**DATA UNDERSTANDING AND DATA PREPARATION**

**INTRODUCTION**

The Tweet2GPT project aims to leverage Natural Language Processing (NLP) to improve train maintenance strategies, with a particular focus on analysis of Thameslink-related tweets. In this report, the Data Understanding and Data Preparation stages of the project, which is a joint work of 'The Fine Tuners' team, are examined.

Our main goal in this section of the report is to examine the processes involved in converting raw, unstructured Twitter data into a refined format suitable for NLP techniques. This includes preparatory steps to examine the data, assess its quality, and ensure it is ready for analytical procedures. Future sections will outline the methodologies and strategies used in these stages and provide a clear roadmap of our approach, from data collection to transformation.

**DATA UNDERSTANDING**

**Data Structure**

The critical step in understanding any data set is understanding its structure. The Tweet2GPT project's dataset is collected from two different platforms: Brandwatch and Sprinklr.

The dataset consists of 16,949 rows and 20 columns, representing a significant amount of data that needs to be work on. This data is in JSON (JavaScript Object Notation) format, which is typically used to store and transport data.

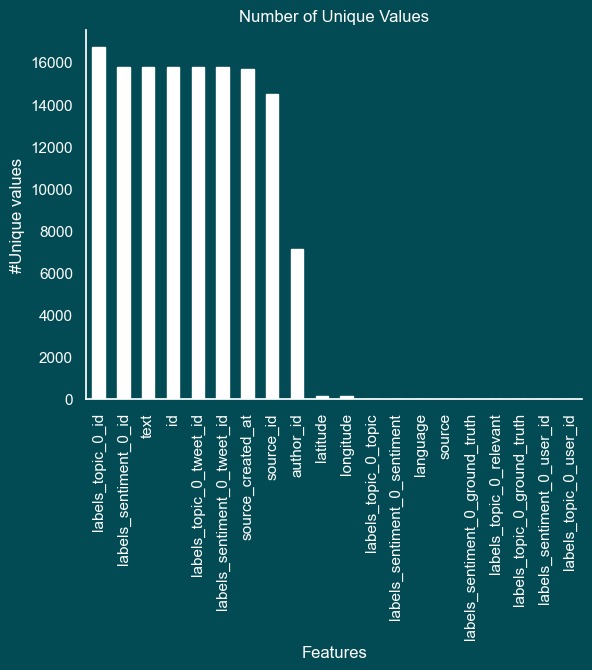
The dataset is diverse in terms of data types. It includes a Date/Time column, which is crucial for analysing tweets chronologically. The dataset also contains four categorical, eight numeric, and one string data type, along with six columns with constant values. This mix of data types provides a well-rounded view of information.

But the dataset is not without its challenges. One notable issue is data quality, with 2,219 instances of duplicate entries. Additionally, there are three columns with missing values.

Another aspect of data structure is its organisation which is a flat file.

**Column Analysis**

In-depth analysis of individual columns within the dataset is a necessary step in understanding the data.



**Graph 1: Analysis of Key Columns – Number of Unique Values**

Six of the 20 columns shown in Graph 1 were chosen to extract relevant and understandable information. These columns include 'text', 'author\_id', 'labels\_sentiment\_0\_topic', 'source\_created\_at', 'labels\_topic\_0\_topic', and 'source'.

Text: This column contains the actual tweet content and is the primary data source for NLP analysis.

Author\_id: Identifier of the author of the tweet and is important for understanding the demographics of the dataset.

Labels\_sentiment\_0\_topic and Labels\_topic\_0\_topic: These columns provide pre-categorised sentiment and topic labels.

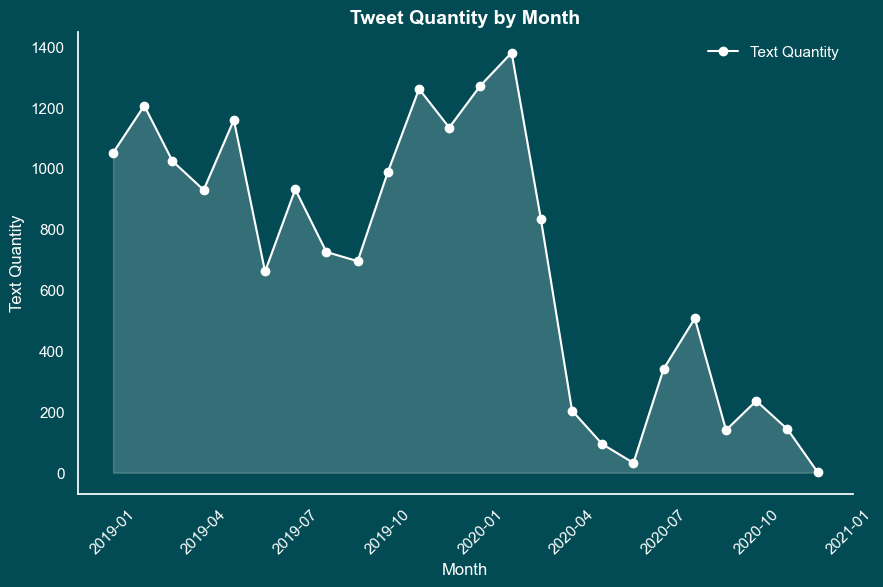
Source\_created\_at: Timestamp of when the tweet was published. This temporal data is essential for trend analysis and understanding feedback timing.

Source: The source of the tweet, whether it was sourced from Brandwatch or Sprinklr.

The text column, containing around 17,000 tweets, is particularly important for extracting meaningful patterns and themes from the public's conversations about Thameslink.

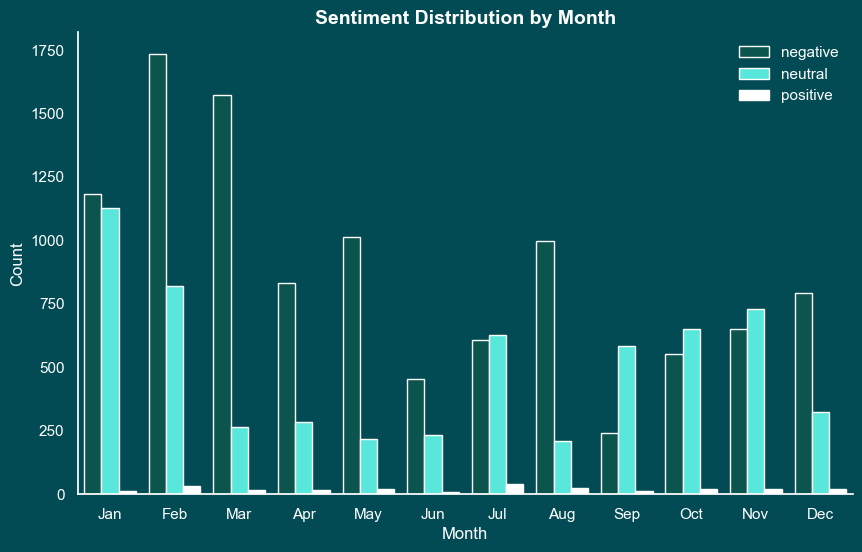
**Time Analysis**

The Thameslink tweet dataset covers the period from the beginning of 2019 to the end of 2020.



Graph 2: Tweet Distribution Over Time – Tweet Quantity by Month

This line Graph, called Graph 2, describes the monthly frequency of tweets. This chart helps identify significant spikes or decreases in Tweet volume.



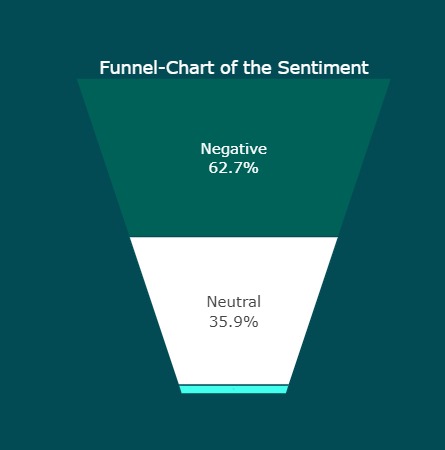
Graph 3: Sentiment Analysis Over Time – Sentiment Analysis by Month

By examining the sentiment expressed in tweets month by month, sentiment analysis over this two-year period can reveal how the public's perception of Thameslink services has changed.

An important observation from this analysis is the consistency of data collection across the twelve months of the year. This coverage provides a more accurate understanding of customer sentiment at different times of the year, minimising the potential for seasonal bias.

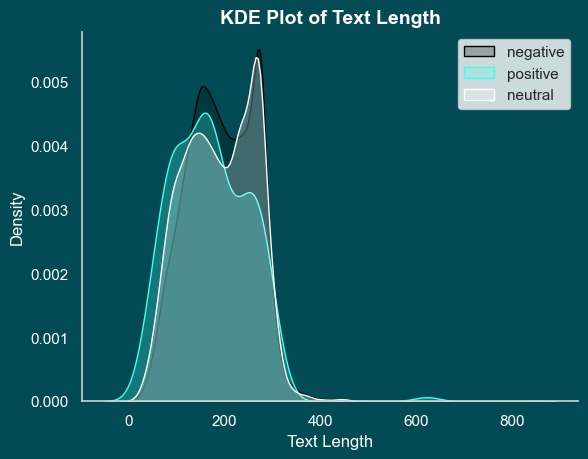
**Sentiment Analysis**

A key component of the dataset is the 'Sentiment' column, which divides tweets into three categories: negative, positive or neutral. This classification is important for understanding the public's perception and response to Thameslink services.



Graph 4: Sentiment Distribution – Funnel-Chart of the Sentiment

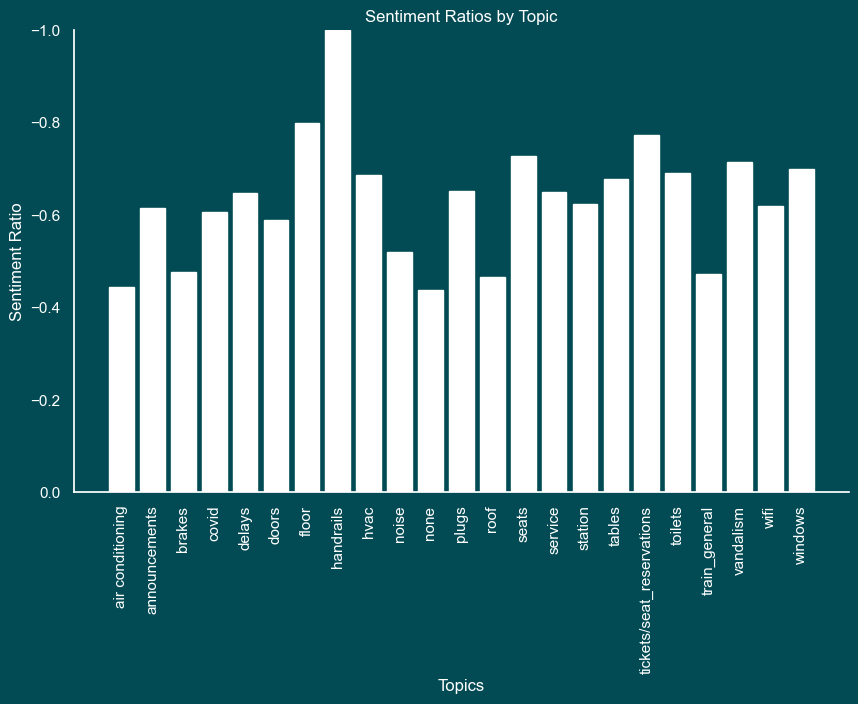
With this funnel chart, we gain insight into general sentiment trends among Thameslink customers by analysing the proportions of negative, positive and neutral sentiments in the dataset.



Graph 5: Sentiment by Text Length – KDE Plot of Text Length

A notable trend observed in the dataset is the correlation between negative sentiments and longer text posts. KDE Plot visualises this trend and highlights how negative emotions such as anger or frustration often show up in longer, detailed tweets.

The analysis of sentiments in relation to different topics mentioned in the tweets provides a more granular understanding of customer feedback. This approach enables the identification of specific areas of service that are sources of satisfaction or dissatisfaction among customers.

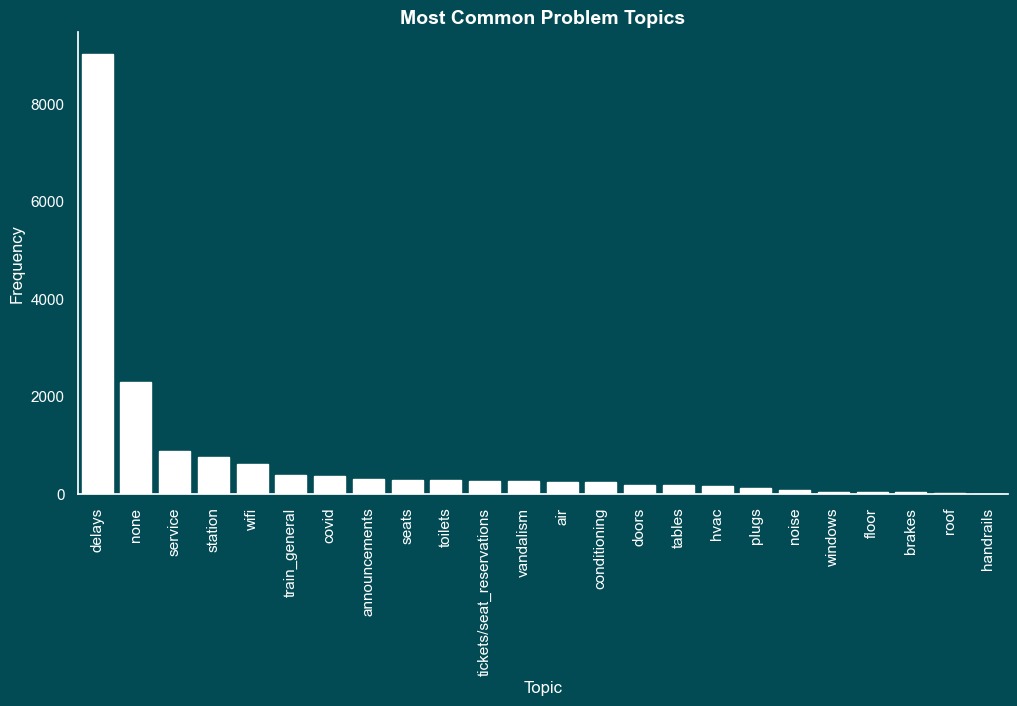


Graph 6: Average Sentiment Across Topics – Sentiment Ratios by Topics

This bar chart visualises how sentiment averages vary based on different topics discussed in tweets. The Graph also indicates aspects of Thameslink services that cause positive or negative reactions.

**Topic Classification and Analysis**

An important aspect of the dataset is the inclusion of categorical annotations for tweet topics. This classification provides a more structured and focused analysis of customer feedback

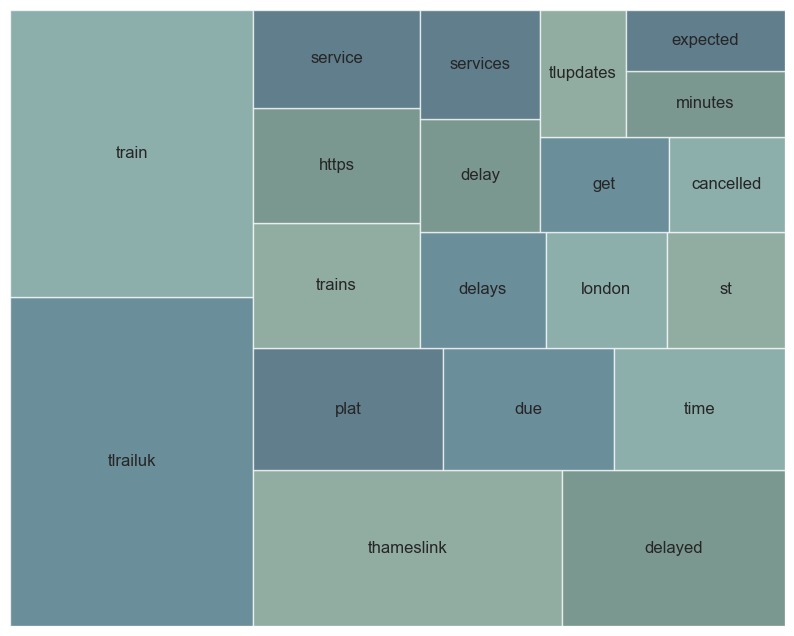
. 

Graph 7: Common Topics in Tweets – Most Common Problem Topics

This bar chart visualises the frequency of topics mentioned in the dataset. By identifying the most common topics we will not only be able to more systematically analyse Thameslink customers' tweets, but also understand the nuances of customer feedback.

**Text Analysis: Word Frequency**

A key component of understanding the textual content of the dataset involves analysing the frequency of word usage. This analysis helps in identifying the dominant terms and themes in the tweets, providing insights into the main issues or topics of interest for Thameslink customers.

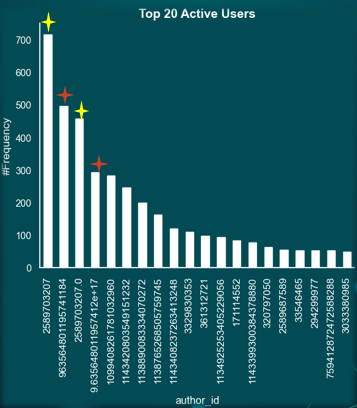


Graph 8: Visualisation of Frequently Used Words

The prominence of certain words can reveal customers' primary concerns and topics of interest. For example, the chart features several variations of the term “delay,” including “delay,” “delays,” and “delayed,” demonstrating the importance of this issue in customer feedback.

**Analysis of Author ID**

A significant aspect of the dataset is the identification of the most active users based on the Author ID. Understanding the activity levels of different users is important for distinguishing between regular customer feedback and official communications from the company.

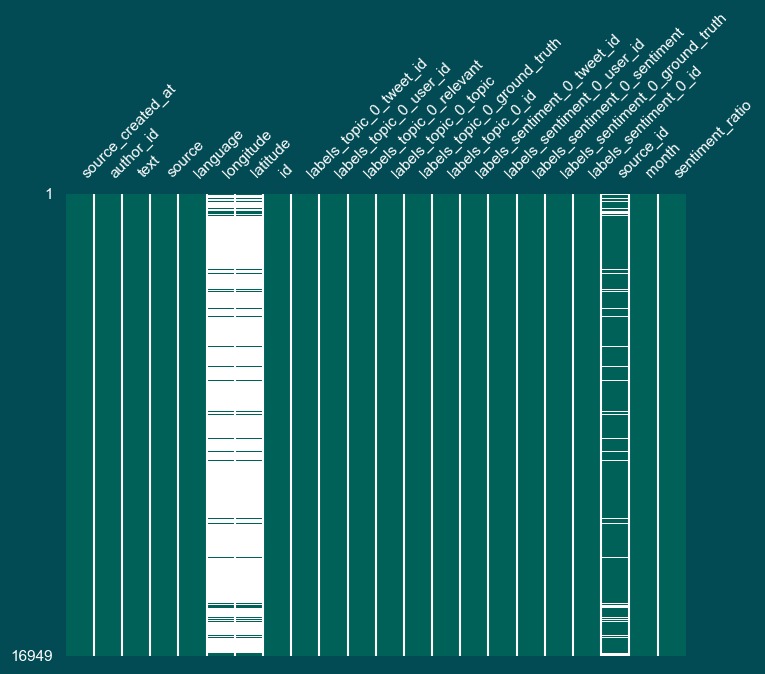


Graph 9: Top 20 Active Users

This Graph provides a clear visualisation of the users who are most active in terms of tweeting about Thameslink. More importantly, it highlights those accounts that serve as official channels for Thameslink communications. Users marked with yellow stars typically use the hashtag #TLUpdates, while those marked with red stars frequently use #ThameslinkUpdate. These accounts helps in segregating the dataset into tweets that are official communications from Thameslink and those that are customer feedback.

**Missing Values Analysis**

One of the challenges in data analysis is handling missing values. Identifying and addressing missing data is important for maintaining the quality and reliability of the analysis.



Graph 10: Missing Values in Each Column

This Graph provides a visual representation of missing values across different columns, highlighting three columns in particular: Longitude, Latitude, and Source ID. The missing values are notably high in the Longitude and Latitude columns. Understanding the implications of these missing values is important for the subsequent stages of data preparation.

**Duplicates and Outliers**

Another crucial aspect of data preparation is the identification and handling of duplicates and outliers. These can significantly affect the conclusions drawn from the analysis if not addressed properly.

The dataset contains 2,219 duplicate entries and 37 retweets. These duplicates need to be addressed to prevent redundancy and potential biases in the analysis. Additionally, the box plot reveals 11 outliers in text length, with some texts having a length of over 800 characters. Outliers in text length can indicate unusual posting behavior or data entry errors, and their handling is crucial for maintaining the dataset's consistency.

The identification of these duplicates and outliers is a key step in ensuring the accuracy and reliability of the analysis.

**DATA PREPARATION**

**Launching Data Preparation**

Data Preparation is a vital step in ensuring the quality and relevance of data for analysis. In the Tweet2GPT project, it is decided to use entire dataset for the analysis without restricting it to a specific time period or subset. This approach was chosen to ensure a comprehensive examination of all available data, providing a complete overview of customer feedback and sentiments across different timeframes and topics.

**ETL Process Enhancements**

The Extract, Transform, Load (ETL) process underwent several enhancements to improve the dataset's quality and relevance for analysis.

|  |  |  |
| --- | --- | --- |
| **Findings/Problems** | **Decision** | **Impact** |
| Unreadable Column Names | Shorten and Standardise | Improved Data Clarity |
| Data Types | Fixed Data Types | Enhanced Data Consistency |
| Redundant Time Information | Remove Second and Millisecond Information | Reduced Data Volume and Improved Data Quality |
| Singleton and Junk Columns | Removed 14 Columns | Simplified Data Structure and Reduced Data Redundancy |
| Duplicate Rows Based on Text Column | Removed 2,219 Rows | Enhanced Data Relevance for Maintenance Improvements |
| Tweets Sent by Thameslink Accounts | Removed 9 Authors | Enhanced Data Purity and Content Relevance |
| Delay Related Topics | Removed 7,645 Rows | Reduced Noise and Enhanced Data Relevance |

Table 1 : Summary of ETL Process Enhancements

*Unreadable Column Names*: These column names are shortened and standardised to improve data clarity.

*Data Types*: Corrections were made to ensure enhanced data consistency.

*Time Information*: Seconds and millisecond information were removed to improve data quality.

*Singleton and Junk Columns*: A total of 14 singleton or junk columns were removed to simplify the data structure.

*Duplicate Rows*: 2,219 duplicated rows were removed to reduce redundancy and enhance data relevance.

*Tweets from Thameslink Accounts*: As part of the project architecture, tweets from Thameslink accounts were removed to focus on customer-generated content.

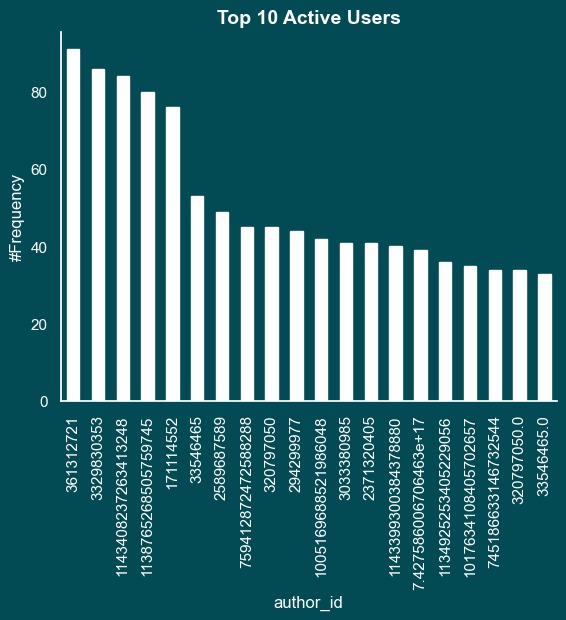
*Delay and None Topics*: A total of 7,645 rows related to these topics were removed to reduce noise and enhance the relevance of data for maintenance improvements.

Each of these steps in the ETL process was critical to improve the quality and relevance of the dataset for the NLP analysis. By addressing issues such as unreadable column names, redundant data, and duplicates, the process ensured that the dataset was optimised for deriving meaningful and accurate insights.

**Before and After ETL Comparison**

An important aspect of data preparation is to visually demonstrate the impact of the ETL process on the dataset. This comparison provides an understanding of how data cleaning and transformation efforts have refined the dataset for analysis.

Refer to Graph 9, as previously discussed in the "Analysis of Author ID" section, originally showed the top 20 active users, including Thameslink-related accounts.



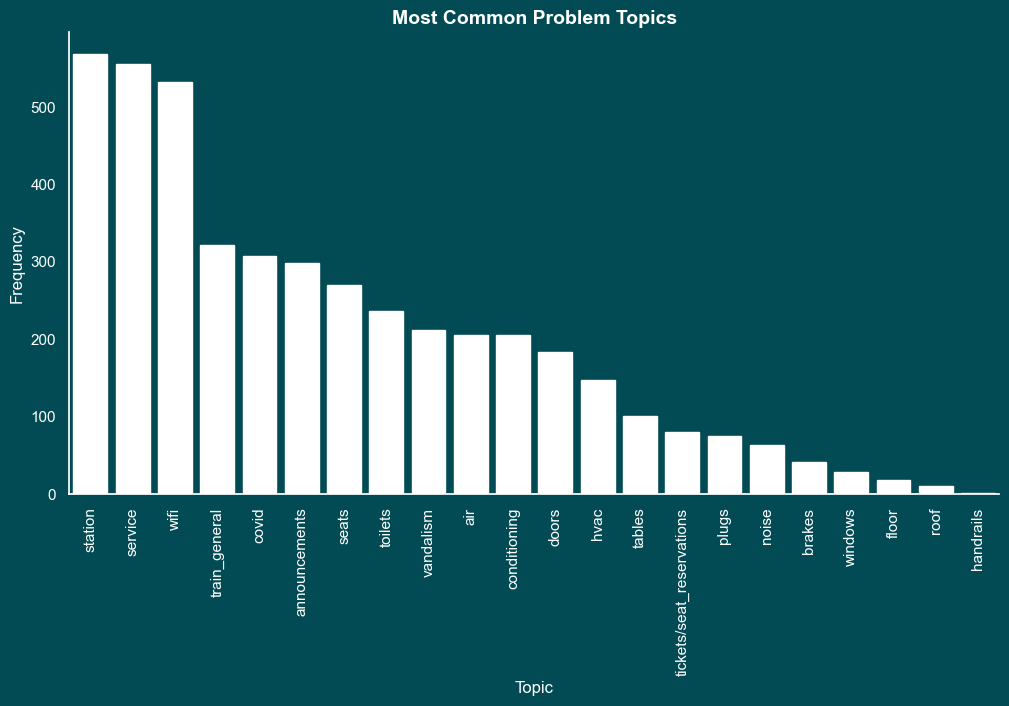
Graph 11: Distribution of Top 20 Active Users (After ETL) – Top 10 Active Users

This Graph illustrates the dataset after the ETL process, specifically showing the impact of removing Thameslink-related accounts. By excluding these official accounts, the focus shifts entirely to customer-generated content, offering a more authentic view of public sentiment and feedback. This Graph demonstrates how the ETL process has refined the dataset, removing potential biases introduced by official communications and ensuring that the analysis is centered on genuine customer experiences and opinions.

**Impact of ETL on Topic Distribution**

The decision to remove topics related to latency in the ETL process significantly changed the topic distribution within the dataset. This change is important to focus the analysis on areas that can provide actionable information for care improvements.

Before the ETL process, the dataset included a significant number of tweets related to delays refer to Graph 7, as previously discussed in the "Topic Classification and Analysis" section. While this is a critical area of concern for passengers, such tweets often do not offer the type of detailed feedback necessary for informing maintenance work.



Graph 12: Most Common Problem Topics (After ETL)

Graph 12 illustrates how the removal of delay-related topics reshapes the focus of the dataset. By excluding these tweets, the dataset becomes more aligned with the project's objective of extracting actionable insights for maintenance improvements. This chart reveals a new distribution of issues, highlighting other areas of concern more directly related to maintenance issues.

Removing delay-related tweets is a strategic decision during the ETL process aimed at improving the suitability of the dataset for the project's goals. Analysis focusing on topics that provide more direct feedback on maintenance-related issues, provides more targeted and actionable information to improve Thameslink services.

**SUMMARY**

The stages of Data Understanding and Data Preparation in the Tweet2GPT project involved a series of methodical and strategic steps to ensure the dataset was primed for effective analysis.

*Data Structure and Annotations*

Initially, we described the data structure in detail, understanding the different types of data present and their organisation. Annotations, such as sentiment labels and topic categories, were carefully checked to ensure they were accurately reflected in the dataset. This step was crucial to set the stage for a more in-depth analysis.

*Data Quantity, Distributions, and Representation*

A thorough evaluation of data quantity, distributions, and representation was undertaken. This process involved assessing the volume of data, understanding how it was distributed across various categories, and ensuring that it was representative of the wider customer sentiment.

*Comprehensive Dataset Usage*

Our approach involved using the entire dataset for analysis without limiting it to specific timeframes or subsets. This decision was made to ensure a holistic view of customer feedback.

*Handling Missing Values and Outliers*

We identified and addressed missing values in the dataset, which is an important step in maintaining the integrity of the analysis. Additionally, outliers were identified and addressed. These actions were important to ensure that the data accurately reflected the reality of customer feedback without being distorted by anomalies.

*Decision Clarification*

Finally, we clarified the logic behind each decision made during the Data Preparation stage. This transparency is key in ensuring that the analysis is replicable and grounded in a solid understanding of the dataset's characteristics and limitations.

**CONCLUSION**

In conclusion, the Data Understanding and Data Preparation stages of the Tweet2GPT project involved a series of careful considerations and actions. From analysing the structure and content of the data to preparing it for analysis by addressing missing values, outliers, and making informed decisions, each step was crucial in ensuring the dataset was optimally prepared for meaningful and accurate analysis. These preparations are the basis on which the suggestions prepared for the maintenance team will be built.