# In-class Lab I

AYHAN SULEYMANZADE, MAHSA AGHAZADEH, 이준모, 김지인

In this Lab session, Trump's tweets were tested to investigate the following hypotheses:

**Hypothesis I:** Tweets that address controversial issues, significant policy changes, or international conflicts receive higher engagement due to their emotional impact and direct relevance to users. However, tweets about ceremonial topics get less interaction, likely due to their predictable nature and limited scope for discussion.

**Hypothesis II:** Tweets with negative sentiments gain more engagement, because of the 'negativity bias' which asserts that negative content has a stronger psychological impact, often leading to higher interaction on social media due to users' reactive behavior toward the content they find concerning.

**Data-Preprocessing:**

In text preprocessing, several key steps were implemented to enhance the quality and consistency of the data, vital for effective text analysis. The text is first converted to lowercase, ensuring uniformity and reducing complexity. Next, the non-contributory elements like URLs, social media handles, and hashtags are removed to eliminate noise and focus on meaningful content. The process also involves filtering out non-alphabetic characters and short words (less than three characters), as these often do not add significant meaning and can be considered noise. Next, extra spaces are removed to maintain textual consistency. Finally, tokenization and stop-word removal are crucial. Tokenization breaks the text into individual words, enabling the analysis of each word's role and significance. Subsequent removal of stop words – common words like "the", "is", and "at" – is essential because these words, while frequent, often add little meaningful information to the text. Each of these steps is essential for refining the text, removing irrelevant components, and standardizing the data.

**Hypothesis I Analysis:**

Our methodology employs Latent Dirichlet Allocation (LDA), a sophisticated topic modeling technique, to analyze a dataset of President Trump’s tweets. LDS assists in discovering abstract “topics” within a large volume of text, based on patterns of word occurrence. Utilizing the LDA on preprocessed text, we obtained 15 outputs with a specific formula in the form of $T = \sum_i I_i \times W_i$, where $T$ denotes the topic $I_i$ and $W_i$ the words and its specific importance in a defined topic. For example, the most common topic is produced with the following output: $$ Topic 1 = 0.030*"states" + 0.029*"united" + 0.027*"trade" + 0.018*"great" + 0.016*"look" + 0.015*"forward" + 0.014*"china" + 0.010*"president" + 0.010*"world" + 0.008*"leaders" $$

Based on this formula, since it is hard for the human eye to detect the topics, we have utilized the LLM (OpenAI GPT4) to give us the topic names as in usual format. We then grouped the tweets based on the topic and averaged the number of likes and the retweets.

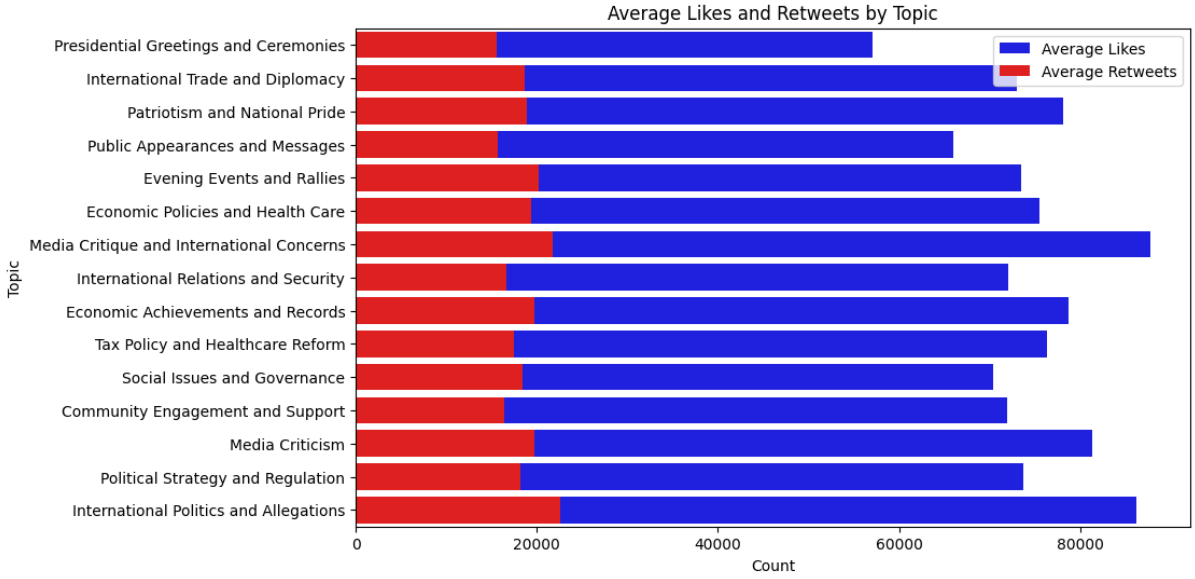


Figure 1. Bar chart. Average Likes and Retweets by most common 15 topics.

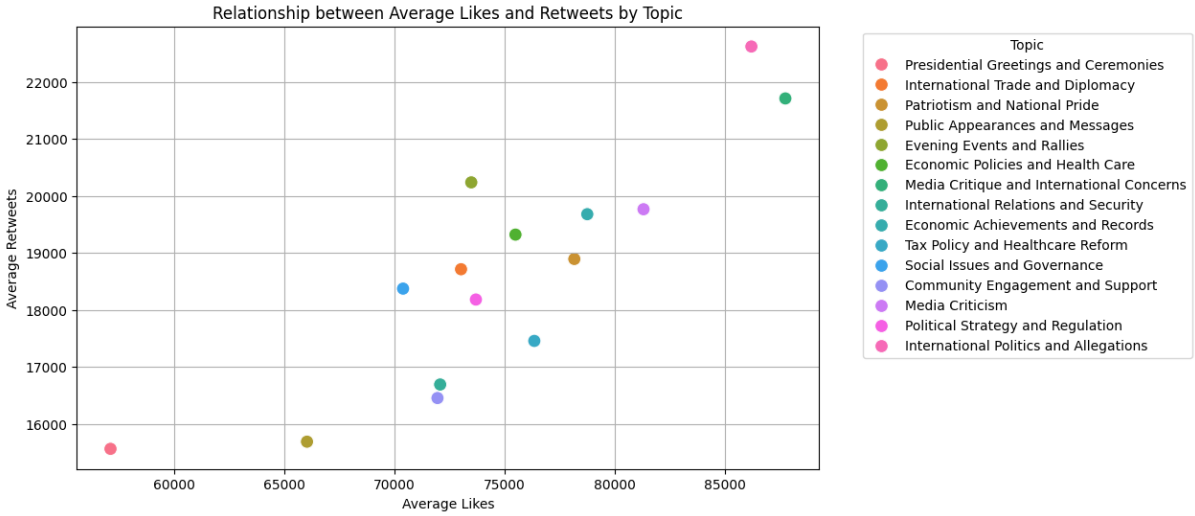


Figure 2. Scatter plot. Relationship between Average Likes and Retweets by Topic.

According to the data visualization in Figure 1, it appears that tweets categorized under 'International Politics and Allegations,' 'Political Strategy and Regulation,' and 'Media Criticism' receive the highest average likes and retweets, indicative of more significant user engagement. In contrast, 'Presidential Greetings and Ceremonies' and 'International Trade and Diplomacy' show notably lower engagement.

The higher engagement in the first three categories may be attributed to several factors:

1. **Provocative Nature of Content**
   Tweets about 'International Politics and Allegations' and 'Media Criticism' often contain provocative content, potentially involving accusations or contentious commentary. These tweets may attract attention because they offer insights into conflicts or challenges to established narratives, sparking discussions and polarized reactions.
2. **Impact on Audience**
   Tweets related to 'Political Strategy and Regulation' likely detail policy decisions or political maneuvers that can have direct consequences for the public. These tweets might galvanize support or opposition, prompting users to express their stance through likes and retweets.

On the other end, possible reasons for less engagement are:

1. **Ceremonial Nature**
   Tweets about 'Presidential Greetings and Ceremonies' are typically more informative and less controversial, resulting in less engagement as they may lack the emotional or confrontational elements that drive user interaction.
2. **Specificity and Complexity**
   'International Trade and Diplomacy' subjects often involve complex negotiations and specialized knowledge, which may resonate less widely with the general Twitter audience, leading to lower engagement.

In addition, the scatter plot in Figure 2 leads us to have the following reasoning:

1. **Correlation Between Likes and Retweets:** Generally, a topic with a high average of likes also tends to have a high average of retweets, indicating a positive correlation between these two forms of engagement. This pattern is expected and suggests that tweets that resonate with the audience are likely to be both liked and shared.

In summary, the analysis corroborates Hypothesis 1, revealing that tweets on Provocative and relevant topics like 'International Politics and Allegations' engage users more than neutral content such as 'Presidential Greetings.' This highlights the role of content's emotional charge and relevance in driving Twitter engagement.

**Hypothesis II Analysis:**

Our sentiment analysis method, utilizing the TextBlob library, quantifies the emotional tone of tweets to assess their impact on user interactions. We assigned a polarity score to each tweet, ranging from -1.0 for negative to +1.0 for positive sentiments, with scores around zero indicating neutrality.

We first generated a scatter plot to visualize the relationship between the sentiment score of tweets and their engagement metrics as follows:

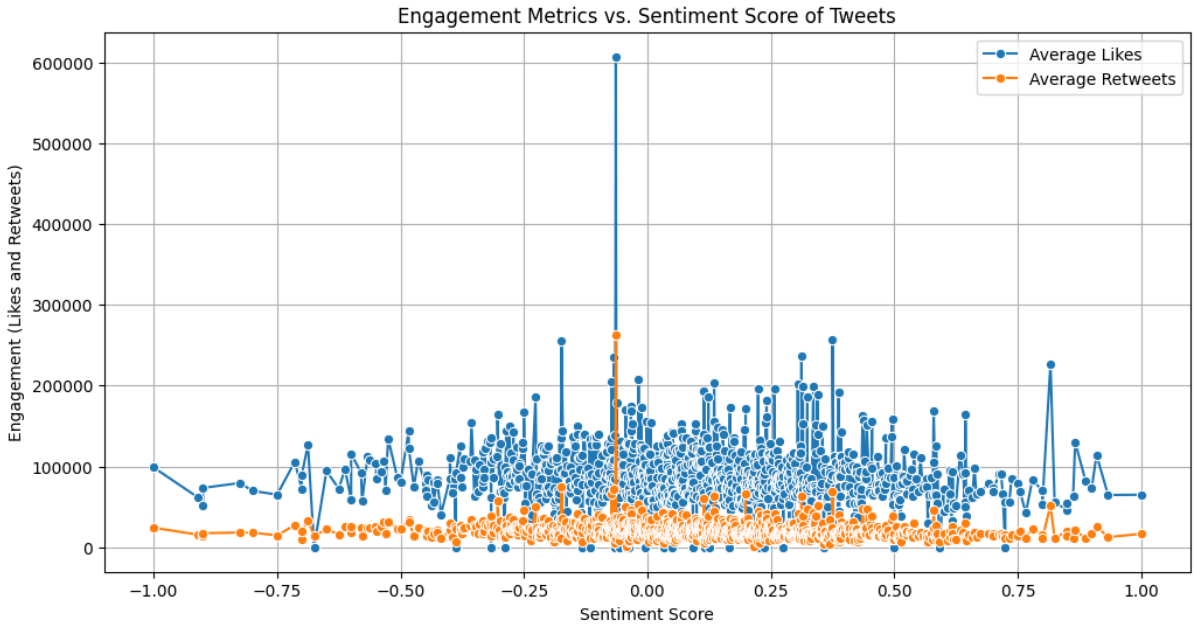


Figure 3. Engagement Metrics (likes and retweets) vs. Sentiment Score of Tweets

The scatter plot correlates tweet sentiment scores with engagement, showing numerous neutral-sentiment tweets. Negative tweets generally receive more likes and retweets, the latter being less frequent than likes, implying users prefer to 'like' rather than share content. Positive tweets see less engagement, and spikes may indicate viral tweets or significant events.

However, due to the presence of outliers with high engagement, a logarithmic scale has been used to normalize the data visualization to help in better understanding the overall trends by reducing the impact of outliers.

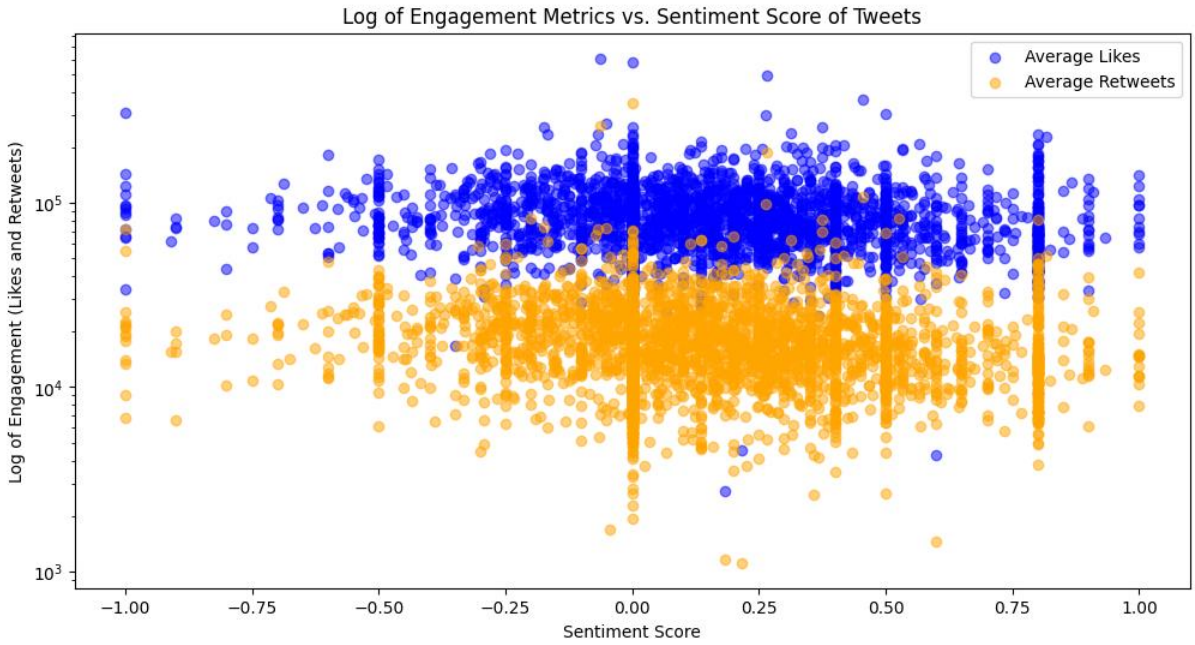


Figure 4. Logarithmic scale on Engagement Metrics vs. Sentiment Score of Tweets

Since using this figure we cannot infer meaningful results, we classified them into five categories: ‘very positive’ for scores above 0.6, ‘positive’ for those above 0.2, ‘neutral’ for scores between -0.2 and 0.2, ‘negative’ for scores below -0.2, and ‘very negative’ for scores below -0.6.

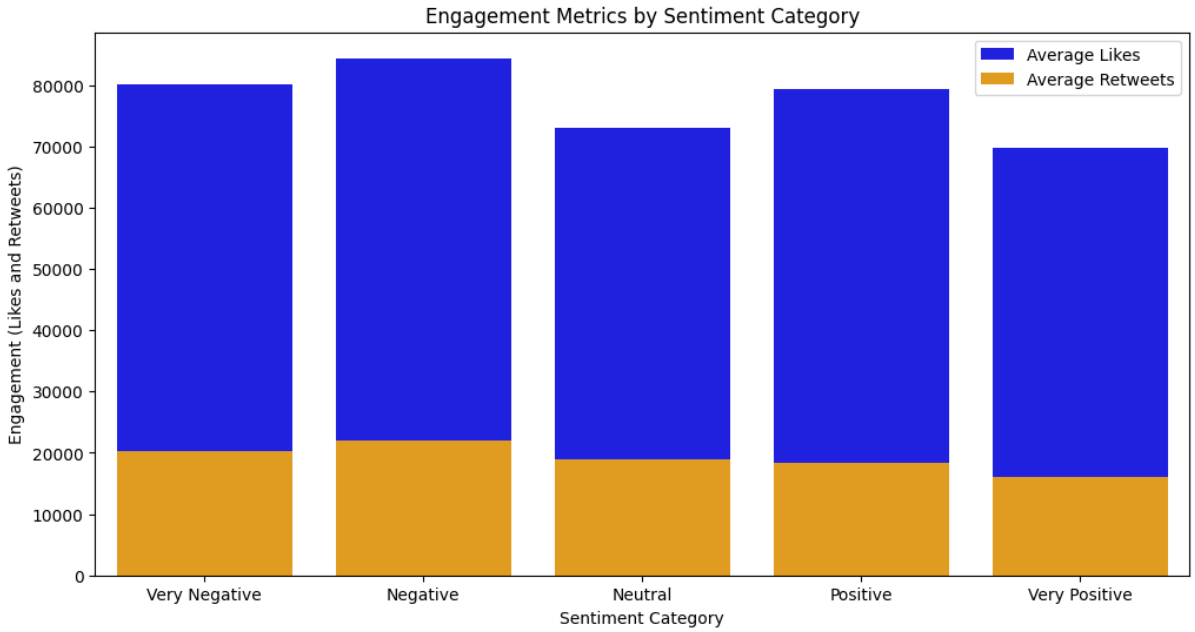


Figure 5. Bar chart. Engagement Metrics (likes and retweets) by Sentiment Category

The bar chart analysis reveals that negative tweets receive the highest engagement, positive tweets better than neutral ones but still fall short of the engagement seen with negative content.

Possible reasons for such results are:

1. **Negativity Bias**
   People have a psychological propensity to pay more attention to negative information. Negative tweets might spark a sense of urgency or concern, prompting users to respond more readily through likes and retweets as a form of acknowledgment or amplification.
2. **Emotional Activation**
   Negative tweets often trigger stronger emotions, which can lead to higher engagement as users are more compelled to react to content that affects them emotionally, whether through support, opposition, or sharing to inform others.

In conclusion, the sentiment analysis suggests that negative tweets tend to engage users more than positive or neutral ones, likely due to a negativity bias and emotional activation, where negative content prompts a stronger psychological response and greater user interaction.

**Discussion: Based on the property of Big Data**

Twitter is known as one of the most algorithmically confounded social media. Their recommendation algorithms do not recommend political tweets to people who have not shown interest in these types of tweets which means this data does not represent the whole society. Since Trump is a politician, we can assume that people that follow Trump are into politics, particularly, American politics. People that do not live in the USA and people that are not into politics would have different properties from the results of Trump's tweets. So, the results of this paper cannot be applied to other areas of twitter.

Moreover, as it was explained above this dataset was dirty and we had to preprocess the data to clean it. We needed to remove stop words , URLs, social media handles etc to get the meaningful content. Also, there could be possible bots or fake accounts that liked or retweeted which we couldn't filter out because the data is already prepared, and this dirtiness in the data may have affected the result.

On the other hand, we analyzed 3,196 updated tweets from Trump, which is online and big enough to reduce noises and generalize the results. Also, the non-reactiveness of the data further enhances the objectivity of the conclusion. Therefore, we can say that the analysis successfully captured the behavior of the twitter users who are political enthusiasts, according to the content of Trump's tweet.