FIT5147 Data Exploration and Visualisation

DEP Part 2: Data Exploration Project

—— How Does MBTI Influence Twitter Engagement and Content?

Student Name: Yinghan Ma

Student ID: 34711783

Major: Master of Data Science

Date: 15 April 2025

Table of Content

1. Introduction

Social media platforms are constantly finding new ways to better understand user behaviour and boost engagement. Personality is one factor that can influence how people interact, express themselves, and connect online.[1] The Myers–Briggs Type Indicator (MBTI) is a popular personality framework that offers insight into these differences.

However, the connection between MBTI traits and online engagement patterns remains underexplored, especially on platforms like Twitter.

In this project, we look at how MBTI personality types relate to engagement patterns on Twitter. Our focus is on three main questions:

1. How do MBTI personality types engage with their social network on Twitter?
2. How does personality influence tweet expression in sentiment and language?
3. How does MBTI distribution on Twitter compare to general demographics?

By exploring these questions, we hope to provide useful insights for social media platforms, marketing teams, and users who want to create more personalised and meaningful online experiences.

2. Data Wrangling and Checking

2.1 Data Sources

This project uses three publicly available or commonly cited datasets:

- **Dataset A: Twitter MBTI Personality Types Dataset**
  Source: Kaggle - Twitter MBTI Dataset[2]
  This dataset includes 8,328 Twitter users who self-reported their MBTI type, along with up to 50 of their most recent tweets. Each record contains a categorical MBTI label and tweet text.
- **Dataset B: MBTI Classification Accuracy Dataset**
  Source: TECLA[3]: A Temperament and Psychological Type Prediction Framework from Twitter Data ([PLOS ONE](https://doi.org/10.1371/journal.pone.0212844.t017))
  This dataset presents the classification performance of various models across the four MBTI dimensions (I/E, N/S, T/F, J/P). It is used to evaluate the model accuracy and understand prediction strengths and limitations across personality traits.
- **Dataset C: MBTI Baseline Distribution Dataset**
  Source: MBTI® Manual, U.S. National Sample (Myers et al., 1998)[4](https://archive.org/details/mbti-manual-a-guide-to-the-development-and-use-of-the-myers-briggs-type-indicator-pdfdrive)
  This dataset provides the baseline personality type distribution based on a national U.S. sample. It is widely cited in psychological and sociological studies and is used in this project as a comparison point to the Twitter population.

2.2 Tools and Environment

All data cleaning, analysis, and visualisation were done using R (version 4.4.3) in RStudio. The following libraries were used:

- dplyr, tidyverse: for data manipulation and summarisation

- ggplot2, wordcloud: for data visualisation
- tidytext: for text tokenisation and stopword filtering
- stringr: for text cleaning and regular expressions
- syuzhet: for sentiment analysis using lexicon-based scoring

2.3 Data Cleaning Procedures

The cleaning process was performed on two datasets: user_info.csv and user_tweets.csv. The key steps included:

1. User Data Cleaning

- MBTI Type Matching: MBTI labels were merged with user info using a shared id column.
- Outlier Removal: Metrics such as followers, friends, tweet count, and favourites were trimmed at the 99.9th percentile to remove extreme values.
- MBTI Dimension Decomposition: Each MBTI type was split into four dimensions (I/E, N/S, T/F, J/P) for dimension-level analysis.

2. Tweet Data Cleaning

- Row Filtering: Only rows with numeric id values were kept to avoid mismatches.
- Reshaping: Tweets were stored across 20 columns and reshaped into long format using pivot_longer().
- Text Normalisation:
  a) URLs, HTML tags, emojis, and symbols were removed using regular expressions.
  b) All text was converted to lowercase for consistency.
  c) Empty or null tweets were filtered out.
- Tweet Length: A tweet_length variable was added to track character counts per tweet.

3. Tokenisation and Word Filtering

- Tweets were broken into word tokens using unnest_tokens().
- Stopwords were removed using the tidytext dictionary.
- Only alphabetic tokens were retained (A–Z) using pattern filtering.
- Word frequencies were calculated separately for extroverts and introverts for use in word clouds.

4. Sentiment Processing

- Each tweet received a sentiment score using the syuzhet method.
- Each score was then aggregated to compute average sentiment scores per MBTI type, which were later visualised to detect personality-emotion correlations.

2.4 Justification and Checking

Several quality checks were performed to ensure data reliability:

- ID and MBTI validation: Only rows with numeric IDs were used. All MBTI types matched the official 16-type set.
- Missing values: Empty tweets and null fields in user metrics were removed.
- Outlier removal: Extreme values were clipped using the 99.9th percentile.
- Manual review: Cleaned tweets were sampled to confirm successful removal of emojis, HTML tags, and non-text elements.
- Token check: Common noise words like "rt" and "amp" were excluded after inspecting token frequencies.
- Sentiment validation: The syuzhet method was applied for short text sentiment, with no missing results after processing.

These checks ensured consistent format and trustworthy inputs for the following analyses.

3. Data Exploration

We structured our data exploration around the three research questions introduced in Section 1. Each set of visualisations and analyses directly addresses one of the core areas of inquiry: engagement behavior (Q1), emotional and linguistic expression (Q2), and population distribution (Q3). Below, we present our findings and interpretation for each research question.

3.1 Engagement Patterns by MBTI Type

3.1.1 Average Engagement Metrics Across MBTI Types

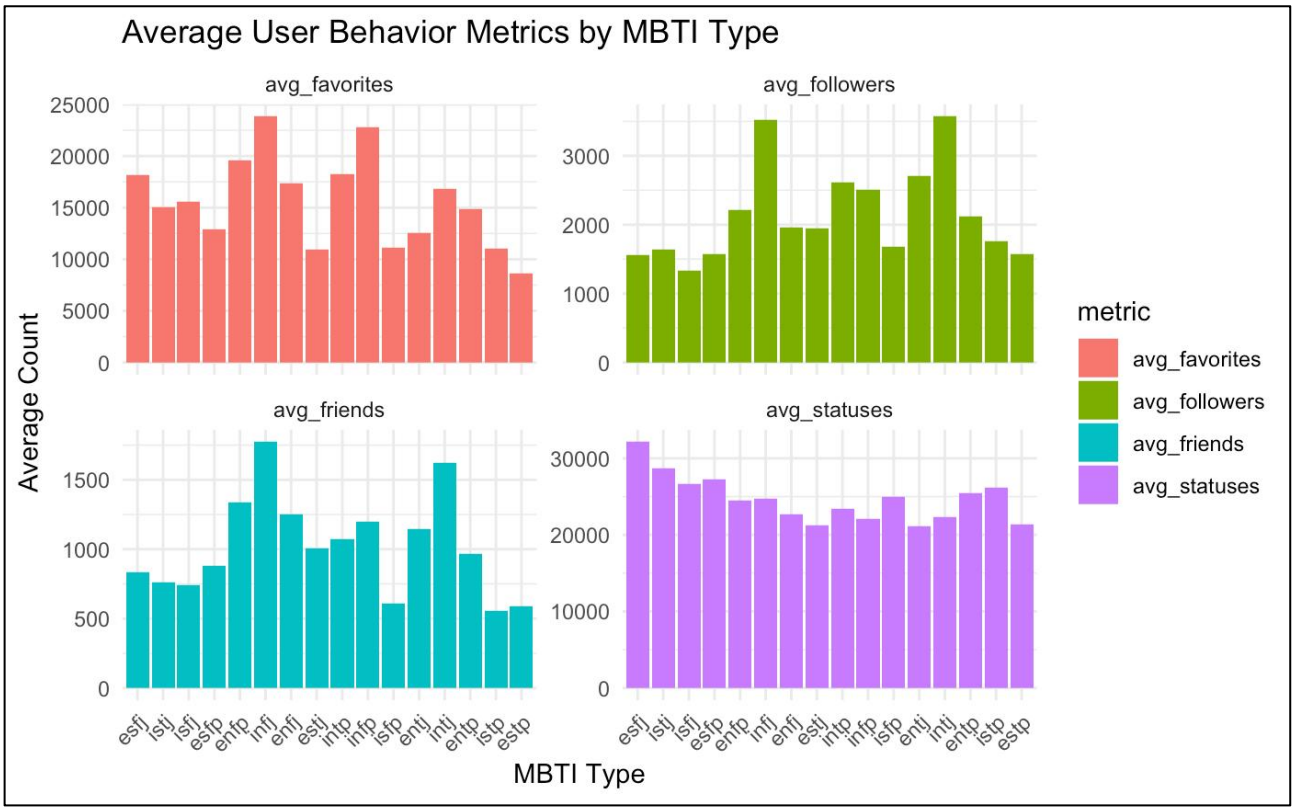

Figure 3.1.1: User Engagement Metrics by MBTI Type

To answer the first research question, we examined how different MBTI personality types engage with their social network on Twitter. Specifically, we calculated the average number of followers, friends (followees), tweets (statuses), and favourites (likes received) for each MBTI category.

The data was summarised using the dplyr package in R, and visualised with ggplot2. We applied group_by() and summarise() functions to compute per-type averages, and presented the results in a multi-panel bar chart (Figure 1) to highlight differences across the four engagement metrics.

Key findings include:

- Favorites: INFJ and INFP users received the highest average number of favorites, and ESTP and ESTJ users received the lowest ones, suggesting that introverted intuitive types tend to generate more engaging or emotionally resonant content.
- Friends (following): INTJ and INFJ also have the highest number of friends, and ISTP, ISFP and ESTP users received the lowest ones, indicating a high level of interest in other users' content despite being introverted.
- Followers: INFJ and INTJ types attracted the most followers on average, and ISFJ and ESFJ users received the lowest ones, highlighting their potential social influence and visibility.
- Statuses (tweets): ESFJ and ISTJ types are the most active in terms of tweet volume, and ESTJ and ENTJ users received the lowest ones, indicating high levels of content output.

These findings highlight that MBTI personality traits are closely associated with distinct social engagement patterns on Twitter. While introverted intuitive types tend to generate more liked and followed content, extraverted sensing types appear more active in terms of posting. This supports our initial hypothesis that personality can shape not only how much users engage, but also how they connect and express themselves online.

3.1.2 Tweet Count Distribution per MBTI Category

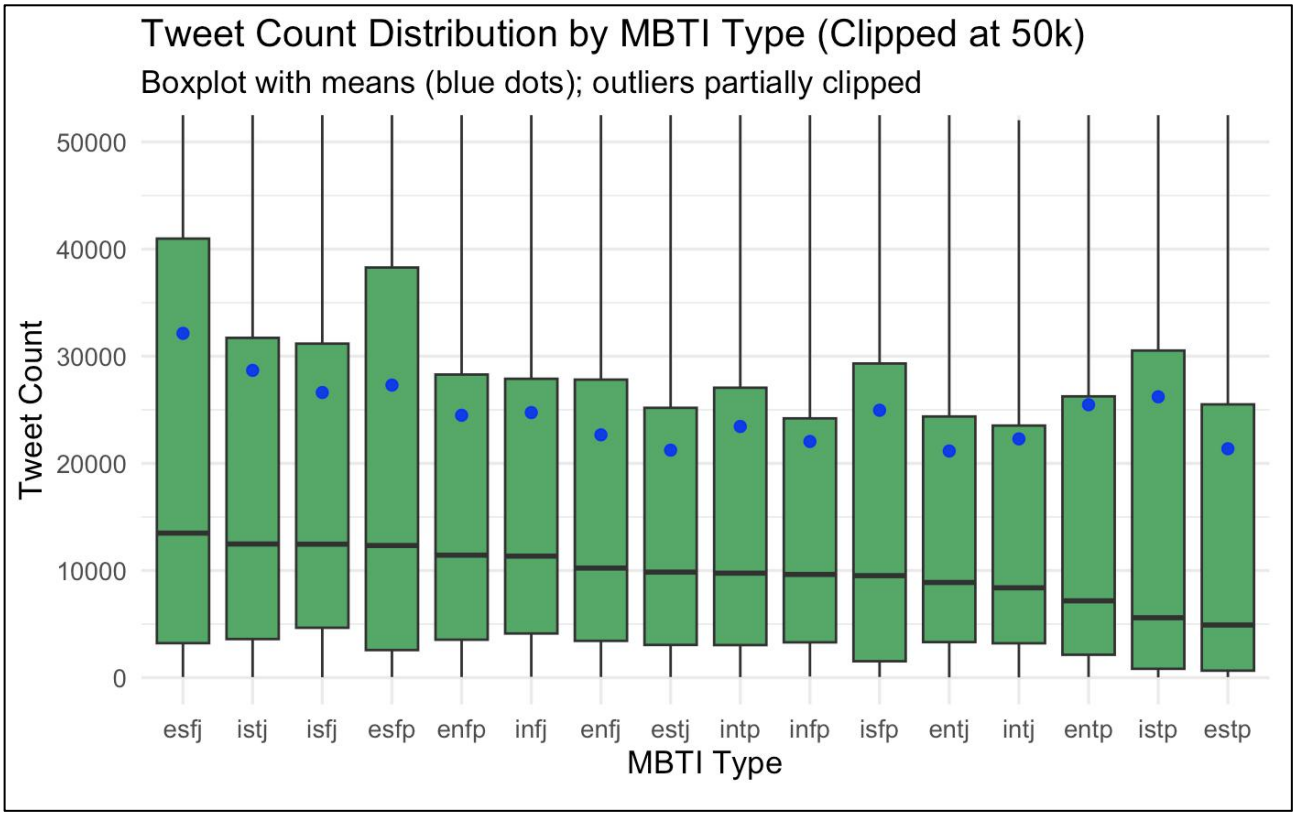

Figure 3.1.2: Tweet Count Distribution Across MBTI Types

To further explore individual tweeting behaviour, we plotted a boxplot showing the distribution of tweet counts across different MBTI personality types. This visualisation reveals how active users are within each personality group, beyond just the averages.

We used the ggplot2 package in R to generate the plot, and dplyr was used to filter out extreme outliers above the 99.9th percentile (clipped at 50,000 tweets) to improve clarity. Each boxplot represents the spread of tweet counts per type, and blue dots indicate the group mean.

Key Findings:

- ESFJ users exhibited the highest overall tweet count, both in terms of average and upper-range values, suggesting consistently high activity.
- ENTJ and INTJ types had lower median tweet counts, but some users within these types still showed high activity, as seen in the upper whiskers.
- ESFP and ENFP types showed wide within-type variation, meaning some users in these groups are highly active while others are not, suggesting heterogeneous engagement patterns.
- Most distributions were right-skewed, with the mean above the median, reflecting a small number of highly active users in each type.

These patterns reinforce earlier observations and suggest that both personality traits and individual differences within types can affect tweet behaviour.

3.1.3 Tweet Frequency and Follower Reach Across MBTI Types

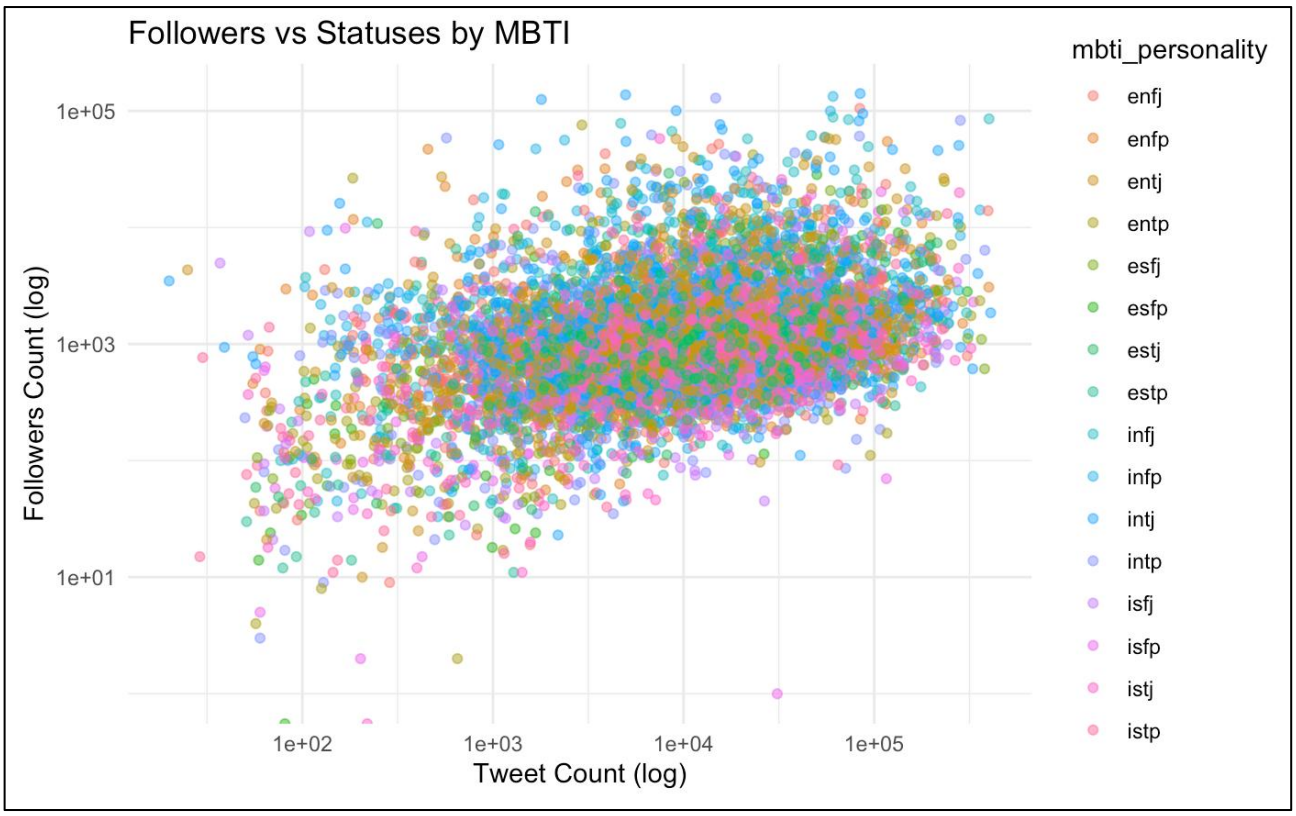

Figure 3.1.3: Tweet Frequency vs Follower Reach

To explore how tweeting activity relates to social visibility across MBTI types, we created a scatterplot comparing tweet count and follower count (both log-transformed). Each point represents a user, coloured by MBTI type.

This plot was generated using ggplot2 in R, with data preprocessed via dplyr. Log scales were applied to handle skew and improve readability.

Key Findings:

- There is a general positive correlation between tweet count and follower count—users who tweet more frequently tend to have more followers.
- However, the relationship appears weak and dispersed, with considerable variability in both dimensions and no clear clustering by MBTI type.
- All personality types show high within-group variation, suggesting that follower count is influenced by additional factors beyond personality traits.
- MBTI types are evenly spread across the plot, indicating that no single type consistently dominates in either tweet frequency or follower reach.

This figure supports our earlier analysis by offering a user-level perspective. While previous charts highlighted group-level tendencies, this scatterplot emphasises that individual variation is substantial, and personality traits alone may not reliably predict social visibility or content activity on Twitter.

3.2 Sentiment and Language Difference

3.2.1 Sentiment Patterns by Personality Type

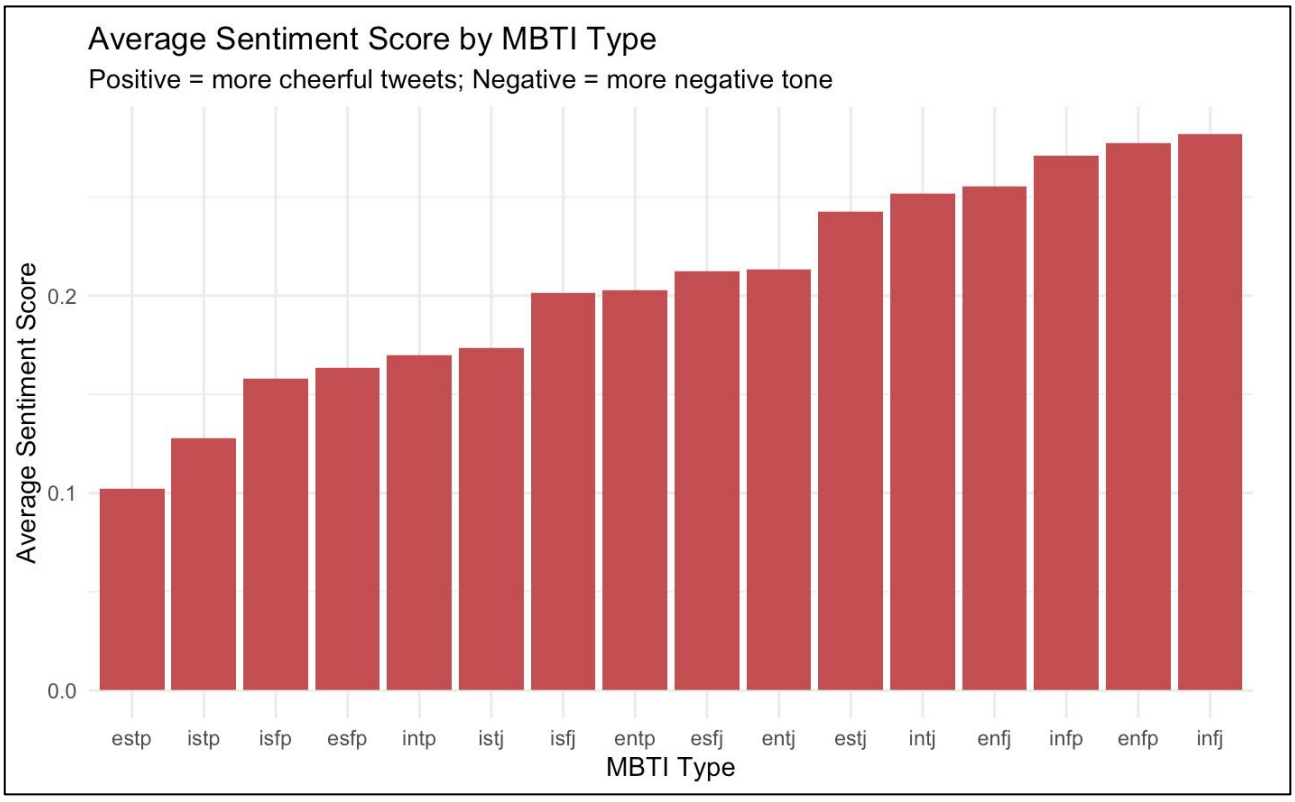

Figure 3.2.1: Sentiment Polarity by MBTI Type

To examine emotional tone across MBTI types, we calculated the average sentiment score of tweets per personality group using the syuzhet package. Higher scores indicate more positive tone; lower scores reflect neutral or negative expression.[3] The plot was created with ggplot2, based on preprocessed English tweets cleaned and summarised using dplyr.

Key Findings:

- INFJ, ENFP, and INFP types showed the highest sentiment scores, suggesting a more cheerful and emotionally expressive tone.
- ESTP and ISTP types ranked lowest, indicating a tendency toward neutral or less emotionally charged language.
- A clear trend emerged: intuitive–feeling types generally express more positive emotions, while sensing–thinking types lean toward a more reserved or factual style.

These results support the idea that MBTI traits influence emotional expression, with feeling-dominant users more likely to use affective language on social media.

3.2.2 Lexical Patterns: Extroverts vs Introverts

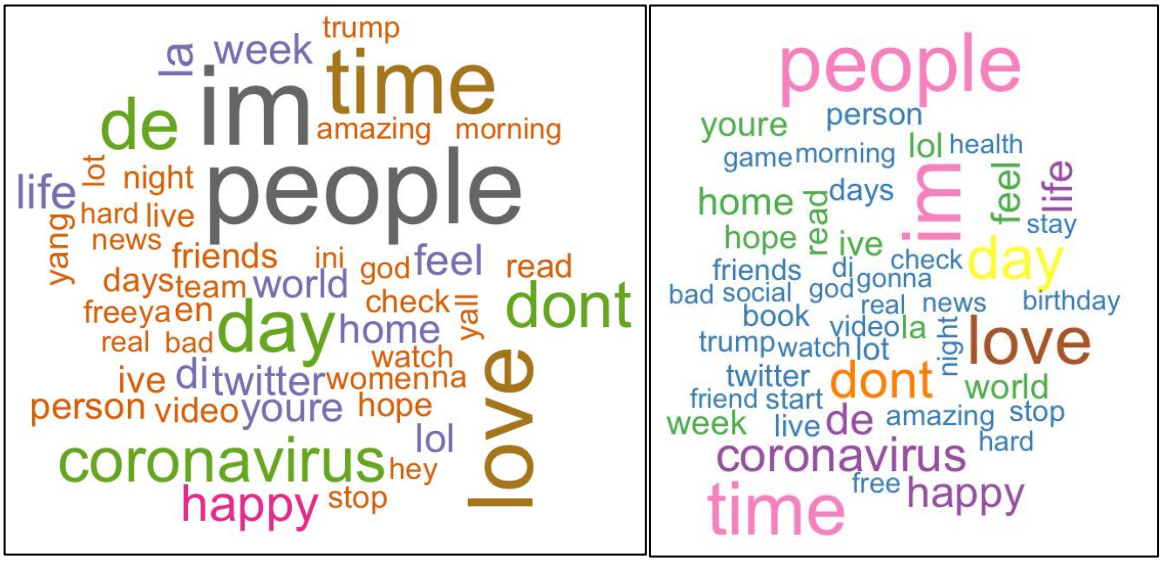

Figure 3.2.2: Top Words Used by Extroverts and Introverts

To examine how language use differs between extroverted and introverted personality types, we generated separate word clouds based on the top 200 most frequent words used by each group. Tweets were preprocessed using dplyr and tidytext, with stop words removed and only alphabetic tokens retained. The visualisation was created using the wordcloud package in R.

Key Findings:

- Shared themes: Words like people, time, love, and coronavirus appear in both groups, reflecting common social and global topics during the data collection period.
- Distinct focus:
  - Extroverts use more socially oriented and event-related terms such as video, birthday, social, and game.
  - Introverts more often mention introspective words like feel, hope, and life, suggesting a more reflective tone.
- Emotional tone: Words with positive valence such as happy and love are slightly more prominent in the introvert cloud, which aligns with their higher average sentiment scores seen in Section 3.2.1.

These lexical patterns offer further support for personality-linked differences in emotional expression and communication style.

3.3 Personality Type Distribution Comparison

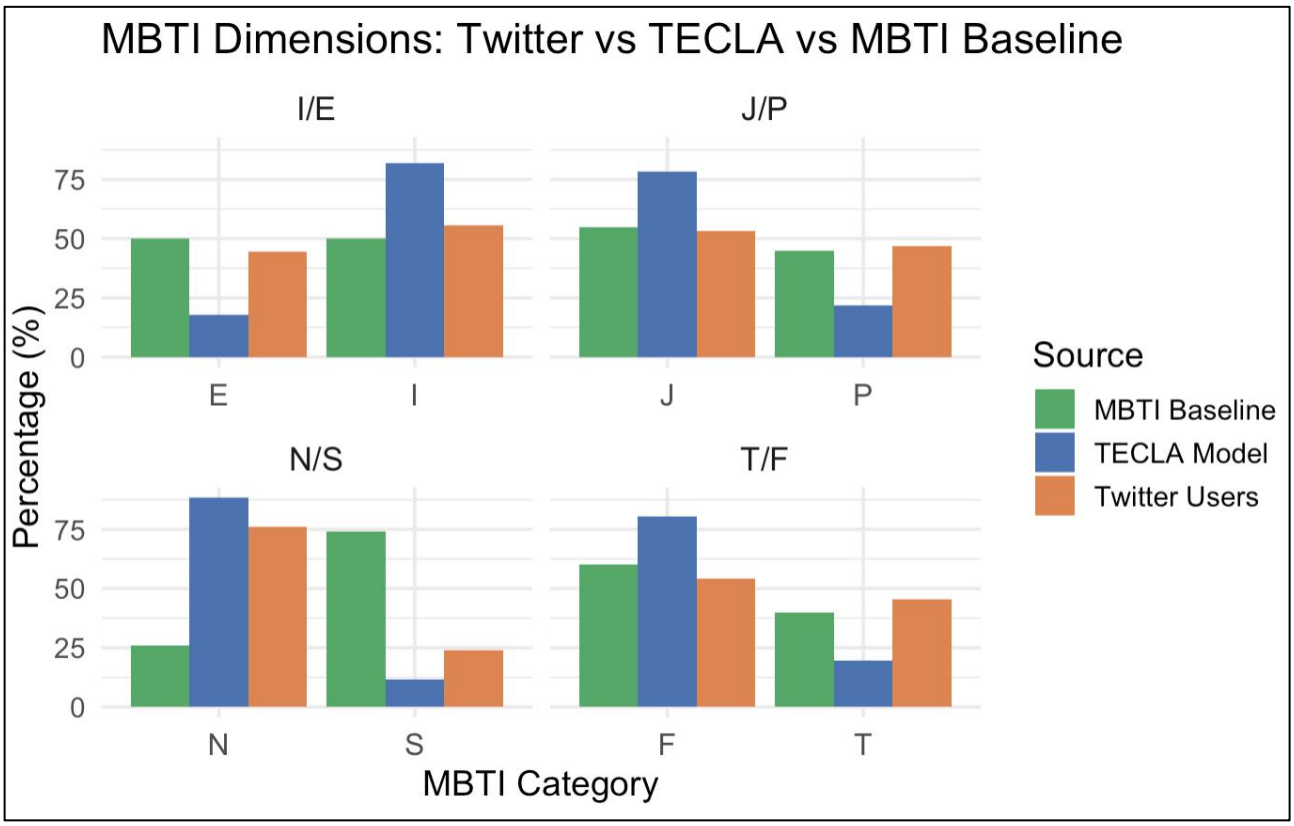

Figure 3.3: Comparison of MBTI Trait Distribution

To explore how Twitter users differ from broader reference populations in terms of MBTI traits, we created a faceted bar chart comparing the proportions of four MBTI dimensions (I/E, N/S, T/F, J/P) across three sources: Twitter users, the TECLA model (from prior research)[3], and standard MBTI baseline statistics[4]. The chart was generated using ggplot2, with proportions calculated using dplyr and manually referenced from the TECLA study and MBTI® Manual.

Key Findings:

- Introversion bias: Over 55% of Twitter users in our dataset are classified as Introverts (I), aligning with the TECLA model but differing from standard MBTI distributions, where the split between I and E is more balanced.
- N/S divergence: Approximately 75% of Twitter users are classified as iNtuitive (N), compared to a majority of Sensing (S) types in traditional MBTI datasets. This suggests that social platforms like Twitter may appeal more to abstract thinkers and concept-driven communicators.
- F and J lean: Our Twitter dataset shows a higher proportion of Feeling (F) and Judging (J) types, which is consistent with TECLA's model predictions and slightly elevated compared to MBTI population benchmarks.
- TECLA–Twitter alignment: The TECLA model closely mirrors our dataset, implying that it effectively captures traits typical of Twitter users rather than the broader population.

Rather than reflecting population-wide personality norms, the MBTI trait distribution on Twitter appears to reflect platform-specific tendencies. Users may be more likely to possess introspective, intuitive, and emotionally expressive traits -- characteristics that support reflective sharing and opinion-driven interaction online.

4. Conclusion

This project examined how MBTI personality traits relate to social media behavior using a Twitter-based dataset. Through a combination of behavioral metrics, sentiment scores, and lexical analysis, we addressed the three guiding research questions:

4. Engagement Patterns: We observed clear behavioral differences across MBTI types. ESFJ and ISTJ users were the most active in terms of tweet volume, while INFJ and INFP users received the highest number of favorites. This suggests that cognitive style may influence both content creation and audience response.
5. Emotional Expression: Sentiment analysis revealed that Feeling–Intuitive types, suah as INFJ and ENFP, used more emotionally expressive and positive language, whereas Thinking–Sensing types, such as ESTP and ISTP, tended toward a more neutral or restrained tone.
6. Personality Distribution: Compared to MBTI baseline statistics, Twitter users in this dataset skewed more toward Introversion and Intuition. This divergence implies that certain personality types may be more drawn to social media platforms like Twitter, possibly due to its format and culture.
7. User Profile Insight: The MBTI distribution on Twitter, when contrasted with survey-based benchmarks, suggests the platform may attract users with different cognitive styles—particularly those who are introspective and abstract in their thinking.

Overall, our findings support the idea that personality traits can influence not only what people share online, but how they interact and engage with others. These insights may inform future research in user profiling, content personalization, and platform design.

5. Reflection

This project taught me how to combine data cleaning, text processing, and visualisation to extract meaningful insights from social media data. I learned how to use tools like dplyr, tidytext, and ggplot2 more effectively, especially for handling messy text data.

If I were to redo this project, I would spend more time refining the token filtering and sentiment scoring steps, since some noise words still appeared in the word clouds. I would also consider adding statistical tests to support the visual patterns observed. Despite these challenges, the process helped me better understand the real-world data exploration.

6. Bibliography

[1] Sajjad, M., & Zaman, U. (2020). Innovative perspective of marketing engagement: Enhancing users' loyalty in social media through blogging. Journal of Open Innovation: Technology, Market, and Complexity, 6(3), 93

[2] Rai, S. (2018). Twitter MBTI Dataset. Kaggle. https://www.kaggle.com/datasets/sanketrai/twitter-mbti-dataset

[3] Plank, B., & Hovy, D. (2019). Personality Traits on Twitter-or-How to Get 1,500 Personality Tests in a Week. PLOS ONE, 14(3), e0212844. https://doi.org/10.1371/journal.pone.0212844

[4] Myers, I. B., McCaulley, M. H., Quenk, N. L., & Hammer, A. L. (1998). MBTI® manual: A guide to the development and use of the Myers-Briggs Type Indicator®. Consulting Psychologists Press.

[5] Rinker, T. W. (2017). syuzhet: Extract Sentiment and Plot Arcs from Text. R package version 1.0.6. https://cran.r-project.org/web/packages/syuzhet/

[6] I acknowledge the use of ChatGPT (https://chat.openai.com/) to refine the academic language and accuracy of my own work.