LA TROBE
BUSINESS
SCHOOL

**BUS5PR1**

**ARTIFICIAL INTELLIGENCE AND
HYPERAUTOMATION**

**SEMESTER 1, 2023**

# ASSIGNMENT 3 - CAPSTONE PROJECT

**SPRINT 2**

**DELIVERED BY DREAM TEAM**

## 1. Introduction

In today's business environment, there has been a growing trend towards promoting healthy food and lifestyle choices, leading to the emergence of a new market segment. As a result, many companies have started targeting the lifestyle customer segment in their marketing strategies. In this report, we conducted an analysis of over 46,587 customer tweets that expressed concerns about food and lifestyle, using natural language processing techniques to identify valuable insights. The process involved pre-processing the data, exploring it thoroughly, and building models to derive useful recommendations for IntelliMed healthcare to leverage this emerging market segment.

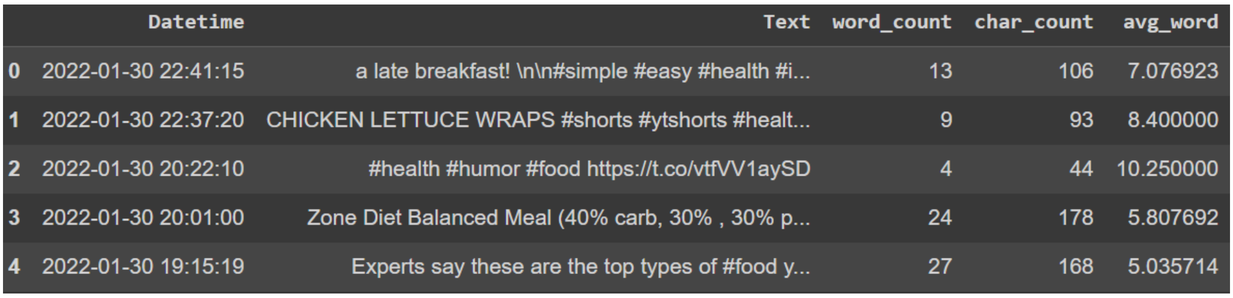
## 2. Data pre-processing

### 2.1 Preliminary analysis

We conducted a preliminary analysis of Twitter data related to food and lifestyle. Our process involved some initial data pre-processing and analysis, which included selecting relevant columns from the DataFrame, creating custom functions to calculate word and character counts, and applying these functions to the 'Text' column of the DataFrame to generate new columns for word count, character count, and average word length.

From the summary statistics, we can see that the number of words per tweet ranges from 4 to 27, with an average of 16.6 words per tweet. The number of characters per tweet ranges from 44 to 178, with an average of 117.8 characters per tweet. The average word length ranges from 5.0 to 10.3 characters, with an average of 7.6 characters per word.

Figure 1: Table showing the first 5 rows of the DataFrame

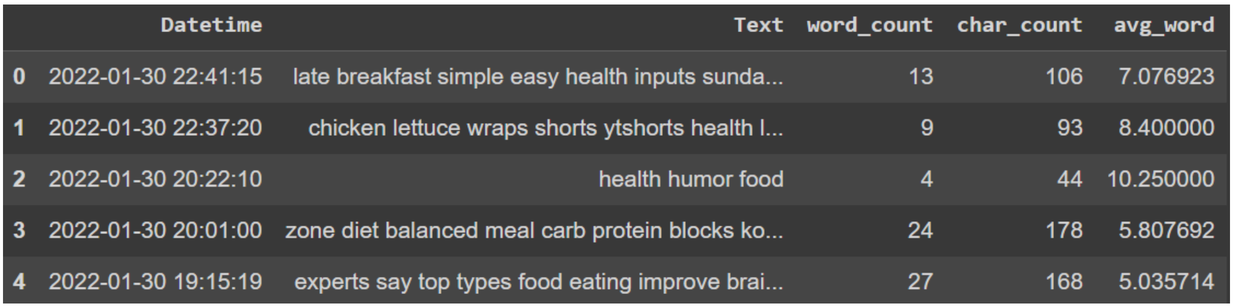


|   | Datetime | Text | word_count | char_count | avg_word |
|---|---|---|---|---|---|
| 0 | 2022-01-30 22:41:15 | a late breakfast! \n\n#simple #easy #health #i... | 13 | 106 | 7.076923 |
| 1 | 2022-01-30 22:37:20 | CHICKEN LETTUCE WRAPS #shorts #ytshorts #healt... | 9 | 93 | 8.400000 |
| 2 | 2022-01-30 20:22:10 | #health #humor #food https://t.co/vtfVV1aySD | 4 | 44 | 10.250000 |
| 3 | 2022-01-30 20:01:00 | Zone Diet Balanced Meal (40% carb, 30% , 30% p... | 24 | 178 | 5.807692 |
| 4 | 2022-01-30 19:15:19 | Experts say these are the top types of #food y... | 27 | 168 | 5.035714 |

## 2.2 Text pre-processing

We initially eliminated URLs present in the Text column by applying a regular expression library. Subsequently, all the text contained in the Text column is converted to lowercase in the second step. In the third step, non-alphanumeric characters are removed from the Text column. In the fourth step, digits present in the Text column are removed. The fifth step involves discarding all emojis from the Text column by referring to a dictionary of emojis and their corresponding meanings. To further refine the Text column, stop words present in the predefined list are eliminated in the sixth step. Lastly, any residual emojis present in the text column are removed in the final step.

Figure 2: Table showing the ‘Text' column cleaned up after removing any emojis



|   | Datetime | Text | word_count | char_count | avg_word |
|---|---|---|---|---|---|
| 0 | 2022-01-30 22:41:15 | late breakfast simple easy health inputs sunda... | 13 | 106 | 7.076923 |
| 1 | 2022-01-30 22:37:20 | chicken lettuce wraps shorts ytshorts health l... | 9 | 93 | 8.400000 |
| 2 | 2022-01-30 20:22:10 | health humor food | 4 | 44 | 10.250000 |
| 3 | 2022-01-30 20:01:00 | zone diet balanced meal carb protein blocks ko... | 24 | 178 | 5.807692 |
| 4 | 2022-01-30 19:15:19 | experts say top types food eating improve brai... | 27 | 168 | 5.035714 |

## 2.3 Common and rare word analysis

The analysis involved determining the frequency of words in the given text data. The figure shows the top 10 most commonly occurring words and the least common words that were found.

Figure 3: Top 10 most frequent words

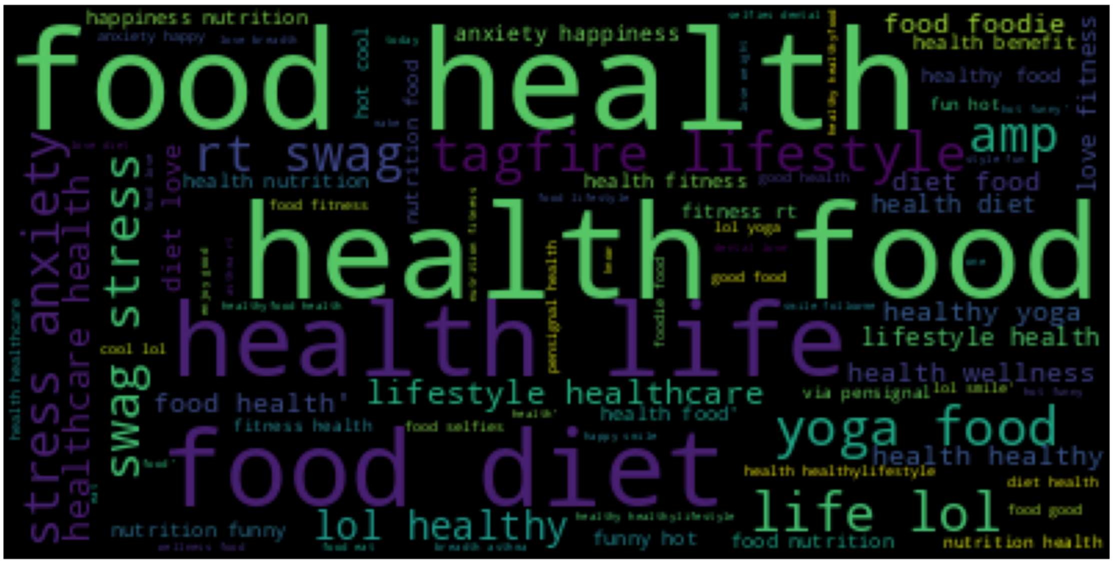
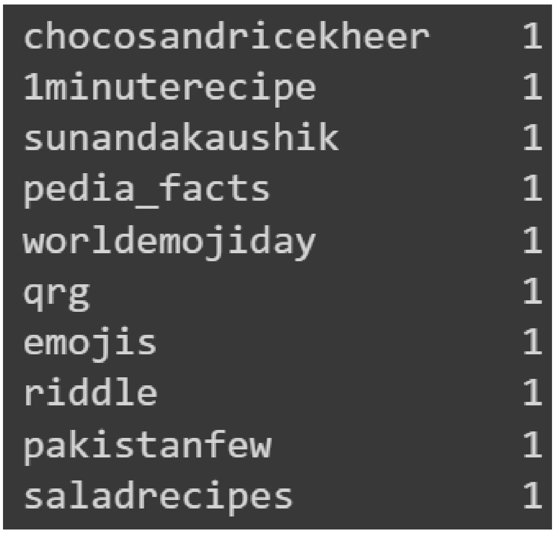


Figure 4: Table showing the top 10 least frequent words

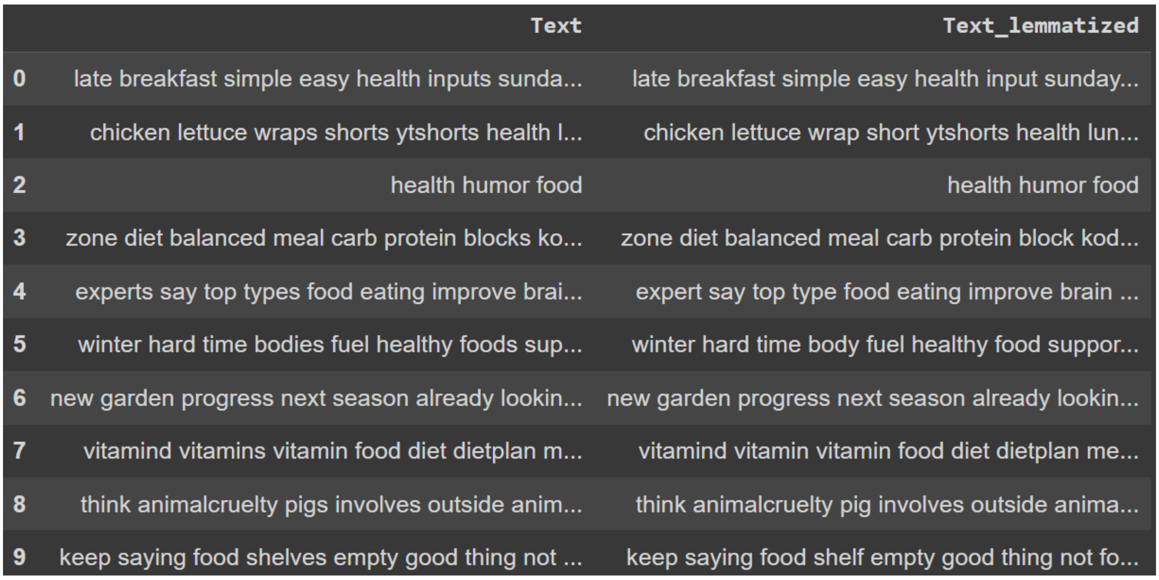


| | |
|---|---|
| chocosandricekheer | 1 |
| 1minuterecipe | 1 |
| sunandakaushik | 1 |
| pedia_facts | 1 |
| worldemojiday | 1 |
| qrg | 1 |
| emojis | 1 |
| riddle | 1 |
| pakistanfew | 1 |
| saladrecipes | 1 |

## 2.4 Lemmatization

In the first row, the word "inputs" was lemmatized to its base form "input" to standardize the text and reduce the number of unique words in the dataset. By lemmatizing the words, it was confirm that they were valid English words and to perform further analysis on them.

Figure 5: Table showing lemmatized text



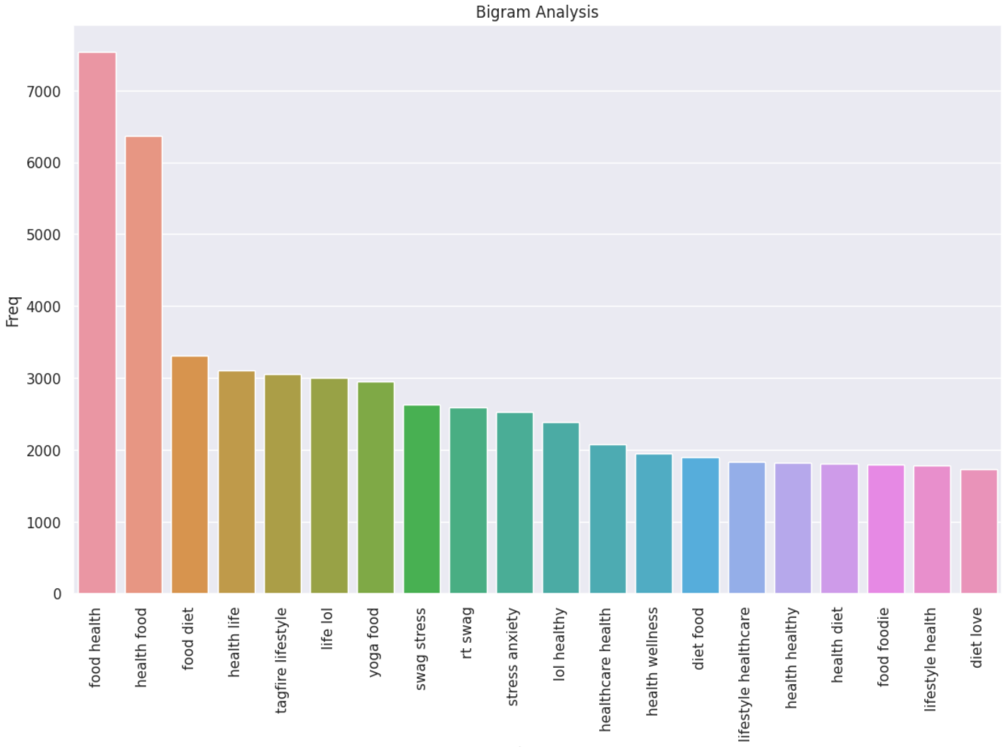
| | Text | Text_lemmatized |
|---|---|---|
| 0 | late breakfast simple easy health inputs sunda... | late breakfast simple easy health input sunday... |
| 1 | chicken lettuce wraps shorts ytshorts health l... | chicken lettuce wrap short ytshorts health lun... |
| 2 | health humor food | health humor food |
| 3 | zone diet balanced meal carb protein blocks ko... | zone diet balanced meal carb protein block kod... |
| 4 | experts say top types food eating improve brai... | expert say top type food eating improve brain ... |
| 5 | winter hard time bodies fuel healthy foods sup... | winter hard time body fuel healthy food suppor... |
| 6 | new garden progress next season already lookin... | new garden progress next season already lookin... |
| 7 | vitamind vitamins vitamin food diet dietplan m... | vitamind vitamin vitamin food diet dietplan me... |
| 8 | think animalcruelty pigs involves outside anim... | think animalcruelty pig involves outside anima... |
| 9 | keep saying food shelves empty good thing not ... | keep saying food shelf empty good thing not fo... |

## 2.5 Bi-grams and Tri-grams

We employed bi-grams and tri-grams analysis techniques to derive more meaningful insights from the text data, by identifying frequently occurring pairs and groups of words, which could potentially convey more significance than individual words on their own.

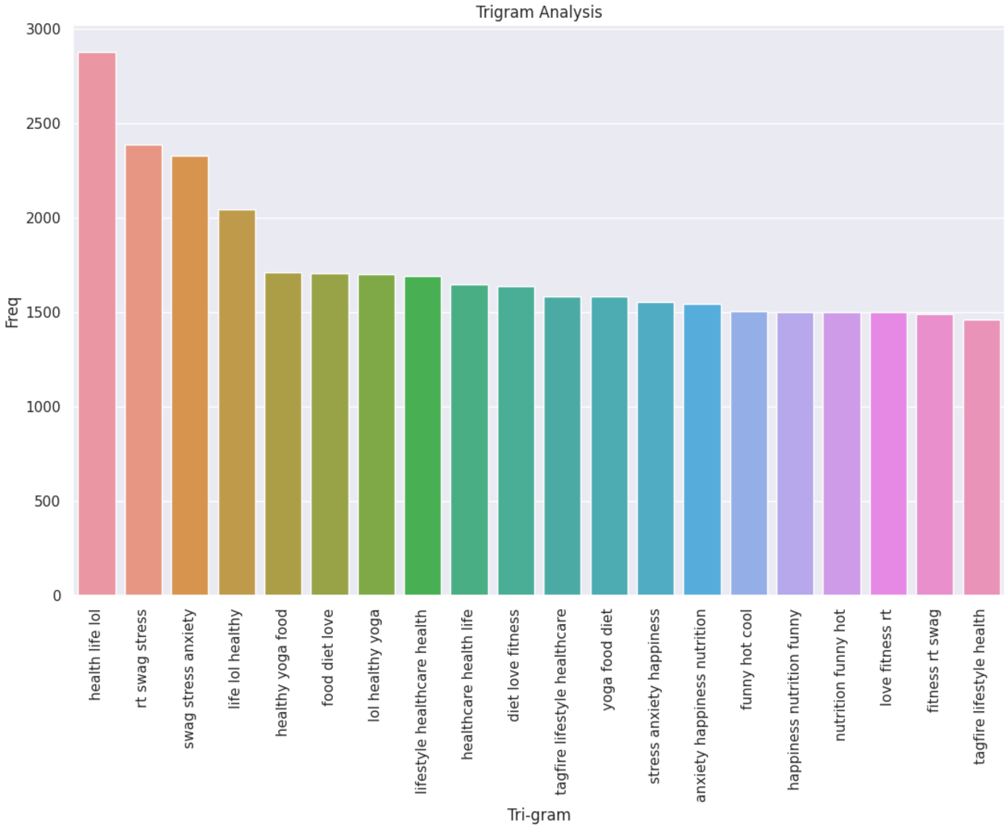
The analysis of bi-grams and tri-grams revealed that the most commonly occurring word pairs in the dataset were related to "food health" and "health food", suggesting a strong emphasis on maintaining a healthy lifestyle through dietary choices. Furthermore, the tri-grams analysis indicated a connection between health and lifestyle, interest in managing stress and anxiety, focus on healthy practices such as yoga/fitness and food, and a desire for a more positive and enjoyable relationship with healthy eating habits.

Figure 6: Bigram Analysis



.

Figure 7: Trigrm Analysis



### 2.6 Bags of Words (BoW)

We adopted the BoW technique, which allows the text data to be transformed into a numerical format that can be used in machine learning algorithms. The BoW technique involves several steps, including tokenization, lemmatization, and building a vocabulary of all unique words in the dataset. The resulting BoW representation will be used for our task such as sentiment analysis.

### 2.7 TF-IDF (Term Frequency-Inverse Document Frequency)

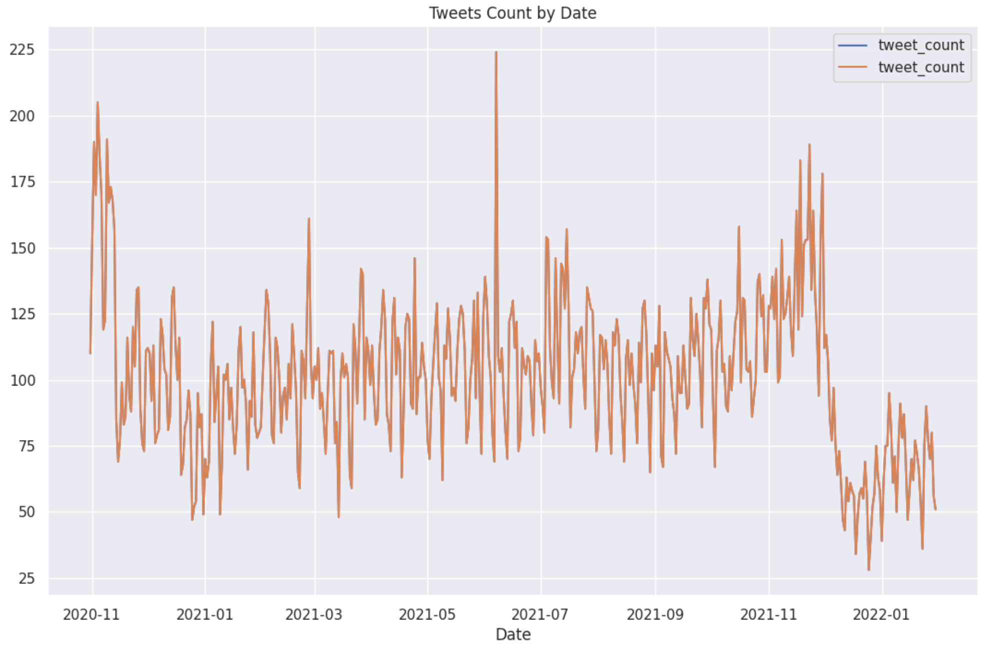
After Bow, we utilized lemmatized sentences as our text data and applied the TF-IDF technique to improve the analysis, which is a weighted scheme that considers not only the word frequency in a document but also its frequency across all documents in the corpus. This approach enabled us to identify words that are significant and pertinent to a particular document.

## 3. Temporal Analysis

### 3.1 Date Analysis

We started the process by importing the datetime module and using lambda functions to split the original 'Datetime' column into 'Date' and 'Time' columns. After that, we dropped the original 'Datetime' column and reduced the resulting dataframe to only include the 'Date' and 'Text' columns. We then grouped the dataframe by 'Date' and calculated the count of tweets per day, which we saved as 'tweet_count'. We used matplotlib to plot the resulting dataframe 'Tweets Count by Date', as shown below the figure.

Figure 8: Tweets Count by Date



Based on the chart depicting the tweets count by date, we can deduce several insights. Firstly, the tweet activity showed a gradual decline from November 2020 to June 2021, with a peak of 225 tweets occurring in June 2021. After the peak, there was a fluctuation in tweet activity, with the number of tweets ranging between 70 to 160 tweets until November 2021, when the count rose to 190. Subsequently, there was a significant drop in tweet activity, and the number of tweets ranged from 25 to 100. These findings suggest that the analyzed Twitter data experienced a period of growth followed by a decline, with fluctuations in between.

Our recommendation for IntelliMed's entry into this new market is to leverage the insights gained from the analyzed Twitter data to gain a better understanding of the market and consumer trends. The gradual decline in tweet activity from November 2020 to June 2021 could suggest a lack of interest in existing health products or solutions. However, the peak in June 2021 and the subsequent fluctuations in tweet activity indicate that there are moments of heightened consumer interest.

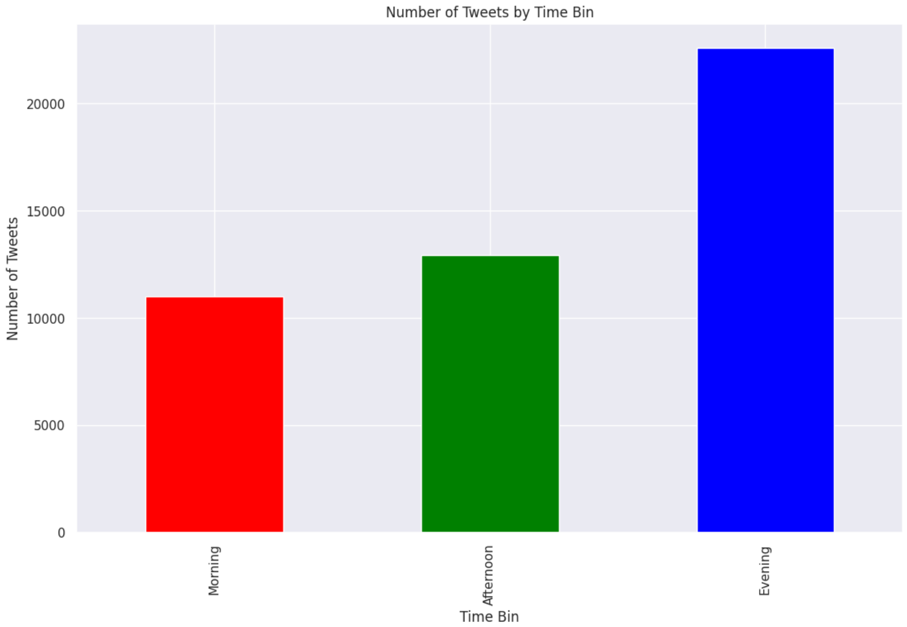
To take advantage of these trends, IntelliMed could focus on creating innovative products or solutions that address the current gaps in the market, particularly during periods of increased consumer interest. Additionally, they could monitor social media activity to track trends and stay up-to-date with consumer preferences and needs, which could inform their product development process and ensure that their offerings align with market demands.

Another possible recommendation for IntelliMed is to investigate the reasons for the fluctuations and declines in tweet activity. This could involve conducting sentiment analysis or tracking related news events to see if there were any external factors affecting Twitter activity. Understanding these factors could help IntelliMed better tailor their marketing and outreach efforts to the needs and interests of their target audience. Additionally, IntelliMed could explore other social media platforms or online communities to engage with potential customers and gather feedback on their products.

## 3.2 Time Analysis

Based on our analysis, we can observe the distribution of tweets across different time periods during the day. We converted the 'Time' column to datetime format and created a new column 'Time Bin' based on the hour of the tweet. The tweets were grouped by their time bin and the number of tweets in each bin was counted. We plotted a bar chart as below, to visualize the distribution of tweets across different time bins. Our analysis reveals that the peak time for tweeting is in the evening, followed by the afternoon and then the morning. As a result,  IntelliMed can use this information to optimize our social media strategy by scheduling tweets during the peak hours to reach a larger audience and increase engagement.

Figure 9: Number of Tweets by Time Bin



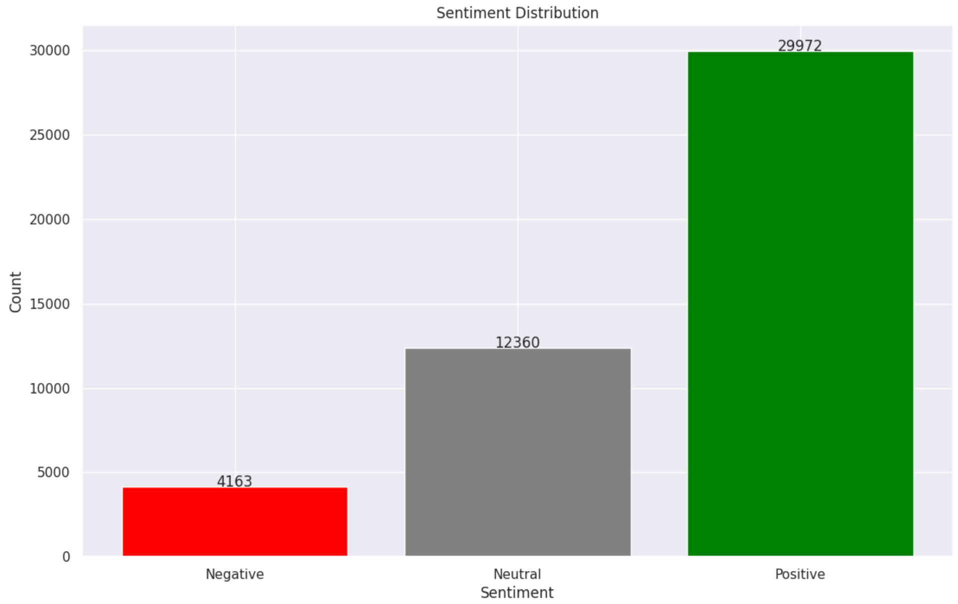
## 4. Sentiment Analysis

In the sentiment analysis, we started by applying the TextBlob library to calculate the polarity scores for each tweet in the 'Text_lemmatized' column. We then categorized these scores into 'Negative', 'Neutral', and 'Positive' by using a loop to convert the polarity scores into these categories. Afterward, we created a new 'sentiment' column in the dataframe to store these sentiment categories for each tweet. We then grouped the data by sentiment category and counted the number of entries in each group to get an idea of the sentiment distribution of the tweets. Finally, we created a bar plot, as shown below, using the sentiment counts and defined colors for each sentiment category. The resulting plot provides an overview of the sentiment distribution of the tweets, with the number of tweets falling into each sentiment category.

The insight gained from the sentiment analysis is that the majority of the tweets are positive, with 29972 tweets having a positive sentiment category compared to only 4763 tweets with a negative sentiment category. Additionally, there were 12360 tweets categorized as neutral.

Based on the sentiment analysis of the Twitter data, we recommend that IntelliMed continue to focus on creating positive and neutral content on their social media channels. The sentiment analysis reveals that there are significantly more positive and neutral tweets than negative tweets related to health and food topics. This information can also be leveraged by IntelliMed to continue to create content that resonates with their audience and helps to build a positive reputation.

Figure 10: Sentiment Distribution



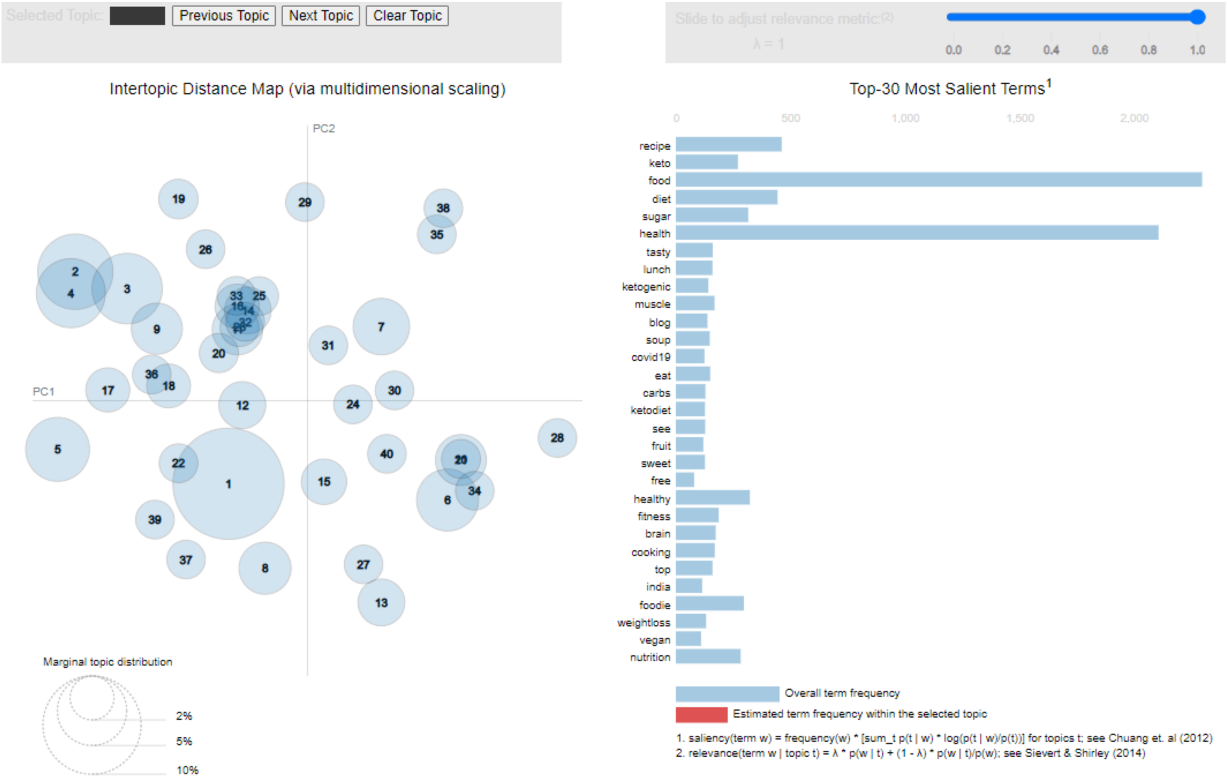
## 5. Topic Modeling

We started by cleaning and preprocessing the tweets, removing any unwanted noise or information to obtain a clean dataset. Next, we constructed a dictionary of words from the cleaned tweets and created TF-IDF features to represent the most important words in the tweets. These features were then transformed into feature vectors, which helped us to better understand the patterns and relationships between the words in the tweets.

Using Latent Dirichlet Allocation (LDA), we were able to identify the main topics present in the tweets and their respective keywords and weights. We interpreted these keywords to understand what each topic was about, enabling us to gain deeper insights into the underlying themes and sentiments expressed in the tweets.

Finally, see the chart below, we used an interactive topic analyzer to visualize and explore the topics further, allowing us to see how the different topics were related and how they evolved over time. Through this process, we were able to extract valuable insights from the tweet data, providing us with a better understanding of the topic and its impact on the online community.

Figure 11: Interactive Topic Analyzer



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

The interactive topic analyzer can generate several insights that can help us better understand the topic of interest. For example, it can help us identify the most prominent topics and keywords in the tweets, and how they are related to each other. We can also explore the frequency of different topics over time, and how they may have evolved or changed.

Furthermore, the analyzer can help us identify the sentiment of the tweets related to each topic, such as whether they are positive, negative, or neutral. This information can be useful in understanding how people are reacting to the topic and the underlying emotions associated with it.

Additionally, the analyzer can provide us with insights into the most active users or influencers in the conversation, as well as the most frequently used hashtags or mentions. This can help us identify key players in the online community and better understand their impact on the discussion.

Overall, the interactive topic analyzer can generate a wealth of insights that can help us gain a deeper understanding of the topic, its impact, and the sentiment and behavior of the online community discussing it.