Information visualization Assignement 1

NUS
National University
of Singapore

Author
Jean Edouard Acker

Assignement 2

Course
CS5346

Professor
Bimlesh Wadhwa

Contents

if you can't launch the react application (available here), you can send me an e-mail: e1374370@u.nus.edu. I'll try to make it available online, a demo video is also provided with this report.

1 Data Visualization with Tableau

2 Introduction

Social media platforms are becoming increasingly influential in our daily lives, with the most popular ones boasting over two billion active users. Each platform is characterized by unique user interaction mechanisms that shape the way content is shared and perceived:

- Twitter is often centered around debate and discussions.
- Instagram is primarily used for showcasing lifestyles, vacations, and personal moments.
- Snapchat is highly popular among younger users as a means of instant communication.

These distinctive mechanics influence the type of content that is promoted and the emotions conveyed on each platform. Suggestion algorithms tend to reinforce specific sentiments, often amplifying the most engaging or polarizing content.

This dashboard leverages sentiment analysis techniques to examine the emotions embedded in social media posts, providing insights into the strengths and weaknesses of each platform. The main objectives of this visualization project are:

- To highlight how some social media networks foster negativity more than others.
- To encourage users to reflect on the influence of these platforms on their thought processes, acknowledging that they do not accurately represent the real world. Social media often amplifies divisive content, where only the most extreme viewpoints gain visibility.
- To assist developers in adjusting their platform policies or identifying opportunities to position themselves in alternative market segments within the vast landscape of social networks.

This project was developed using D3.js to create interactive visualizations that provide a comprehensive overview of sentiment distribution across different social media networks.

3 Datasets and Pre-processing

This section details the datasets used in the project and describes the pre-processing steps applied to them.

3.1 Bar Race Dataset

The dataset for the first visualization, the Bar Race, was manually compiled from multiple sources into a single JSON file:

- Facebook (2021). Monthly Active Users of Facebook.

- https://backlinko.com/twitter-users.
- https://www.demandsage.com/instagram-statistics/.
- https://www.statista.com/statistics/545967/snapchat-app-dau/.
- https://www.statista.com/statistics/234038/telegram-messenger-mau-users/.

The data is generally averaged per quarter, though for some platforms (notably Telegram), the resolution might be lower. Additionally, recent data points are not always tabulated, which explains why Facebook's user count does not change after 2022.

3.2 Treemap Dataset

The second visualization, the Treemap, was made possible by the following dataset:

> https://www.kaggle.com/datasets/emirhanai/social-media-usage-and-emotional-well-being.

This dataset contained 1000 entries with the following features:

- `User_ID`
- `Age`
- `Gender`
- `Platform`
- `Daily_Usage_Time (minutes)`
- `Posts_Per_Day`
- `Likes_Received_Per_Day`
- `Comments_Received_Per_Day`
- `Messages_Sent_Per_Day`
- `Dominant_Emotion`

Only the `Platform` and `Dominant_Emotion` features were retained. The data was then parsed into the JSON format expected by D3, with the `value` attribute counting the number of posts associated with each emotion.

3.3 Bar Chart Dataset

The third visualization is a bar chart showing the proportion of positive, negative, and neutral sentiment on Instagram, Twitter, and Facebook. The dataset used is:

- Social Media Sentiments Analysis Dataset

It contains 500 entries with the following features:

- Year

- Month
- Day
- Time of Tweet
- text
- sentiment (positive, neutral, negative)
- Platform (Instagram, Facebook, Twitter)

Only `Platform` and `sentiment` were retained. The data was parsed into JSON format:

```
[{"socialmedia": <social_media>, "positive": [<proportion of positive per year>],
"negative": [<proportion of negative per year>],
"neutral": [<proportion of neutral per year>]}, {...}]
```

3.4 Histogram Dataset

The fourth visualization is a histogram showing the number of tweets from Russian troll factories per week. The dataset comes from:

- fivethirtyeight/russian-troll-tweets

Originally containing 3,000,000 tweets with 21 features, half of the dataset was retained and then filtered by `account_category` to keep tweets classified as *RightTroll*, *LeftTroll*, *Fearmonger*, *NewsFeed*, *Commercial*, and *HashtagGamer*. The timestamps were formatted, and data was grouped by week to compute the number of trolls per week, which was then exported in CSV format.

3.5 Parallel Plot Dataset

The final visualization is a parallel plot with five axes, made possible by:

- Jon Bruner's Twitter User Dataset

From 600,000 entries corresponding to unique users, 30,000 were randomly sampled. Out of the original 24 features, six were retained:

- `name` – Used to identify each line by the account name.
- `account_created_at` – Date of account creation.
- `followers_count` – Number of Twitter accounts following this account.
- `following_count` – Number of Twitter accounts this account follows.
- `statuses_count` – Number of public posts created by this account.
- `listed_count` – Number of lists on which this account appears.

The timestamp column was formatted into a consistent format, and the data was exported as CSV.

4 Visualizations

4.1 Bar Race: Monthly Active Users

The first visualization is a **bar race** representing the number of monthly active users (in millions) for each social media platform from January 2017 to October 2024. It highlights the rapid growth in user adoption, emphasizing the universal expansion of social media across different platforms. The data comes from multiple sources and has been manually aggregated into a JSON file. In most cases, user counts are averaged per quarter, but some platforms—such as Telegram—have data with a lower resolution. For recent dates, some figures are not consistently reported, which explains why Facebook's user count remains unchanged after 2022.

4.2 Sentiment Analysis: Treemap and Bar Chart

The next two visualizations are:

- A **treemap** showing the proportion of sentiments for each social media platform.
- A **bar chart** illustrating the evolution of sentiments over time across different platforms.

These two visualizations are **complementary**: one provides an overview of the **dominant sentiments** on each platform, while the other tracks the **evolution of these emotions over time**.

The main limitation of these plots lies in the **small dataset size**: the **treemap** is based on **1,000 messages**, while the **time series** uses **500 messages from three social networks** spanning **14 years**.

Finding a dataset that includes **both labeled sentiments and message timestamps** proved challenging. For more meaningful insights, **training a model** to label a **larger dataset** of posts from Instagram, Twitter, and Facebook would likely be necessary.

Despite this limitation, the treemap remains insightful, as it reveals a **dominance of negativity on Twitter**, in contrast to a **prevalence of happiness on Instagram**.

4.3 Russian Troll Tweets: Histogram

The fourth visualization is a **histogram** displaying the number of tweets sent by Russian troll accounts per week.

This histogram represents a sample of tweets identified as originating from Russian troll bots. The bars are aggregated on a weekly basis, with their height reflecting the number of tweets posted each week.

The goal is to highlight the impact of troll factories and their strategic targeting of critical periods, such as U.S. elections, to influence public opinion. This histogram also demonstrates the reality of the influence games played on social networks, showing that a significant portion of social media content is designed to sow discord and trap users in recommendation loops of divisive content and fake news.

Significant spikes in activity can be observed on key dates:

- **October 2016**: Likely related to the October 7, 2016 WikiLeaks release of emails from the Clinton campaign.
- **November 2016**: Corresponding to the U.S. presidential election.
- **Summer 2017**: A shift in focus to a specific type of troll known as the *Right Troll*.

4.4 Twitter Users: Parallel Plot

The final visualization is a **parallel plot** that analyzes Twitter users, highlighting that the vast majority of users remain largely passive on the platform. As a result, most of the content comes from a small number of highly active users.

Each line in the parallel plot represents a single user, with a total of 30,000 users displayed. The plot consists of five axes:

- **Account creation date**: The year the Twitter account was created.
- **Followers count**: Number of Twitter accounts following this account.
- **Following count**: Number of Twitter accounts that this account follows.
- **Statuses count**: Number of public posts created by this account.
- **Listed count**: Number of lists on which this account appears (Twitter lists are curated collections of accounts grouped by users for easier access to specific content streams).

The parallel plot reveals that most lines are concentrated towards the right side of the graph, indicating that the majority of accounts have low activity and a small follower base. In other words, **a small number of highly active users produce most of the content**, which contributes to the polarization of social media discussions.

5 Visualization Photos

To better illustrate the visualizations, the following images are included, a **demo video is provided with this report**:

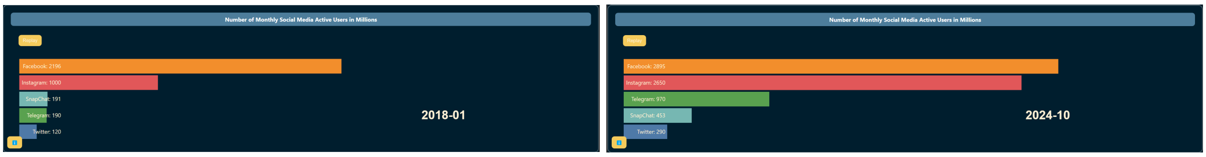

Figure 1: Number of Monthly Social Media Active Users in Millions

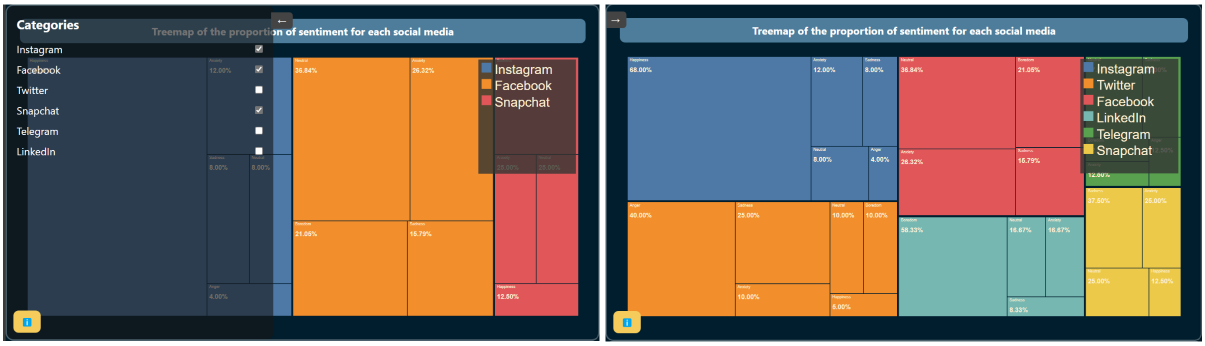

Figure 2: Treemap of the proportion of sentiment for each social media

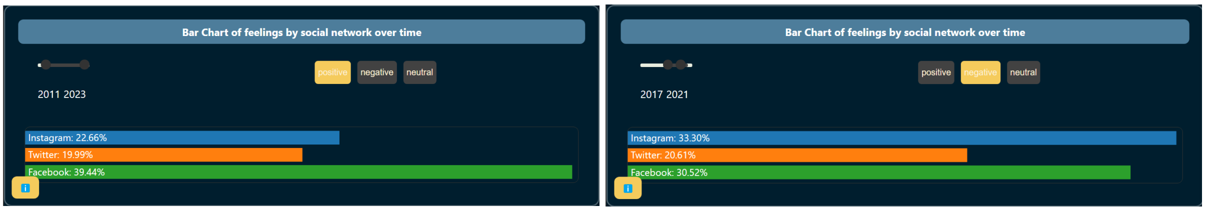

Figure 3: Bar Chart of feelings by social network over time

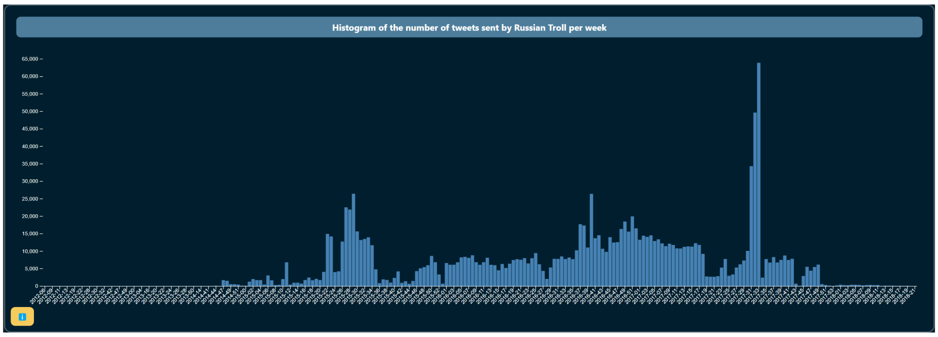

Figure 4: Histogram of the number of tweets sent by Russian Troll per week

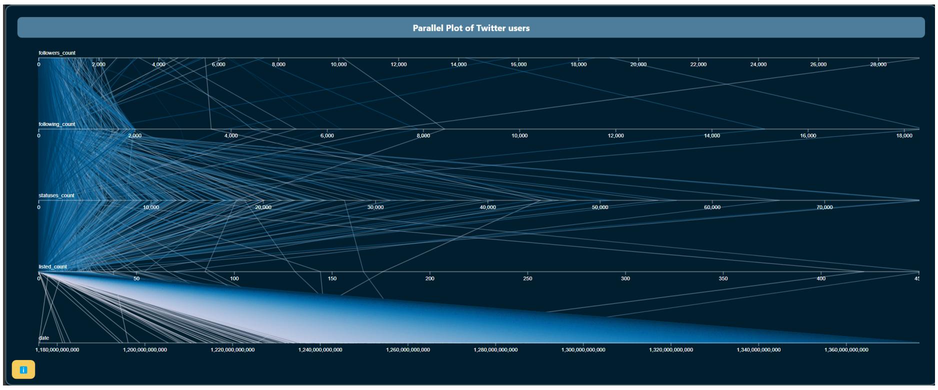

Figure 5: Parallel Plot of Twitter users

6 Conclusion

This project allowed me to deepen my knowledge of D3 and challenge myself with technical visualizations. I found the exercise of creating interconnected visualizations particularly interesting. Additionally, identifying appropriate and meaningful visualizations was not an easy task, especially considering the difficulty of finding datasets suited to the intended representations.

References

[1] Facebook. (2021). Monthly Active Users of Facebook [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/2565605

[2] X (Twitter) Statistics: How Many People Use X? (2025). https://backlinko.com/twitter-users

[3] How Many People Use Instagram 2025 (New Statistics). https://www.demandsage.com/instagram-statistics/

[4] Snapchat Daily Active Users 2024. https://www.statista.com/statistics/545967/snapchat-app-dau/

[5] Telegram Global MAU 2024. https://www.statista.com/statistics/234038/telegram-messenger-mau-users/

[6] Emirhan BULUT. (2024). Social Media Usage and Emotional Well-Being [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/8460631