SI 618: Data Manipulation and Analysis

Final Project Report
on
Tweeting the Fast Lane: Sentiment Analysis in the World of F1 Racing

Contributors:
Sudhanshu Agarwal (sudhagar)
Shitanshu Bhushan (sbhushan)
Divyam Sharma (divyams)

Contents:

1. Motivation

We are F1 enthusiasts who follow F1 from the *lights out to the checkered flag*. Being students of SI 618 gave us a good opportunity to merge our passionate interest in F1 with the final project work granting us a coherent alignment of applying the course material with the practical application of data analysis for uncovering unique relationships between the two datasets. The project involves the analysis of two publicly available datasets related to F1 Racing, encompassing Twitter data and comprehensive F1 race details. The project aims to perform EDA, create custom queries for establishing correlations between the datasets, and use NLP to delve into emotion analysis using the library NRClex.

The primary project goals are to analyze sentiment in F1-related tweets and understand its relation with Formula 1 on-track results. This includes assessing emotion during races, determining their relationship with race results, and the effect of external factors. The study also delves into the influence of keywords and phrases, sentiment/emotion trends over time, and how sentiments/emotions change with drivers' grid positions. Moreover, it aims to assess the effect of sentiment on the dynamics of F1 racing and its social media presence.

Reference Work: [Linked Here] - This work does basic EDA on the F1 tweets data. We significantly expand on this with a coherent merge with race data and attempt to answer questions on sentiment relations with race results.

2. Data Sources

Primary Dataset: The "F1 Twitter Dataset" is a rich collection of Twitter data related to Formula 1, comprising more than 500,000 records with 13 columns. This dataset offers insights into F1-related Twitter activity, including user information, tweet content, and timestamps between 25 Jul'21 to 20 Aug'22.
The features of interest in this dataset are the tweet texts and the date of their creation.

Estimated Size: The dataset size is approximately 239.68 MB.

Location: The dataset is hosted on Kaggle, [Linked Here]

Format: The data is provided in CSV **Access Method:** Download

Secondary Dataset: This dataset provides a wide array of data on Formula 1 races, drivers, constructors, qualifying, circuits, lap times, pit stops, championships, and more, spanning from the inaugural season in 1950 to the latest available data in the 2023 season. The dataset consists of 14 CSV files, each with relevant information for in-depth F1 analysis.
The features of interest in this dataset are raceId, driverId, grid position, finishing position, quali_date, constructor points, round, year and race name.

Estimated Size: The dataset size is approximately 20.63 MB.

Location: The dataset is hosted on Kaggle, [Linked Here]

Format: The data is provided in CSV **Access Method:** Download

3. Data Manipulation Methods

First, we start by taking the races dataset and limiting it between 25 July 2021 to 20 Aug 2022 as we only have tweets for this time range. Then we join the races and results datasets on 'raceId' and then join this combined df with the drivers and constructors dataset on 'driverId' and 'constructorId' to get a combined dataset, named 'combined_Df', which contained all our races between the dates specified above and the driver and constructor name of all the results for each race.

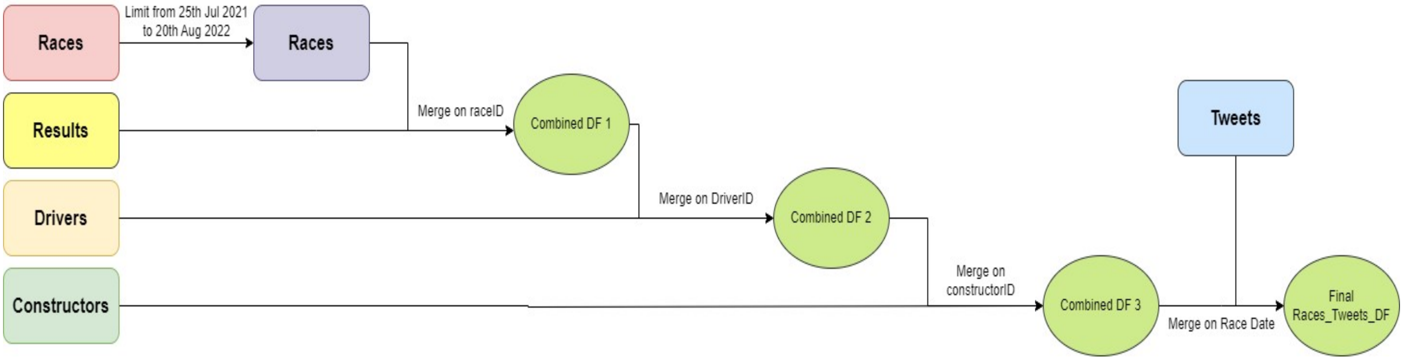


Figure 1: Flowchart of Data Merging

We take the tweets dataset and convert the data column to datetime format, dropping all rows which could not be converted to the datetime format. We also clean the tweets by removing URLs, '@', '#', punctuations, numbers and extra spaces. Then we merged the tweets dataset with 'combined_Df' to arrive at Races_Tweets_DF which we have used for gathering insights.

For each question, specific data manipulation is done as follows:
Q1)

- The races dataset was first pivoted with *index* as raceId, year, round, circuirId, race name, date, time, *columns* as driverName and *values* as position where position is the race position for drivers.
- Then drivers' columns were selected only for **Max Verstappen (Redbull), Lewis Hamilton (Mercedes) and Charles Leclerc (Ferrari)** and then it was inner joined with tweets dataset on date to come up with a merged dataframe with races and the three driver's positions for those races and the tweets for all those race days.
- Then a keyword search was done for the three drivers in the dataset and mentions of them in tweets on each race day were stored in columns as driver_mentions. So, the dataframe had drivers, their positions on races and the number of mentions. Sentiment analysis (positive/negative) was also done for tweets mentioning each driver and that was stored as **dominant sentiment** and **percentage of tweets** with the dominant sentiment out of positive and negative. The results have been plotted with matplotlib to showcase the results and make valid inferences.

Q2)

- First we found the drivers who **gained the most positions** in each race using the 'combined_Df' and this was joined with tweets data on the basis of race date.
- Then we found the **percentage of all tweets** for that race that mentioned that driver and took the mean for each driver over all races he gained the most positions.
- Then we found the **percentage mention** in the tweets each driver got on every race day and took the mean for each driver.
- Now we had the average percentage mentions a driver gets on the race day where he gained the most positions and also have the average mentions of a driver over our entire race's data and plot these 2 to discern any pattern.

Q3)

- First using the races dataset, we found raceIds for 2021. Then joined this with the constructor standing dataset and summed all the **constructor results using groupby** and made the constructor year results for 2021. Then using the results dataset and the drivers dataset we found the drivers for those constructors that finished 4-7.
- Once we had a list of drivers, then we found all tweets over 2021 that mentioned any alias of the driver and then did emotion analysis on those tweets using NRCLex. Then we finally plotted the emotion analysis scores and made our inferences.

Q4)

- From the races_merged_df, we created a pivoted df with columns as ‘DriverRef” and values as “Position”
- Then 4 columns(drivers) were selected from the pivoted_df ('latifi', 'russell', 'max_verstappen', 'perez') and put in a new df named selected_drivers_df.
- Now we merged the df created in step 2 with the tweets df on quali_date by performing an inner join between them. (**quali_merged_data** df created)
- Next we filtered out all the tweets from 2021, **cleaned the tweets** and filtered out tweets for redbull and williams F1 team by picking keywords for both.
- Lastly, we used the **NRCLex Sentiment Analyser** to get 10 sentiments for tweets for both the teams and did a **comparative analysis** for them

Q5)

- The date was limited to the Abu Dhabi Grand Prix, i.e. 2021-12-12 and then tweets were extracted in separate data frames for Max Verstappen and Hamilton using spacy’s tokenization and keyword extraction. Then sentiment analysis was done using NRCLex for both the drives and plots have been generated with plotly.

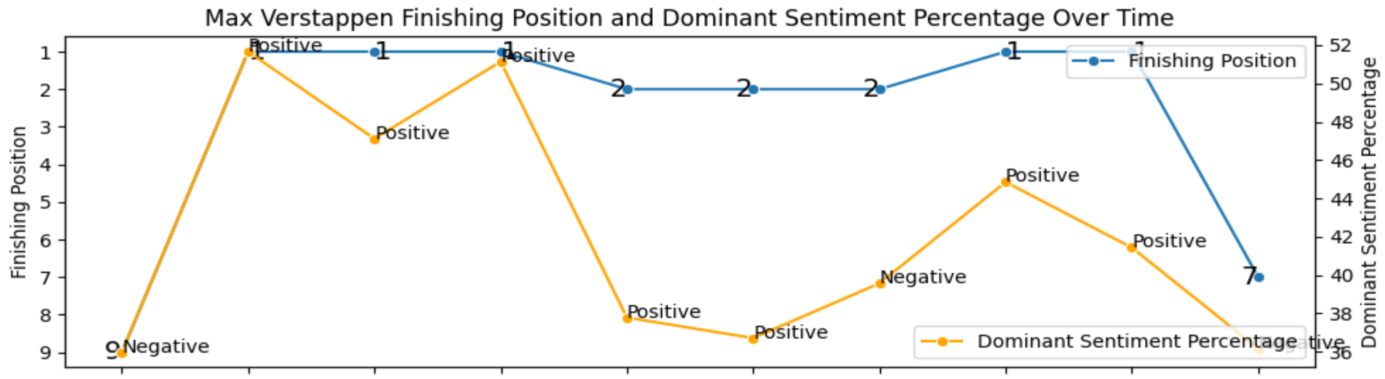
Q6)

- First we extracted the tweets from the tweets_df for the date 27th March 2022 and stored them in a new df
- Then we defined some keywords related to the Saudi Attack like “safety”, “missile”, “attack” etc.
- Now we filtered the tweets containing these keywords and stored them in a new df
- Then using the cardiffnlp/twitter-roberta-base-sentiment from hugging face, performed sentiment analysis on these tweets and categorized them into negative, neutral and positive sentiments
- Lastly, we did some analysis on the sentiments like **word cloud**, bar plots etc.

4. Analysis & Visualization

Q1) How is Finishing Position of the top 3 F1 drivers related with the sentiment in the tweets for them over each race?

Significance: We want to capture the fans’ sentiments on twitter during a race with the finishing performance of the top 3 drivers of F1: Max Verstappen, Lewis Hamilton and Charles Leclerc.



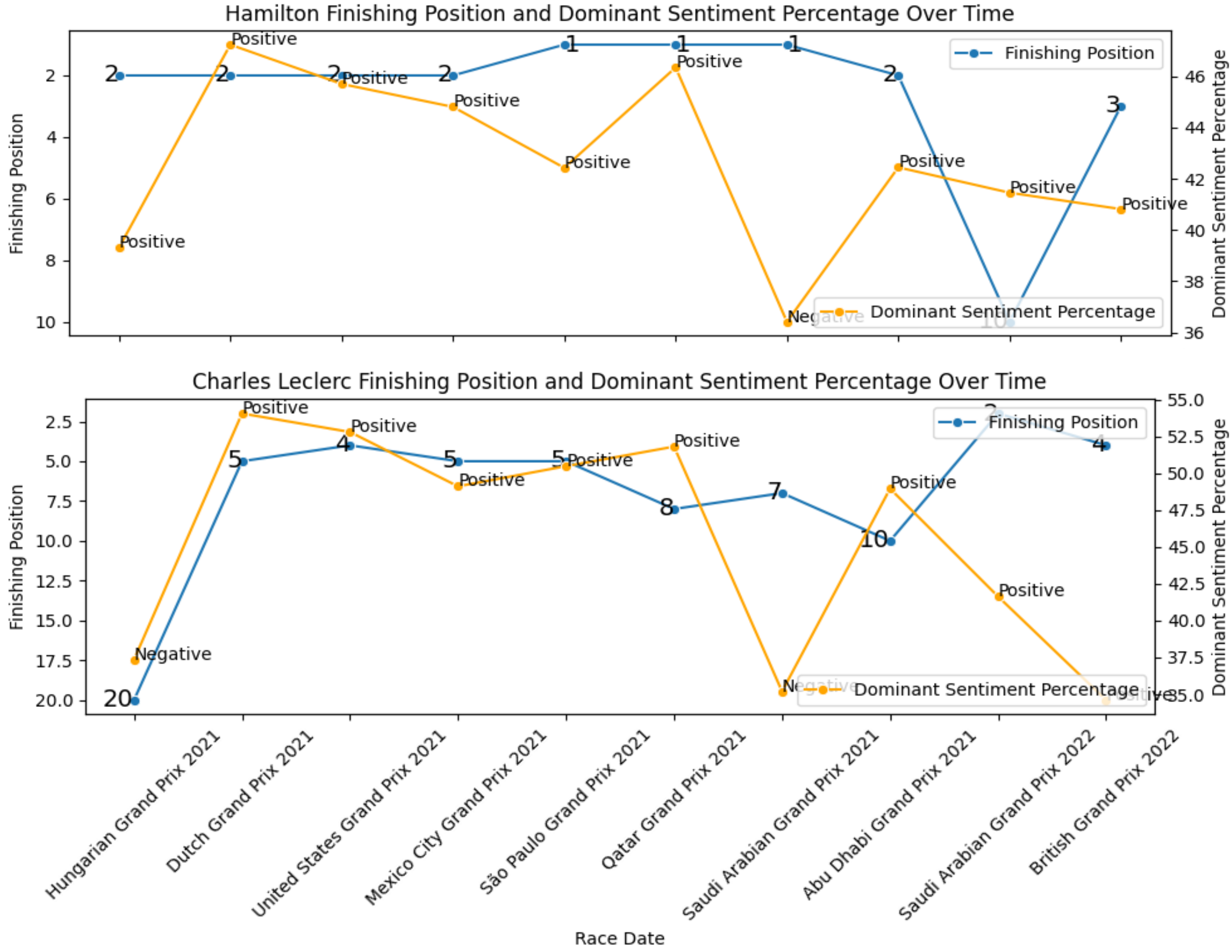


Figure 2. Drivers Positions and Associated Dominant Sentiment Percentage

Inference from the above plot for driver's Race Positions vs Tweet Sentiments:

1. Overall we observe that the sentiment for the top 3 players is strictly correlated with their finishing positions, as the dominant sentiment turns to negative if any of these top 3 drivers finish below the top 5 mark, which can be understood as their fans always have a high expectations from them. Also, the percentage of positive sentiment in the tweets rises or drops with the driver's performance and hence is a strong function of their results.
2. Special negative sentiment case for the race of Saudi Arabian'21 : All the three players, Max, Hamilton and Leclerc have negative sentiments showing in the tweets due to the eccentric nature of this race. From the start of the race, Max and Hamilton were having a brutal fight with each other. Shortly into the race, Max made an illegal overtake on Hamilton by going off the track which was condemned by everyone. Later in lap 37, Max Verstappen made an illegal overtake on Hamilton, for which he was asked to let Hamilton pass by. Also, Max and Hamilton had a slight collision leading to negative sentiment for both. Charles Leclerc was also involved in an incident shortly into the race and despite starting 4th on the grid, could only finish 7th in a Ferrai. That was nothing short of a disappointment for Ferrari Fans and hence his sentiment in the tweets also stayed negative.

Q2) Does most position gained in a race leads to higher trend in tweets on a race day compared to all other race days ?

Overtakes are some of the most exciting things to watch in an F1 race but does gaining the most positions in a race lead to a higher number of mentions on X or not ? We examine the percentage of mentions received by drivers who make the most position gains in a race, comparing it to the percentage of mentions they typically receive in an average race.

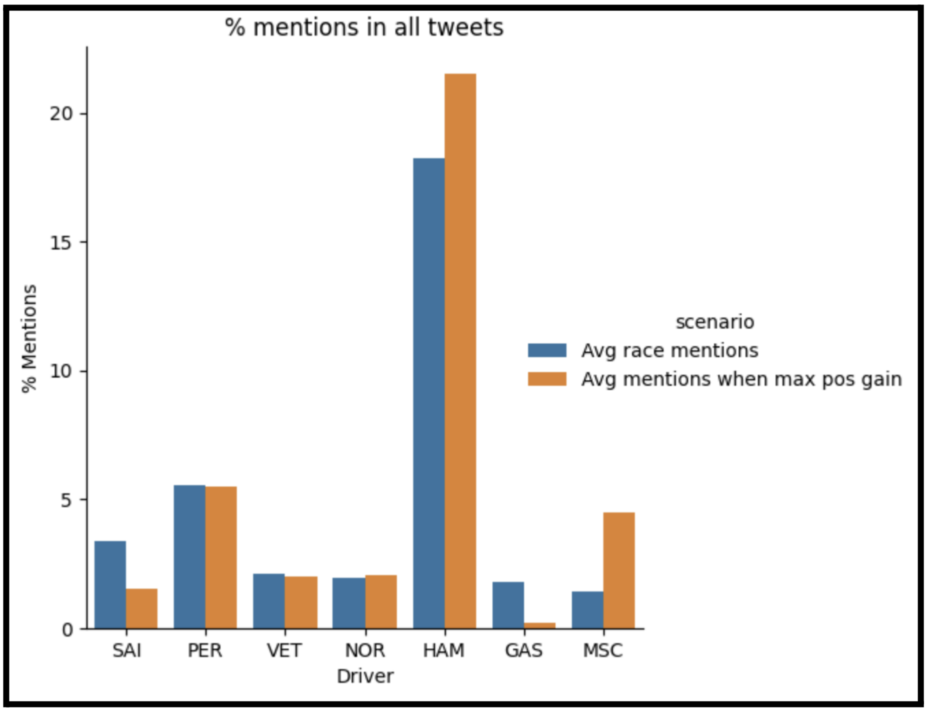


Figure 3. % Mentions of Drivers on Average Race Day vs on the Race with Most Positions Gained

We can see that for the majority of drivers there is no significant difference in the amount of mentions a driver gets on an average race day vs the races where he gains the most positions. This could be due to the fact that F1 usually does not highlight the driver gaining most positions in a race unless they come from way back in the field to the front. We do have 2 drivers though in Hamilton and Schumacher who are showing an increase in mentions though, let us understand why for each:

1. Lewis Hamilton (HAM) : There are 2 races in our dataset where Hamilton gained the most positions, the 2021 Brazil Grand Prix and the 2022 Saudi Grand Prix. In the 2021 Brazil Grand Prix, Hamilton famously started from last position and still managed to win the race. This amazing performance meant that he received an overwhelming amount of mentions on twitter, pushing his average higher.
2. Mick Schumacher (MSC) : In the 2022 British Grand Prix, Schumacher gained the most positions in the race but his increase in trend is most likely due to the fact that it was his first point scoring finish in F1 for which he received increased attention.

We also have Pierre Gasly (GAS) who actually received less % of mentions in the race where he gained the most positions. This is because Gasly gained the most positions in the 2021 Abu Dhabi Grand Prix where all attention was on Hamilton VS Verstappen. Thus there was actually a decrease in the % of mentions he usually gets.

Q3) What is the emotion towards Mid-order drivers for the 2021 season ?

Track position and race results are some of the most important things in an F1 race. We analyze the emotions of people towards the drivers who are neither at the top of the table nor are they at the bottom but can still cause chaos in a race and also win some races.

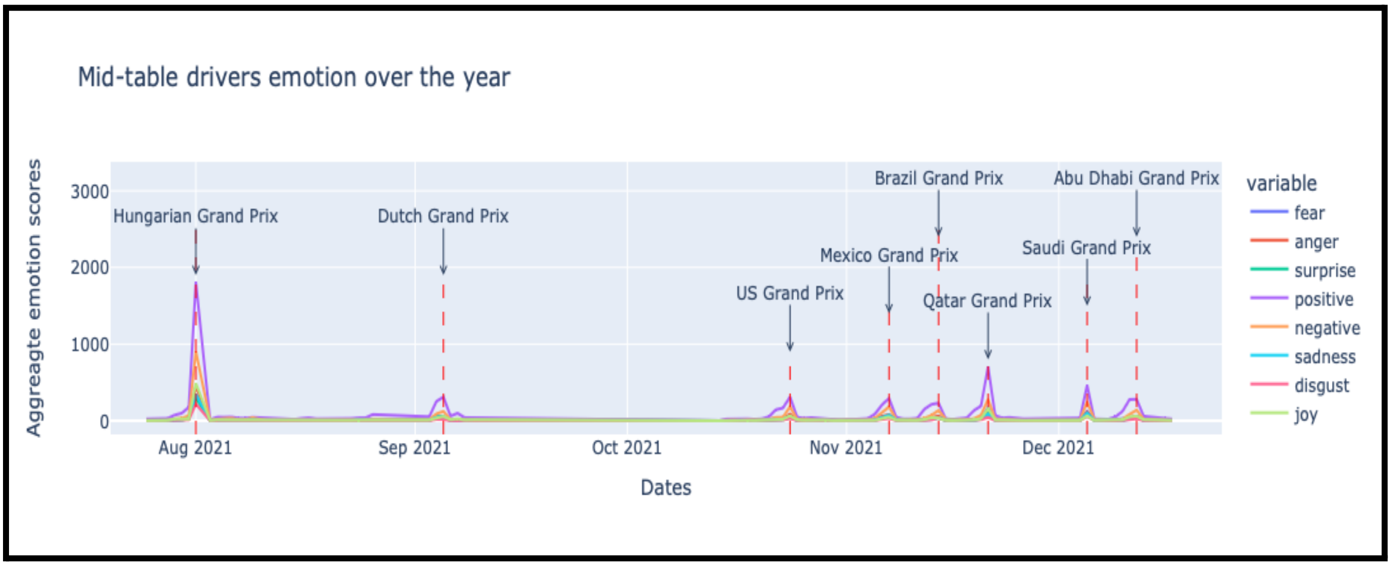


Figure 4. Mid-order Drivers' Emotions Over the Year for 2021 Season

We can see that over the entire 2021 season, the dominant emotion was always that of positive towards the mid-order drivers. One way to understand this would be that as these are mid-order drivers, the spikes in emotion scores typically happen when they do something significant to affect the race results. Like in the 2021 Hungarian Grand Prix, where Esteban Ocon (Alpine) won over Hamilton and Fernando Alonso (Alpine) managed to hold Hamilton behind him for a long time, and then in the 2021 Qatar Grand Prix, where Alonso finished third.

Apart from these over an average race day, as they get very low screen time and usually are not part of any significant narrative, most tweets would just be from the fans of each driver and thus it is understandable that they are majorly positive towards their favorite driver.

After positive we can see that the other major emotions are of negative, surprise, joy, and fear. This shows that people are usually surprised or in joy when these mid-order drivers become part of some race-changing event and they also get negative tweets most probably from the fans of the leading drivers whose race results get affected from such events.

Q4) Is there a direct correlation between twitter sentiments of a Formula 1 team's qualifying vs race day performance ?

This question seeks to examine the relationship between the sentiment of tweets on qualifying positions and race day results of specific teams, namely Red Bull and Williams. The analysis will compare the Twitter activity on both qualifying days and race days to determine if there's a noticeable difference in social media sentiment |based on the team's on-track performance.

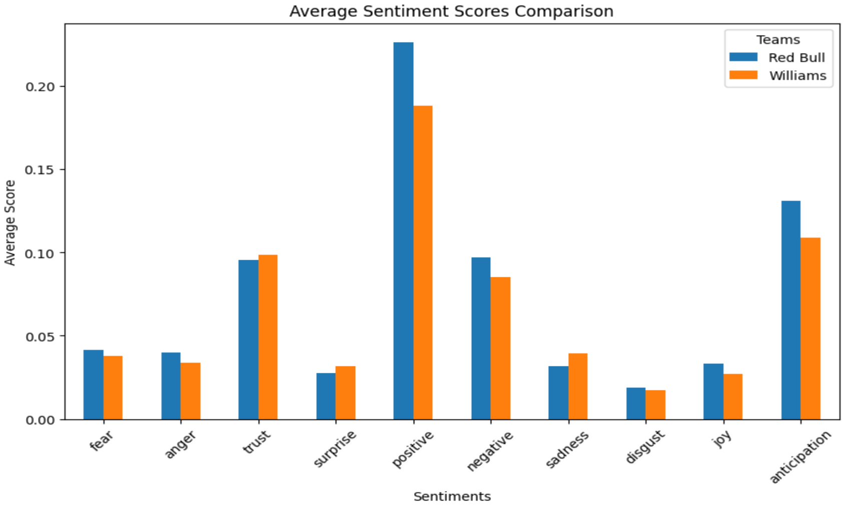


Figure 5. Average Sentiment Score Comparison of Red Bull & Williams on Qualifying Day

The sentiment analysis graph, reflecting Twitter data during the 2021 F1 qualifying sessions, captures the fan reactions toward Red Bull and Williams. Red Bull not only achieved higher average positive sentiment scores, suggesting fans were pleased with their performance, potentially due to their competitive lap times and pole positions, but they also saw heightened levels of joy, possibly in response to their drivers'—such as Max Verstappen's—strong qualifying finishes.

Williams, on the other hand, while having lower positive sentiment, ranked higher in trust, indicating fans held a cautious optimism for their race day prospects, perhaps inspired by their historical underdog status or standout qualifying efforts like those from George Russell. Their trust scores slightly outpaced Red Bull, suggesting a solid fan belief in their reliability, despite a slight uptick in negative sentiments like anger and sadness, which could be tied to close calls or qualifying mishaps.

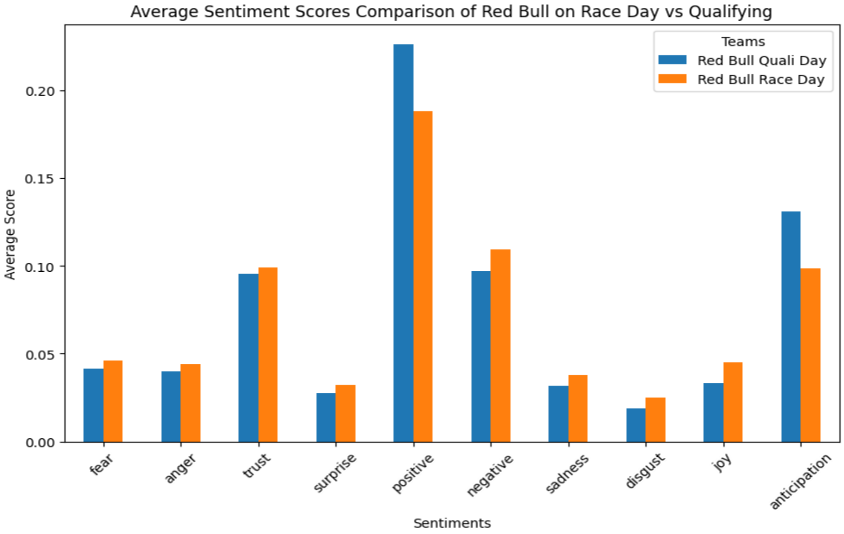


Figure 6. Average Sentiment Score Comparison of Red Bull

The sentiment analysis for Red Bull during the 2021 F1 season reveals that the team was met with greater positive sentiment on race days (just under 0.20) than on qualifying days (just above 0.15). Similarly, joy was higher on race days, indicating fans' happier reactions to the actual races compared to the qualifiers. Anticipation was more elevated during qualifying days, suggesting higher expectations ahead of the races. Trust sentiment was slightly stronger on race days, whereas negative sentiments remained low but were a tad higher during qualifying. Overall, the data suggests that Red Bull's race performances resonated more positively with fans, with race days seeing a stronger approval than qualifying days.

Q5) How did the emotions in F1-related tweets evolve during the 2021 Abu Dhabi Grand Prix, the most-watched race in F1 history, with the changing race lead?

Significance: The Abu Dhabi Race was quite an interesting one as it was the last race of the 2021 season and the tussle for the Driver's Championship was still on between Lewis Hamilton (Mercedes) and Max Verstappen (Redbull). In the race as well, the fight went on till the last lap where, with an interference of FIA's race director Michael Masi, a controversial verdict was given by the FIA after the safety car deployment which advantaged Max Verstappen to take the race home in the very last lap and winning the maiden Driver's Championship of his buzzing career.
So, a lot went on in this race, and we try to capture it through the sentiment analysis of the fans and users on twitter by focusing strictly on the tweets posted during the race +/- 1 hour.

Sentiment Plot for Max Verstappen (Redbull):

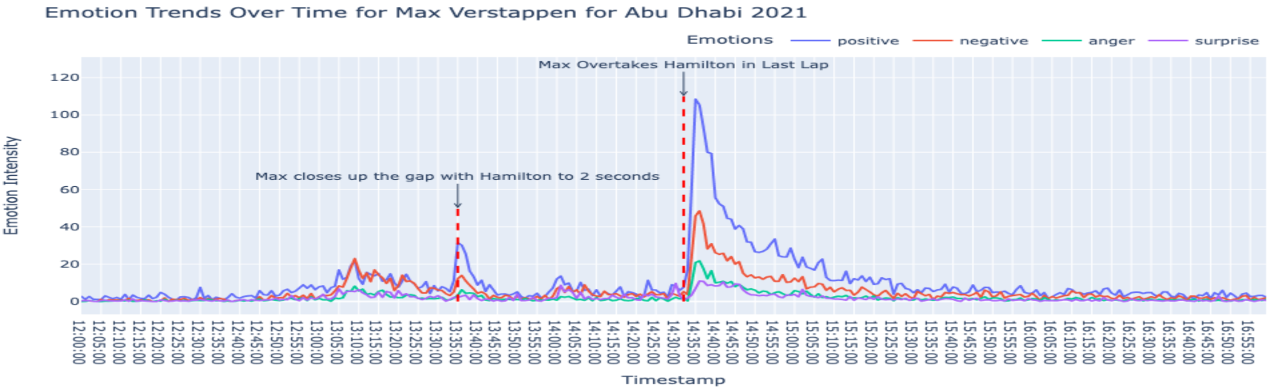


Figure 7. Emotion in Tweets for Max Verstappen during 2021 Abu Dhabi Grand Prix

<u>Inference for Fig. 1 for Max Verstappen:</u>

1. At the start of the race at 13:00:00, though Max started first on the grid however lost his lead in the starting few seconds to Hamilton and hence a rise of negative tweets is observed for him. After a pit stop, at 13:35:00 in Lap 21, while Perez kept on tussling with Hamilton to hold him for Verstappen to bridge the gap, and Max was able to close the gap to 2 seconds emerging as a potential winner as half of the race still remained, leading to a surge in positive sentiment for Max.
2. At the thrilling moment of 14:35:00, in the very last lap 58, Max overtook Hamilton to give his fans a burst of joy and hence the significant rise in positive sentiment in the tweets for Max Verstappen.

Sentiment Plot for Lewis Hamilton (Mercedes):

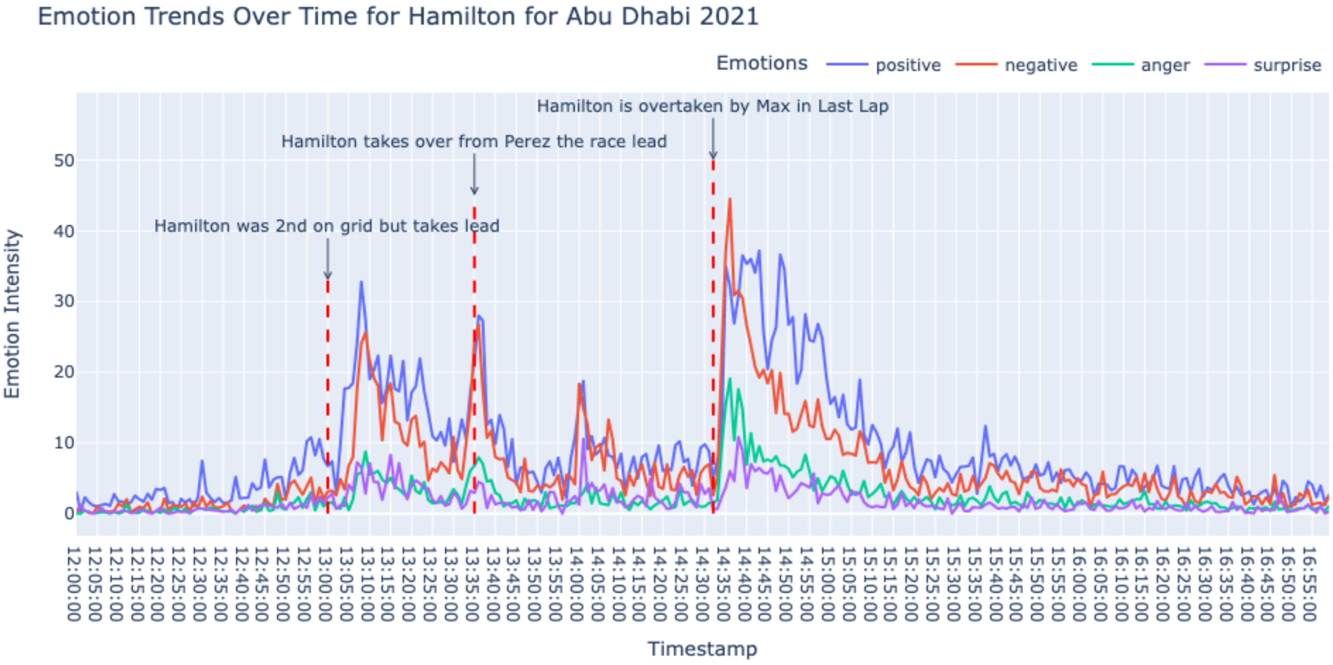


Figure 8. Emotion in Tweets for Lewis Hamiton during 2021 Abu Dhabi Grand Prix

<u>Inference for Fig. 2 for Lewis Hamilton:</u>

1. At the start of the race at 13:00:00, Hamilton started second but took the lead right at the start of the race but Max tried to come back furiously with a close shave collision. That led to a divisive sentiment in tweets for Hamilton.
2. At 13:35:00 in Lap 21, Hamilton overtook Perez after a lengthy tussle between the two. Hence the nature of the tweets for Hamilton remained high in both positive and negative.
3. On lap 53 a Williams Driver crashed leading to the safety car being deployed. The safety process went on till lap 58, 14:32:00, when the race began with a controversial verdict by FIA which

advantaged Max, enabling him to overtake Hamilton on the last lap of the race and win the Driver's Championship for the first time, leading to a spike in negative and anger filled tweets among Hamilton's fans.

4. However, despite the wrong protocol by FIA, which cost Hamilton the championship, the way Hamilton handled the situation without creating any fuss, he won the hearts of his fans again and gathered a lot of new fans for himself. This is validated by a persistent positive sentiment in tweets for Hamilton.

Q6) What specific words, phrases, or hashtags in Formula 1-related tweets can be linked to positive or negative sentiments, and how do these correlate with particular occurrences in the sport?

In the backdrop of the 2022 Saudi Grand Prix, an unprecedented event unfolded as a missile strike targeted an oil facility near the Jeddah race circuit. This alarming incident sent shockwaves through the Formula 1 community, casting a shadow over the weekend's racing activities. The tense atmosphere was palpable both on and off the track, as the safety of the event and its participants was brought into question amidst the high-stakes environment of an F1 race weekend.

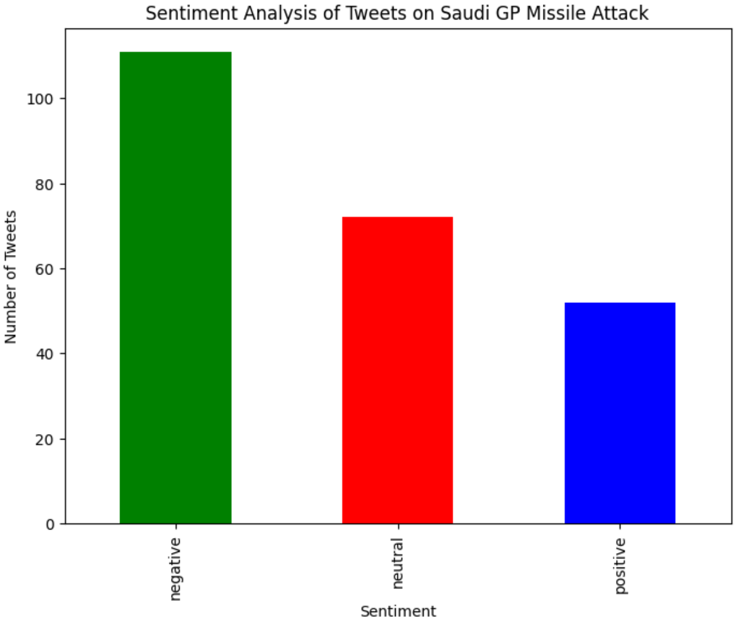


Figure 9. Categorization of Tweets Related to Saudi Missile Attack Based on Different Sentiments

The graph shows that negative sentiments dominate, with the highest number of tweets. This reflects the concerns and shock within the Formula 1 community over the security risks posed by such a nearby attack. Neutral sentiment tweets, which may reflect news reporting or factual updates without emotional commentary, are slightly fewer. Positive sentiment tweets are the least, which could indicate expressions of relief or gratitude for the safety of individuals.

Real-life scenarios that occurred during the Saudi attack include the immediate response by the F1 authorities and the drivers' discussions about whether to proceed with the event. Drivers like Lewis Hamilton and Max Verstappen, leading figures in the sport, were likely involved in discussions about the race's continuance.

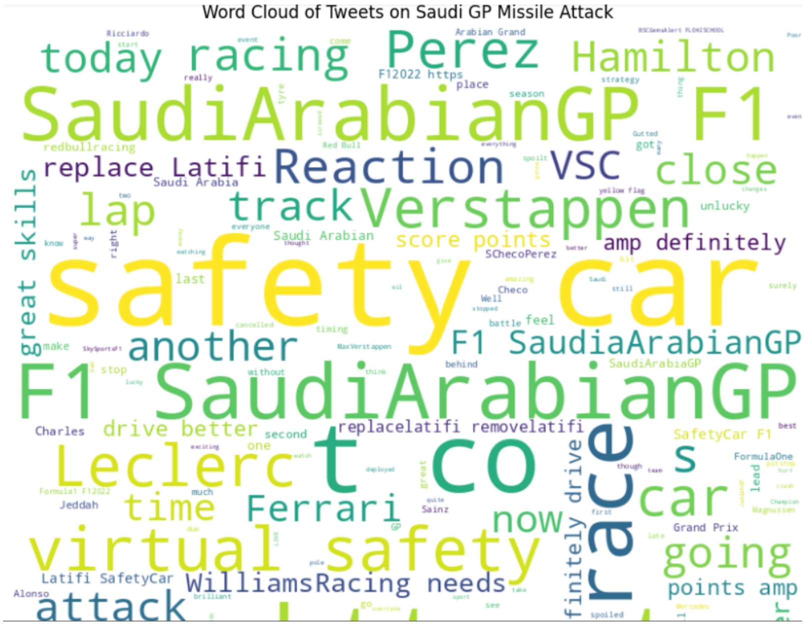


Figure 10. Word Cloud of the most occurring word in filtered tweets related to Saudi GP

The word cloud generated from tweets about the Saudi GP missile attack reveals a focus on "safety" and driver names like "Leclerc" and "Hamilton," indicating significant discussion around driver safety and reactions. The prominence of words such as "attack," "reaction," and "car" suggests a high level of engagement with the incident's impact on the race and the drivers.

5. Statement of Work

In this collaborative project, each team member played a pivotal role in its success. Regular meetings were conducted to discuss strategies and steps, ensuring a cohesive approach. While individual responsibilities included tackling specific questions, one team member took the lead in establishing common steps and standards, fostering consistency across all components. The synergy created through our effective communication and shared efforts greatly contributed to the overall achievement.

Each one of us did 2 questions comprehensively:

Sudhanshu - Q.4. and Q.6.

Shitanshu - Q.2. and Q.3.

Divyam - Q.1. and Q.5

Once each one of us came up with the steps required to solve our respective questions we found the common steps required between all questions and did them just once to avoid repetitive efforts and to standardize code among us all.

In future work, we can plan our timelines ahead, and use better collaborative tools as this time we struggled a little in coherent collaboration with Google Colab and Deepnote. We would expect to find and then make use of better collaborative tools.

We also had a few issues while merging our code in this collaborative work, for the future, we will try coming up with coding checkpoints and milestones, on accomplishing which we will sync up our code using git.