into unigrams, bigrams and trigrams, which means breaking down the text into single words, pairs of words, and triplets of words.

(ii) Hashtag and User Mention Analysis

A consolidated list of hashtags from all users was created and analysed to identify unique hashtags, the total number of hashtags used, and count the frequency of each hashtag. This enabled the identification of the most commonly discussed topics. The same approach was used for analyzing user mentions in the tweets. The top hashtags and user mentions are “Starbucks” “Poshmarkapp” as seen from Figure 2 and Figure 3. Poshmark is a leading social marketplace. The high number of its mentions is due to the fact that “Poshmark” was selling the Starbuck’s products for discounted price during that time.

(iii) Sentiment Analysis

Sentiment analysis was performed to gauge people’s reactions towards the brand. To do this, transfer learning was utilized from pre-assessed text that was already classified into positive and negative sentiments since there was no ground truth to build our text analytics model. The NLTK library includes the pre-built VADER (Valence Aware Dictionary and Sentiment Reasoner) lexicon, a pre-trained tool that maps lexical features to emotion intensities and provides sentiment scores for commonly used words and phrases in social media. VADER was chosen due to its ability to handle social media text well. As shown in Figure 6, more than half of the tweets had positive sentiments.

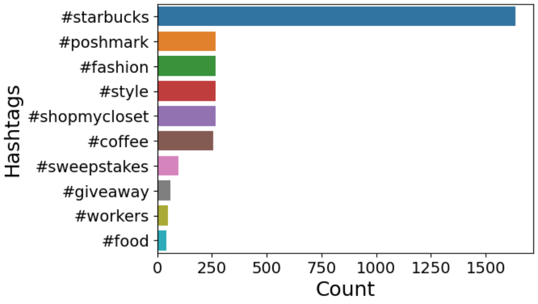

Figure 3: Top 10 Hashtags

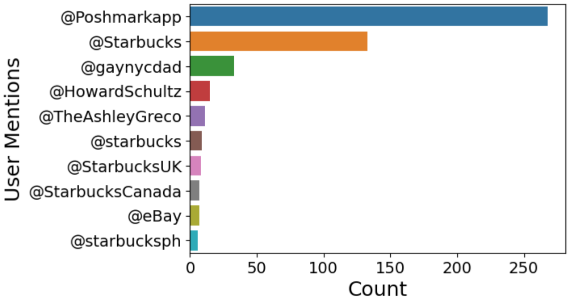

Figure 4: Top 10 User Mentions

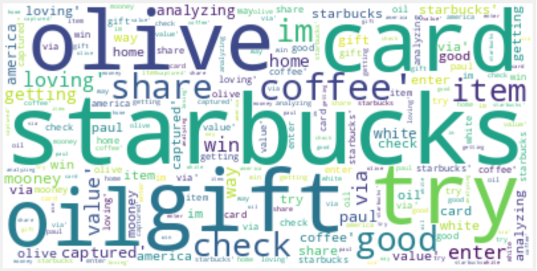

Figure 5: Word Cloud of all Tweets

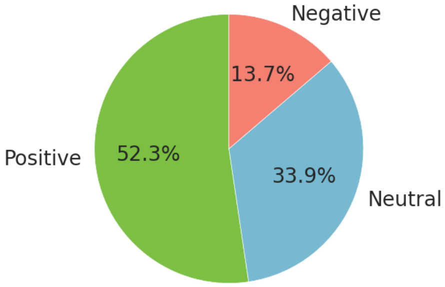

Figure 6: Sentiment analysis of tweets

(iv) Topic Modelling

Topic modelling was performed to uncover the topics being talked about by the users regarding the brand. The 4 major topics obtained after applying topic modelling are:

- **Topic 1: Sharing and positivity**, with words like "loving", "share", "good morning", & "post day". It can be related to the fact that majority of tweets are positive.
- **Topic 2: Gift cards and promotions**, with words like "gift", "win", "value", and "starbucks"
- **Topic 3: Coffee and drink products**, with words like "coffee", "starbucks", "olive", and "infused". It pertains to Starbucks' recent launch of Oleato beverages in Italy in February. These beverages are made by infusing coffee with Partanna extra virgin olive oil.
- **Topic 4: Legal issues and labor disputes**, with words like "union", "judge", "firing", "ruling", and "labor law". It relates to the recent news of union conflict with Starbucks. Specifically, Starbucks fired over 85 workers in the US who were involved in organizing worker unions against unfair labor practices

E Micro-influencer Recommendation

For selecting the micro-influencer, the tweet level data-set was transformed to user-level data-set, for getting insights at user-level. The user-level data-set was obtained by aggregating the features of the tweet level data-set thus reducing the granularity of the actual data-set. The following criterias were used for selecting the micro-influencer:

- **Verified user**: The influencer must be a verified twitter user, as it shows the genuinity of the user, so that people will have trust on the tweets from that account as the account is not fake.

- **Popularity**: The influencer should have large followers, more favorites count and re-tweet count per tweet to influence more number of people, hence it must have more than the average number of followers favorites count and re-tweet count per tweet of a user.
- **Positive Tweet Sentiment Score**: The influencer's tweets sentiment score must be above the average sentiment score of all users, which describes that the user has been tweeting more positively than the average.
- **Active user**: The influencer must be an active user and hence the number of tweets and number of statuses updated by user must be greater than the average value per user.

After applying the above criteria to filter the users from the user-level data-set, only one user with "author id" as 146958 was found who satisfied all the criteria. Therefore, this user can be considered as a potential micro-influencer. The user is located in "U.S., Canada, Puerto Rico", implying he has reach in these regions. The user has tweeted about the labor unions which has got maximum number of re-tweets as compared to other users. The other features of this potential micro-influencer are shown in the Table 3.

Table 3: Comparison of potential micro-influencer features against average features

User	Tweets	statuses_count	followers_count	listed_count	favorite_count	sentiment score
146958	2	64581	229398	2317	35	0.6
Avg. User	1.3	35065.7	11705.7	83.4	2.8	0.3

F Conclusion

In conclusion, the Twitter analytics project on Starbucks aimed to uncover insights related to brand recognition among people. Through exploratory data analysis and text analytics, various insights were uncovered related to people's sentiments towards the brand, popular topics being talked about, and users who tweeted about Starbucks. The overall sentiment towards Starbucks was positive. The Twitter users discussed several topics related to Starbucks such as promotions, coffee products, and labor disputes. The word cloud generated from the tweets highlighted the descriptive words that people often associated with Starbucks. Also, the highest number of tweets originated from India and the United States, indicating the popularity of the brand in these regions. Finally, based on user-level dataset and selection criteria, a micro-influencer was identified for Starbucks. The influencer was chosen based on their high engagement rates, a large number of followers, and a positive sentiment towards the brand. This influencer has the potential to positively impact the brand's image making them an ideal candidate for future marketing campaigns. Following are the recommendations for the company to extend this pilot study into a full analysis:

- **Increase the dataset**: The current dataset covers only a period of 2 weeks. To obtain more insights and better statistical significance, it is recommended to increase the dataset size by collecting tweets for a longer period of time.
- **Expand the analysis to other social media platforms**: While Twitter is a widely used social media platform, it is also recommended to expand the analysis to other platforms like Facebook and Instagram to gain a more comprehensive understanding of customer preferences and opinions.
- **Incorporating social network analysis**: The provided dataset does not provide any information regarding the social network. If the company tries to gather information about the social network which captures the intercation between users, then it will significantly help in identifying the potential micro-influencers who are well-positioned in the network who will have higher influence and reach.