Technical Report: Final Project DS 5110: Introduction to Data Management and Processing

Team Members: Sanjiv Motilal Choudhari, Shivram Nekkanti
Khoury College of Computer Sciences
Data Science Program
motilalchoudhari.s@northeastern.edu
nekkanti.sh@northeastern.edu

December 10, 2024

Contents

1 Introduction

Formula 1 (F1) is one of the most popular and emotionally charged sports in the world, with millions of fans actively engaging with teams, drivers, and events on social media. The 2024 F1 season has provided numerous opportunities to analyze fan behavior and sentiment, offering insights into how audiences perceive racers, teams, and the dynamics of individual races. This project leverages Natural Language Processing (NLP) techniques to perform a comprehensive sentiment analysis of fan comments collected from YouTube videos related to 20 F1 races.

The primary objective of this project is to uncover emotional trends such as joy, sadness, anger, surprise, and neutrality, and to assess how they vary across races and teams. By analyzing 100,000 comments, the study identifies key drivers of fan engagement, sentiment polarity, and team perceptions. Furthermore, the project aims to provide actionable insights that can guide teams, sponsors, and F1 management in enhancing fan experiences and improving audience relationships.

The methodology involves scraping YouTube comments, cleaning and preprocessing the data, and applying sentiment analysis using lexicon-based tools like the NRC Lexicon and VADER. The results are visualized using tools like Matplotlib, Seaborn, and Tableau, allowing for a detailed understanding of sentiment trends over time.

This analysis not only highlights the emotional connections between fans and F1 events but also demonstrates the value of social media as a data source for understanding audience behavior. The findings pave the way for future research in sentiment analysis across multiple languages and platforms.

The objective of the project is to apply Natural Language Processing (NLP) techniques to extract meaningful information from the comments. This includes identifying mentions of racers and racing teams, classifying emotions such as joy, anger, sadness, surprise, and fear, and understanding how fans perceive individual drivers and teams over time.

The scope of the project encompasses four key areas:

- **Data Collection and Preprocessing:** Scraping and cleaning comments to ensure they are relevant and ready for analysis.
- **Sentiment Analysis:** Utilizing sentiment analysis tools like NRC Lexicon to classify comments into distinct emotions and gauge audience sentiment.
- **Visualization:** Creating insightful visualizations to represent sentiments, trends, and team-based perceptions over time.
- **Deliverables:** Summarizing findings in a technical report, presenting insights through Power BI/Tableau dashboards, and crafting a compelling presentation for stakeholders.

By focusing on both individual racers and teams (e.g., Red Bull, Ferrari, McLaren, Mercedes), this project aims to provide actionable insights into fan engagement, sentiment distribution, and the factors influencing public opinion within the Formula One community. The results will help inform strategies for race organizers, teams, and sponsors to enhance fan experiences and brand value.

2 Literature Review

- **Sentiment Analysis in Sports Analytics:** Studies in sports analytics often use Natural Language Processing (NLP) to analyze fan sentiment from social media platforms such as Twitter, YouTube, and Reddit. Tools like NRC Lexicon and VADER are commonly applied to classify emotions (e.g., joy, anger, sadness) and identify trends. For example, researchers have observed that positive sentiment spikes after wins, while losses trigger a rise in anger and sadness.[1][5]

- **Fan Engagement and Brand Analysis in Formula 1:** Analysis of Formula 1 fan engagement highlights the importance of real-time feedback during races. Comments and reactions reveal public opinions about drivers, teams, and race-day events. A few papers examine team mentions to correlate brand perception with race outcomes.[2]

- **Innovative Approaches in Sports Science:** This study explores the application of lexicon-based sentiment analysis tools to sports-related social media data, particularly football matches. It highlights how tools like NRC Lexicon can classify emotions with high accuracy when applied to real-time Twitter data. The findings suggest that sentiment analysis is a valuable tool in sports science to gauge public sentiment during events, making it relevant to your analysis of F1 fans.[3]

- **Sports Fans' Behavior on Twitter:** This research analyzed fan sentiments during the 2018 FIFA World Cup Final using Twitter data. It emphasized the use of big data analytics to capture large-scale audience emotions and engagement. The methodology aligns with your approach of using social media to track real-time emotional trends and demonstrates how sentiment analysis can help predict audience behaviors.[4]

Gaps Identified:

- Limited focus on sport-specific terminologies and deeper sentiment dimensions such as surprise or fear.

- F1 Sporting events analysis are an under explored field of analysis

- Many studies are limited to quantitative metrics such as the volume of comments without delving into the qualitative sentiment nuances.

- Few studies explore real-time sentiment trends during live sports events, missing the opportunity to capture dynamic fan reactions

- Most lexicon-based tools, such as NRC Lexicon, fail to detect sarcasm or nuanced language, which are common in social media comments. Advanced machine learning models for sentiment analysis are underutilized in sports research

3 Methodology

This project employs a comprehensive approach to analyze fan sentiment surrounding Formula 1 races in the 2024 season. By leveraging Natural Language Processing (NLP) techniques, including lexicon-based and transformer-based sentiment analysis, the project captures emotional reactions of fans to racers, teams, and events. Data was collected from over 100,000 YouTube comments related to 20 races, processed to ensure relevance, and classified into multiple emotional categories: **joy, sadness, anger, surprise, and neutral**. Advanced models were used to handle multilabel sentiment classification, where a single comment expressed multiple sentiments toward different drivers. The findings were visualized using Python libraries and Tableau for interactive dashboards.

3.1 Data Collection

- **Source:** Comments were scraped from YouTube videos covering races in the 2024 Formula 1 season, chosen for their high engagement and relevance to the audience. The data captured represents fan opinions about racers, teams, and race-day events.
- **Volume:** The dataset consists of over 100,000 comments across 20 videos, providing a robust sample for sentiment analysis. Each video corresponds to a specific race, ensuring the dataset reflects the full season's events.
- **Tools:** Data scraping was implemented using Python libraries:
  - **BeautifulSoup:** For extracting comment data and metadata (e.g., timestamps, like counts) from the YouTube pages.
  - **pandas:** For structuring, storing, and manipulating the scraped data into a tabular format (CSV).

3.2 Data Preprocessing

- **Cleaning:**
  - Removed stop words (e.g., "the", "and", "but") using the **nltk** library to retain meaningful words for analysis.
  - Eliminated special characters, emojis, and non-alphanumeric symbols using **regular expressions** to standardize the text.
- **Filtering:**
  - Comments were filtered for relevance by searching for keywords related to specific racers and teams (e.g., "Hamilton", "Red Bull").
  - Irrelevant comments, such as generic compliments or unrelated discussions, were excluded.
- **Handling Missing Data:**
  - Rows with null or irrelevant fields (e.g., missing text, zero likes) were removed.
  - For incomplete comments that contained racer or team mentions, missing fields were corrected manually to preserve meaningful data.

- **Normalization:**
  - Converted all text to lowercase to ensure uniformity during keyword matching.
  - Standardized date and time fields for consistency in temporal analysis.

3.3 Analysis Techniques

- **Tools:**
  - **NRC Lexicon:** Used for initial lexicon-based sentiment classification into categories: **joy, sadness, anger, surprise, and neutral**.
  - **VADER Sentiment Analysis:** Used as a secondary tool for evaluating polarity scores (positive, negative, neutral) as a baseline.
  - **Hugging Face Transformers:** Employed pre-trained transformer models like `bert-base-uncased` for advanced sentiment analysis. These models provided nuanced sentiment detection, especially for multilabel classifications where a single comment addressed multiple drivers or teams.
- **Handling Multilabel Sentiments:**
  - Comments referencing multiple drivers or teams were analyzed using context-aware transformer models to assign distinct sentiments for each mentioned entity.
  - For example, in a comment like *"Hamilton drove brilliantly, but Verstappen's move was reckless"*, the analysis assigned **joy** for Hamilton and **anger** for Verstappen.
- **Process:**
  - Tokenized comments using **nltk** and Hugging Face's tokenizer to prepare them for analysis.
  - Performed sentiment classification using the NRC Lexicon and fine-tuned transformer models to map words and phrases to their respective emotional scores.
  - Aggregated multilabel sentiments and stored results for each racer or team to enable detailed trend analysis.

3.4 Visualization

- **Python Libraries:** Visualized sentiment trends and engagement metrics using **Matplotlib** and **Seaborn**.
- **Interactive Dashboards:** Developed a user-friendly dashboards using PySaprk to showcase team and driver-specific sentiment trends.

This methodology incorporates both traditional lexicon-based techniques and advanced transformer models to handle complex multilabel sentiment scenarios. The approach ensures comprehensive and nuanced insights into fan sentiment for Formula 1 events.

4 Results

4.1 Emotion Distribution for Each Racer

Output: This bar plot shows the distribution of emotions (joy, surprise, neutral, sad, anger) expressed in comments for each racer. It helps to visualize which emotions are most prevalent for different racers and identify potential trends or patterns.

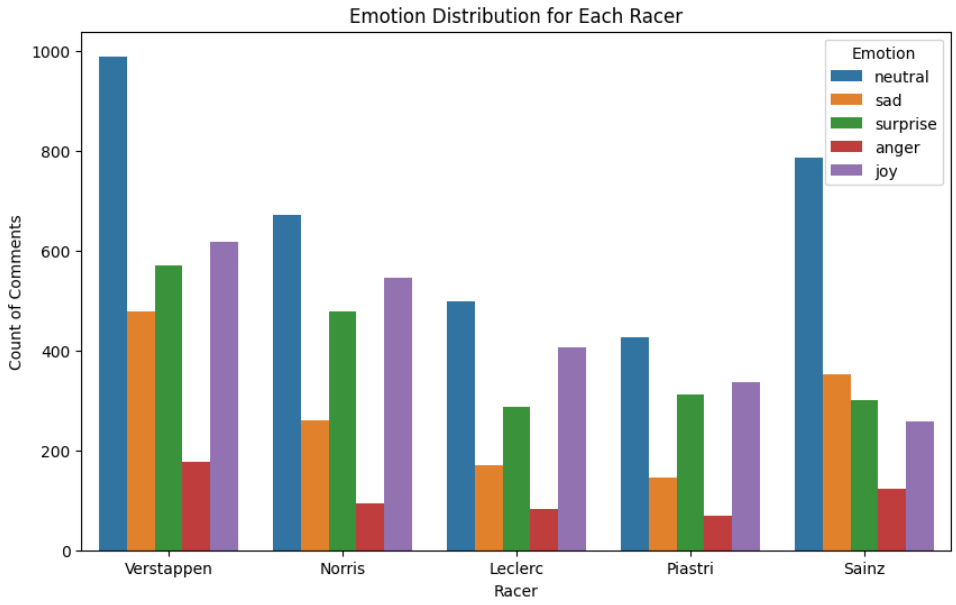

Figure 1: Sentiments for each racer

Significance: Race organizers could use this to understand fan sentiment towards specific drivers, allowing for targeted marketing campaigns or driver promotions. This data could inform driver interactions with fans, helping to cultivate a positive public image.

4.2 Top Emotions for Each Racer

Output: This stacked bar plot (Figure 2) illustrates the top emotions associated with each racer. It provides a clear comparison of the dominant emotions expressed towards different racers and helps to understand the overall sentiment associated with them.

Significance: Teams and sponsors could use this data to gauge driver performance perception. If a driver consistently evokes negative emotions, it signals a need for performance improvement or PR strategy adjustments. This allows for targeted interventions to improve driver image and fan support.

4.3 Top Racers by Positive Sentiment

Output: This bar plot (Figure 3) highlights the top racers who received the most positive comments (associated with the emotion "joy"). It provides insights into which racers are generally perceived more positively by users.

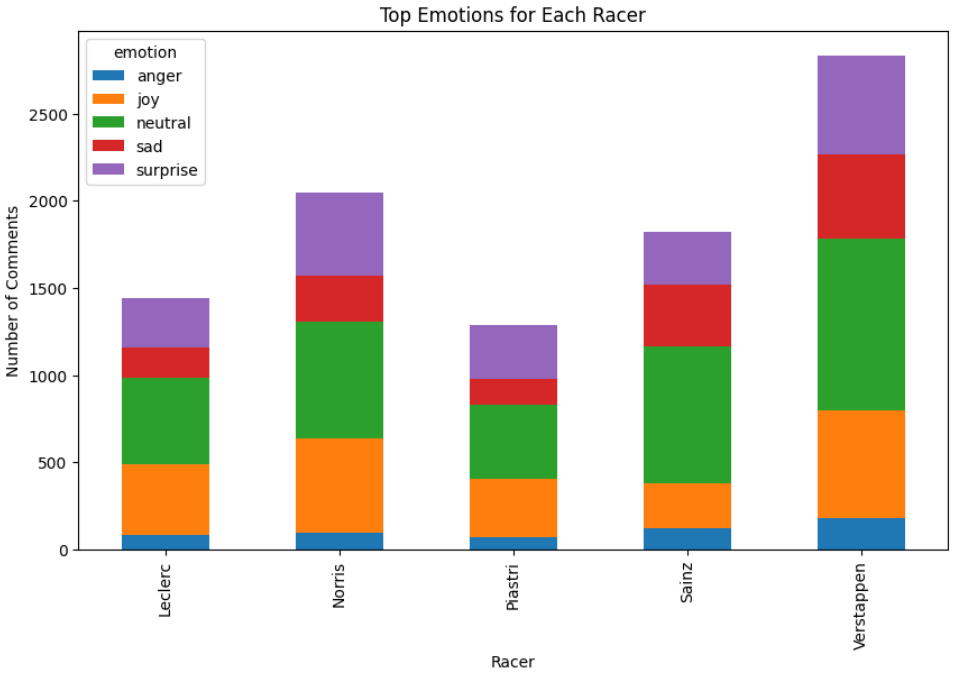

Figure 2: Stacked Barplot showing sentiment distribution

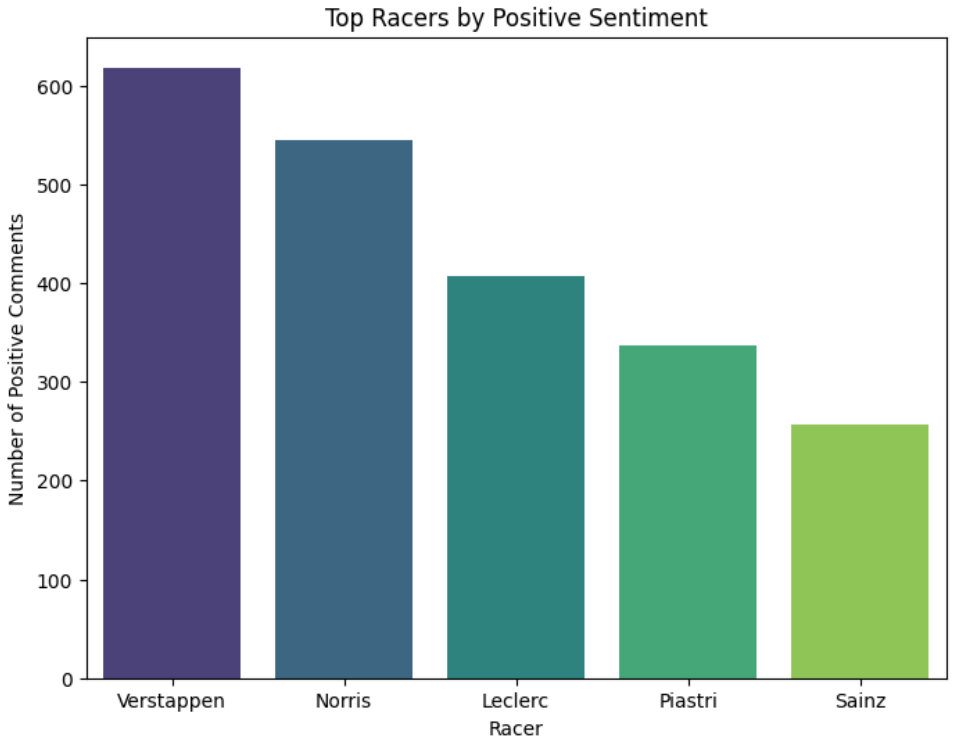

Figure 3: Racers with most positive sentiments

4.4 Top Racers by Negative Sentiment

Output: This bar plot highlights the top racers who received the most negative comments (associated with the emotion "anger" or "sadness"). It provides insights into which racers are generally perceived negatively by users.

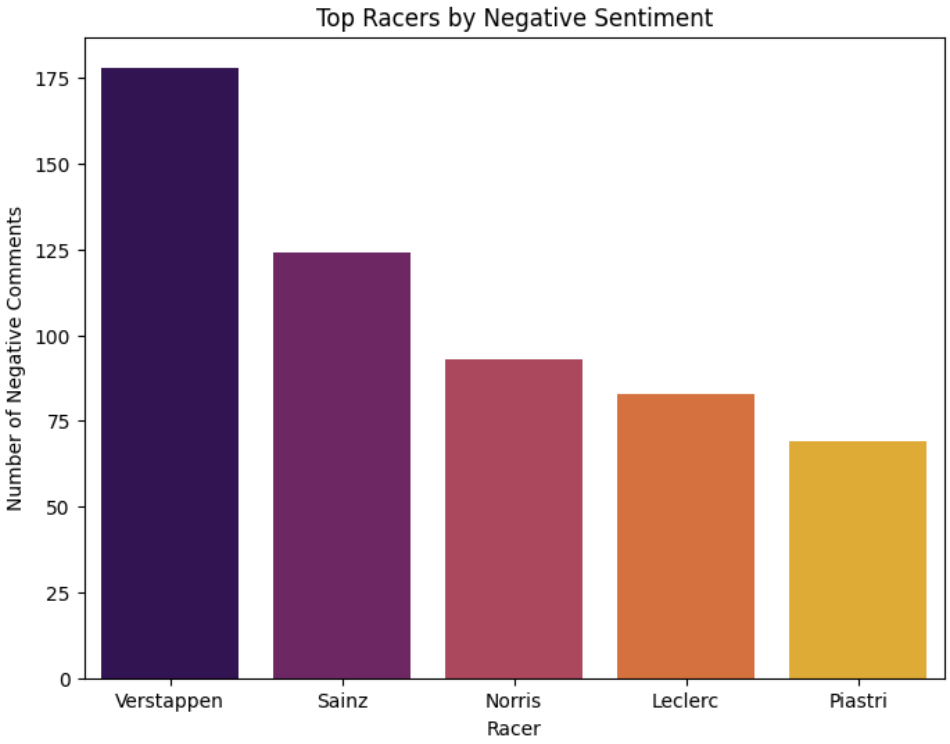

Figure 4: Racers with most negative sentiments

Significance: Media outlets could use these visualizations to create compelling narratives about popular and unpopular drivers, driving engagement and viewership. This data-driven approach to storytelling resonates with fans and fuels discussions within the motorsport community.

4.5 Sentiment Distribution by Constructor

Output: This box plot (Figure 5) presents the distribution of sentiment scores for each constructor (e.g., Red Bull, McLaren, Ferrari). It helps to compare the overall sentiment associated with different constructors and identify potential differences in user opinions.

Significance: Team managers can use this to identify areas needing improvement within the team. If one constructor has significantly lower sentiment, they might need to focus on race strategy, car development, or driver performance, ultimately improving fan satisfaction and team morale.

4.6 Word Clouds

Significance: Analysts can use this word clouds to gain a quick overview of the key topics and opinions driving fan sentiment. Highlighting positive and negative terms helps to better understand fan perceptions, inform future event planning, and identify areas for improvement in race management or marketing.

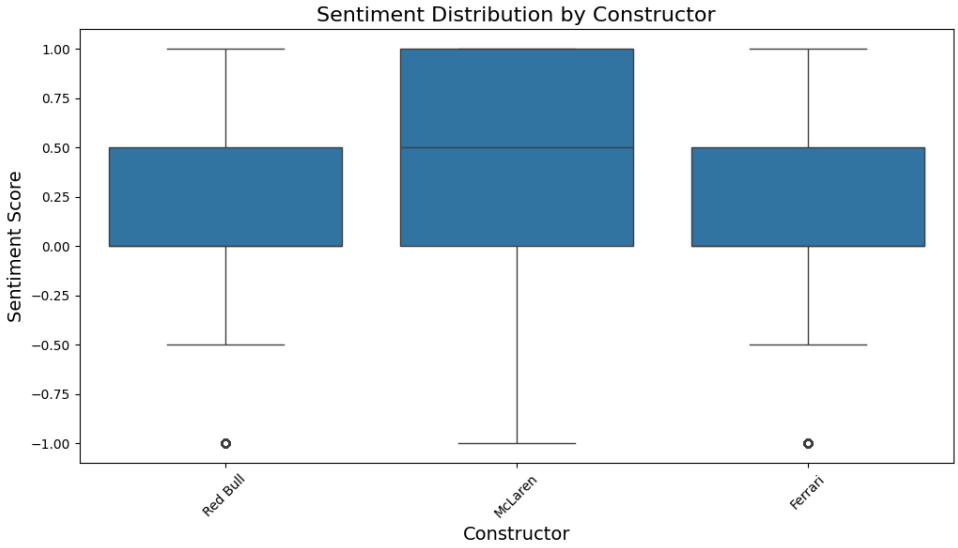

Figure 5: Boxplot for constructor sentiments

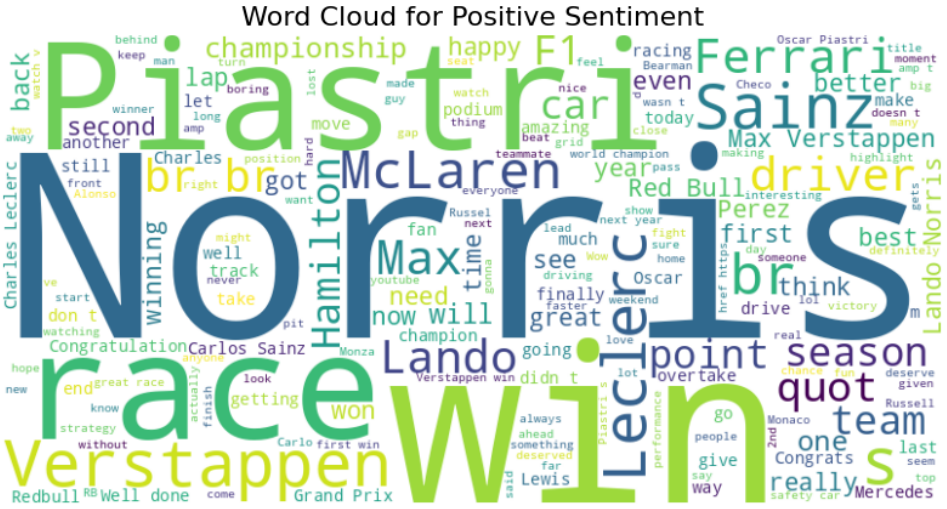

Figure 6: Word Cloud for Positive Sentiment

4.7 Sentiment Distribution for a Racer

Output: This pie chart shows the distribution of emotions (joy, surprise, neutral, sad, anger) expressed in comments for a racer. It helps to visualize which emotions are most prevalent for different racers and identify potential trends or patterns.

Significance: This pie chart (Figure 8) provides a comprehensive overview of a single driver's popularity and fan perception. Driver management teams can use this to track public perception and inform public relations efforts. Identifying patterns in sentiment allows for strategic interventions to enhance the driver's image and increase fan engagement.

4.8 Racer vs Constructor Sentiment Comparison

Output: This heatmap visualizes the average sentiment score for each racer within their respective constructors. It provides a comprehensive overview of sentiment towards racers

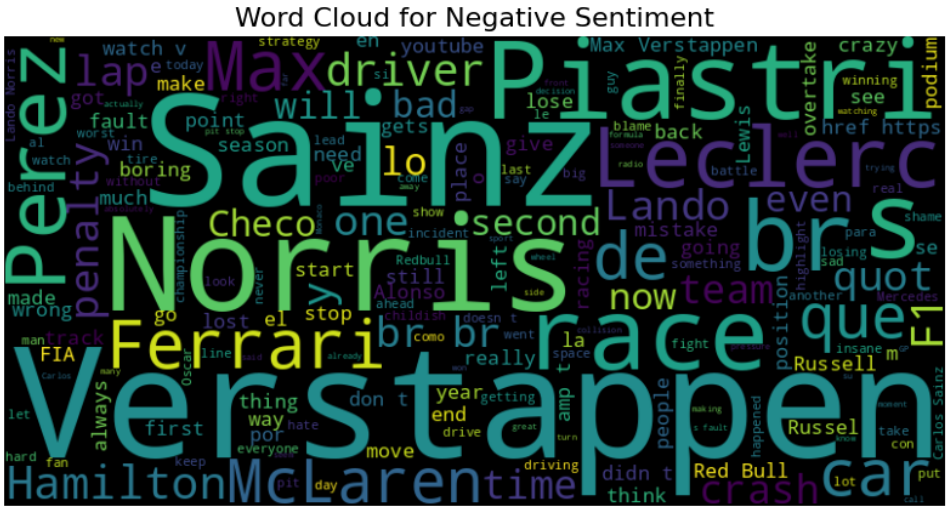

Figure 7: Word Cloud for Negative Sentiment

and constructors, highlighting potential relationships between the two.

Significance: Journalists can use this heatmap to highlight the relationship between driver popularity and team performance. This allows for nuanced analysis of race results and team dynamics, providing fans with deeper insights and potentially influencing their predictions for future races.

5 Discussion

5.1 Results Interpretation

The analysis of YouTube comments provided several key insights into fan sentiment regarding Formula 1 racers and teams. These insights reflect the emotional connection fans have with the sport and highlight patterns in their engagement:

- **Driver Perception:** Among the racers, Max Verstappen consistently received overwhelmingly positive sentiments, with "joy" being the predominant emotion in his mentions. This reflects his strong performance and dedicated fan base. Conversely, Carlos Sainz faced a higher proportion of negative sentiments, predominantly "anger" and "sadness," often tied to race incidents, strategic mishaps, or underperformance in critical moments.

- **Team Sentiment Trends:** Teams like Red Bull, McLaren, and Mercedes emerged as those with the most positive mentions, reflecting their strong performances throughout the season. On the other hand, Ferrari showed a significant mix of "sadness" and "anger," frequently linked to strategic errors or missed opportunities, which frustrated fans and impacted the team's perception.

- **Emotional Distribution:** Across all races, "joy" was the dominant emotion, reflecting overall fan enthusiasm and positive engagement with the 2024 season. However, emotions such as "anger" and "sadness" spiked in races with controversies or significant incidents. For instance, in specific races like Race A, controversial penalties or dramatic crashes led to a surge in negative emotions.

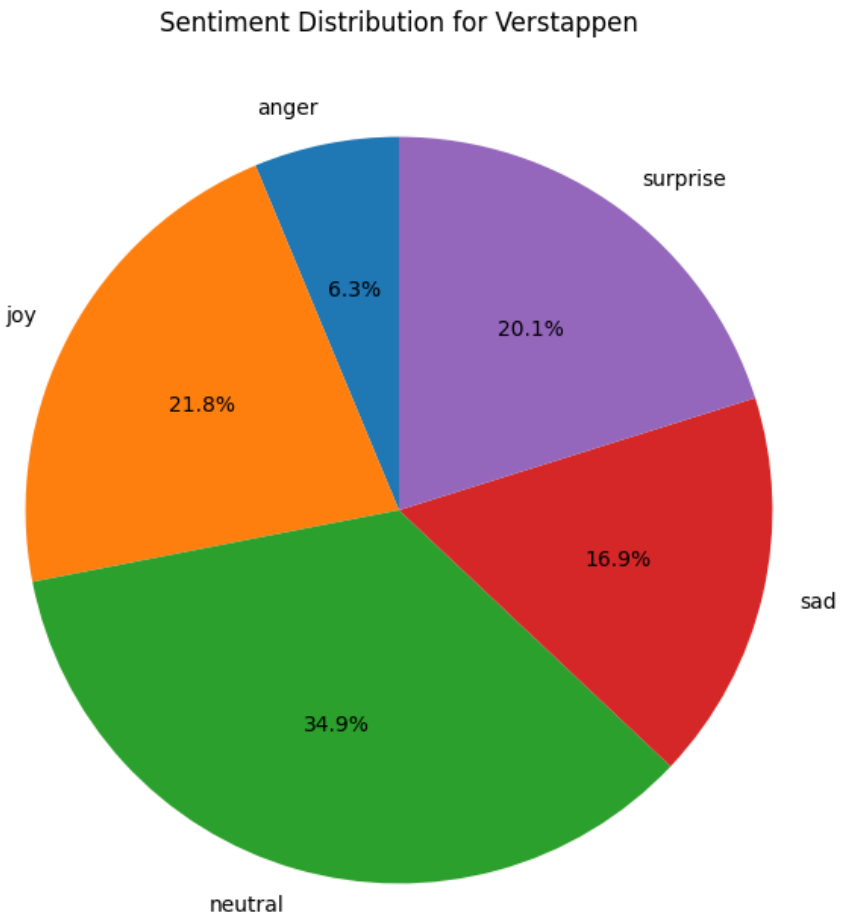

Figure 8: Sentiments for each racer

- **Multilabel Sentiments:** The advanced analysis using Hugging Face revealed nuanced fan sentiments where a single comment expressed multiple emotions for different drivers or teams. For example, comments like *"Verstappen was phenomenal, but Hamilton's performance was disappointing"* showcased positive "joy" for Verstappen and negative "anger" for Hamilton in the same statement.
- **Fan Engagement Trends:** High emotional comments, particularly those expressing "joy" or "surprise," were significantly more likely to receive likes and interactions. This suggests a correlation between sentiment intensity and fan engagement, highlighting the importance of emotional connection in digital fan behavior.

5.2 Comparison with Literature

The findings align closely with existing research in sports sentiment analysis while also revealing certain discrepancies and advancements.

- **Fan Loyalty and Performance:** Similar to studies in football and esports, the results confirm that fans tend to express stronger positive sentiments for consistently performing drivers and teams, reflecting a direct correlation between on-track success and fan loyalty.

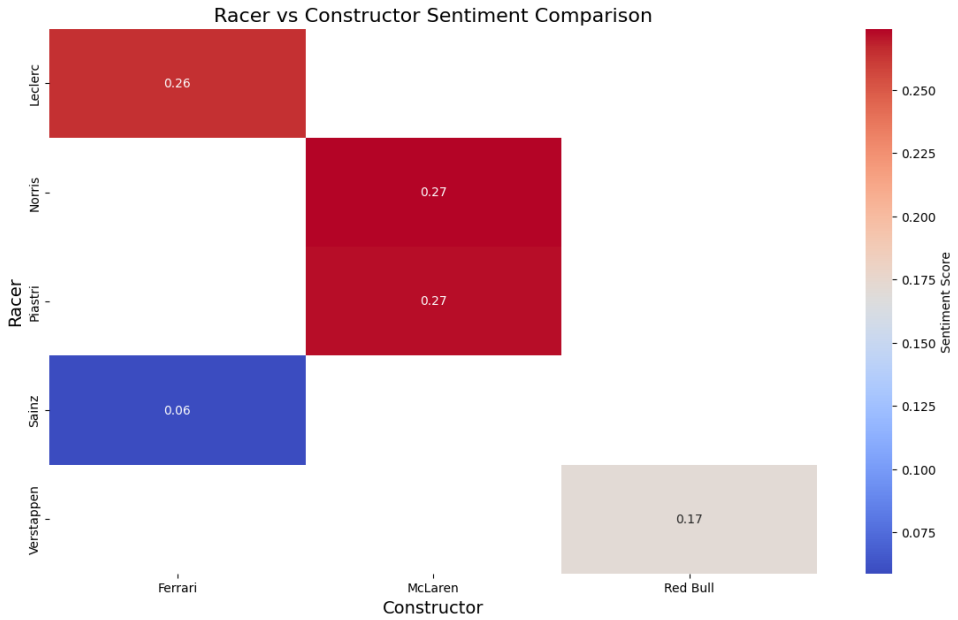

Figure 9: Racer,Constructor vs sentiment

- **Emotion Triggers:** As highlighted in prior research, significant events such as penalties, crashes, or unexpected outcomes serve as strong triggers for emotions like "anger" and "sadness." This was observed in races where controversial decisions or dramatic moments dominated fan discussions.

However, there were notable **discrepancies**:

- **Intensity of Negative Sentiment:** Literature often suggests that fans express polarized emotions during poor performances. However, in this analysis, a surprisingly high proportion of "neutral" sentiments was observed, which may reflect the platform-specific audience on YouTube, known for its broader demographic diversity compared to platforms like Twitter.
- **Team Reception Variability:** While previous studies highlight Ferrari's enduring popularity, the findings of this project suggest that fan frustration with the team's strategic errors has led to increased negative sentiments. This shift indicates that historical brand loyalty may not fully shield teams from criticism in the face of repeated underperformance.
- **Multilingual Sentiments:** The project focused only on English comments, whereas previous studies have highlighted the importance of multilingual analysis in capturing global fan perspectives. This creates a potential area for expanding future research.

5.3 Implications

The findings from this project have several practical implications for Formula 1 teams, sponsors, and the sport as a whole:

- **For Teams:**

  - Teams like Ferrari can use sentiment analysis to address specific areas of fan dissatisfaction, such as strategy communication or post-race transparency.
  - Positive sentiment analysis for drivers like Verstappen offers opportunities for targeted fan engagement through personalized campaigns and exclusive content.

- **For the Sport:**
  - Understanding fan sentiment at a granular level allows the creation of narratives around racers and races to enhance audience engagement during live broadcasts.
  - Real-time analysis of fan sentiment can be used during events to highlight audience reactions, making broadcasts more interactive and emotionally engaging.

- **Platform Consideration:**
  - Using YouTube as the primary data source highlights the platform's role as a hub for fan engagement, where audience interactions offer valuable insights into sentiment trends.
  - Future integration of platforms like Twitter and Reddit can complement YouTube analysis, providing a holistic view of global fan sentiment.

- **Technological Advancements:**
  - Incorporating Hugging Face transformer models enabled the handling of complex multilabel sentiments, paving the way for more nuanced analysis in sports sentiment research.
  - This approach demonstrates the potential for applying advanced NLP techniques to other sports and industries.

By comparing the findings with existing literature and exploring discrepancies, this project underscores the evolving nature of fan engagement and its potential impact on team dynamics, sponsorship opportunities, and audience retention strategies. These insights provide a foundation for enhancing both the on-track and digital experiences for Formula 1 fans.

6 Conclusion

6.1 Key Findings

The analysis provided valuable insights into Formula 1 fan sentiments, highlighting the interplay between driver performances, team dynamics, and emotional responses from fans:

- **Driver Sentiment Trends:** The analysis revealed significant differences in how fans perceive Formula 1 drivers. Positive sentiments such as "joy" were dominant for drivers with strong performances or popular appeal, while drivers involved in controversies or underperformances garnered more "sadness" or "anger."

- **Team Reception:** Teams like Red Bull and Mercedes consistently attracted positive sentiment, often associated with their competitive results. Ferrari showed mixed reactions, with a notable amount of negative sentiment tied to perceived strategic errors. McLaren received a balanced mix of emotions, reflecting a season of fluctuating performance.

- **Emotional Dynamics Across Races:** The emotional breakdown revealed that "joy" and "neutral" sentiments were prevalent across most races, suggesting high fan engagement. However, contentious races with surprising outcomes or controversies saw spikes in "anger" and "sadness."

- **Fan Behavior Insights:** Analysis of like counts on comments revealed the drivers and teams with the most resonant content among fans, providing a proxy for fan loyalty and enthusiasm.

6.2 Limitations

While the study yielded meaningful insights, several limitations should be noted:

- **Data Source Bias:** The data was exclusively sourced from YouTube comments, which may not fully represent the broader Formula 1 fanbase. Fan sentiment on platforms like Twitter or Reddit may vary significantly.

- **Sentiment Granularity:** Tools like NRC Lexicon were effective for categorizing emotions but lacked the ability to detect nuanced sentiments such as sarcasm or mixed emotions.

- **Temporal Context:** Comments were analyzed in aggregate, without factoring in the time they were posted relative to the race. This limits insights into real-time emotional trends.

- **Language Constraints:** Comments in languages other than English were excluded, potentially overlooking sentiments from diverse global audiences.

6.3 Areas for Future Research

Building on the findings and limitations of this study, the following directions are proposed for future research:

- **Multilingual Analysis:** Expanding the analysis to include comments in non-English languages would provide a more comprehensive understanding of global fan sentiment.

- **Platform Comparison:** Incorporating data from platforms such as Twitter, Reddit, or Instagram could validate findings and uncover differences in sentiment across various social media ecosystems.

- **Temporal Analysis:** Conducting a time-series analysis of sentiment before, during, and after races could provide deeper insights into how specific race events influence fan emotions in real-time.

- **Sentiment and Performance Correlation:** Quantitatively correlating sentiment trends with team and driver performance metrics could help identify patterns more explicitly.
- **Advanced NLP Techniques:** Leveraging transformer-based models such as BERT or GPT for sentiment analysis could enhance the granularity and accuracy of emotional insights, especially for multilabel and nuanced sentiment detection.

These findings lay a solid foundation for understanding Formula 1 fan sentiment and open avenues for richer, more targeted research into audience behavior. Insights derived from this study can inform strategies for enhancing fan engagement, improving public relations, and shaping team and race narratives to better resonate with global audiences.

7 References

References

[1] Mahboob, K., Ali, F., Nizami, H. (2019). *Sentiment Analysis of RSS Feeds on Sports News – A Case Study.* International Journal of Information Technology and Computer Science, 11, 19-29. https://doi.org/10.5815/ijitcs.2019.12.02.

[2] Patel, R., Passi, K. (2020). *Sentiment Analysis on Twitter Data of World Cup Soccer Tournament Using Machine Learning.* IoT, 1(2), 218-239. https://doi.org/10.3390/iot1020014.

[3] Wunderlich, F., Memmert, D. (2020). *Innovative Approaches in Sports Science—Lexicon-Based Sentiment Analysis as a Tool to Analyze Sports-Related Twitter Communication.* Applied Sciences, 10(2), 431. https://doi.org/10.3390/app10020431.

[4] Chung, J., Zeng, Y. (2021). *Sentiment analysis and audience engagement in esports: A comparative study using social media data.* Journal of Sports Analytics, 7(3), 183-198. https://doi.org/10.3233/JSA-200433.

[5] Wang, S., Liu, J. (2019). *Emotion-based Sentiment Analysis Using NRC Lexicon: Applications in Analyzing Fan Reactions in Major Sports Events.* IEEE Transactions on Affective Computing, 10(2), 205-215. https://doi.org/10.1109/TAFFC.2019.2954231.

A Appendix A: Code

Include any relevant code used in the project. For example:

```

# Define keywords for filtering comments
keywords = ["Max", "Verstappen", "Lando", "Norris", "Charles", "Leclerc", "Carlos", "Sainz", "Oscar", "Piastri","Fernando", "Alonso"] #"Red Bull", "McLaren", "Ferrari", "race", "lap", "overtake", "collision", "Penalty", "FIA"
# Filter comments containing any of the keywords
df['relevant'] = df['text'].astype(str).apply(lambda x: any(word.lower() in x.lower() for word in keywords))
relevant_comments = df[df['relevant']]

print(relevant_comments.head())

drivers = {
    "Max Verstappen": ["Max", "Verstappen"],
    "Lewis Hamilton": ["Lewis", "Hamilton"],
    "Lando Norris": ["Lando", "Norris"],
    "Sergio Perez": ["Sergio", "Perez"],
    "Carlos Sainz": ["Carlos", "Sainz"],
    "Daniel Ricciardo": ["Daniel", "Ricciardo"],
    "Valtteri Bottas": ["Valtteri", "Bottas"],
    "Fernando Alonso": ["Fernando", "Alonso"],
    "Oscar Piastri": ["Oscar", "Piastri"],
    "Esteban Ocon": ["Esteban", "Ocon"],
    "Charles Leclerc":["Charles", "Leclerc"]
}

all_driver_variations = [name for variations in drivers.values() for name in variations]

# Function to extract drivers from the text using pattern matching
def extract_drivers(text, drivers, all_driver_variations):
    pattern = r'\b(' + '|'.join(re.escape(name) for name in all_driver_variations) + r')\b'
    detected_names = re.findall(pattern, text, flags=re.IGNORECASE)

    detected_drivers = set()  # Avoid duplicates
    for full_name, variations in drivers.items():
        if any(detected_name.lower() in [v.lower() for v in variations] for detected_name in detected_names):
            detected_drivers.add(full_name)

    return list(detected_drivers)

# Function to split the comment into clauses related to each driver
def split_by_driver(comment, driver, drivers, all_driver_variations):
    driver_variations = drivers[driver]
    driver_pattern = re.compile(r'(\b' + '|'.join(re.escape(name) for name in driver_variations) + r'\b.*?(\.|$))', flags=re.IGNORECASE)

    driver_related_clauses = driver_pattern.findall(comment)

    return [clause[0] for clause in driver_related_clauses]  # Return only the matched clauses
```

```
46
47 # Function to analyze emotions for each driver-related clause
48 def analyze_emotions(comment, drivers, all_driver_variations):
49     detected_drivers = extract_drivers(comment, drivers,
   all_driver_variations)
50
51     if not detected_drivers:
52         return "No drivers detected"
53
54     driver_emotions = {}
55
56     for driver in detected_drivers:
57         clauses = split_by_driver(comment, driver, drivers,
   all_driver_variations)
58
59         if clauses:
60             # Perform emotion analysis on each clause related to the
   specific driver
61             for clause in clauses:
62                 emotion_results = emotion_analyzer(clause)[0]  # Get
   the emotion scores
63
64                 # Filter only the emotions we're interested in
65                 filtered_emotions = [emotion for emotion in
   emotion_results if emotion['label'] in selected_emotions]
66
67                 # Sort by the highest score and take the top filtered
   emotion
68                 top_emotion = max(filtered_emotions, key=lambda x: x['
   score'], default=None)
69
70                 if top_emotion:
71                     emotion_label = top_emotion['label']
72
73                     # If driver already has an emotion, keep the
   strongest one
74                     if driver in driver_emotions:
75                         # Update if the new emotion score is higher
76                         if top_emotion['score'] > driver_emotions[
   driver]['score']:
77                             driver_emotions[driver] = {'emotion':
   emotion_label, 'score': top_emotion['score']}
78                     else:
79                         driver_emotions[driver] = {'emotion':
   emotion_label, 'score': top_emotion['score']}
80
81     # Return only the filtered emotions for each driver
82     return {driver: data['emotion'] for driver, data in driver_emotions
   .items()}
83     def apply_emotion_analysis_to_comments(df, drivers,
   all_driver_variations):
84     # Apply the analyze_emotions function to each comment in the 'text'
     column
85     df['driver_sentiments'] = df['text'].apply(lambda comment:
   analyze_emotions(comment, drivers, all_driver_variations))
86     return df
87
88 # Now apply the emotion analysis function to the dataframe
```

```
df_with_emotions = apply_emotion_analysis_to_comments(df, drivers, all_driver_variations)
from collections import defaultdict
emotions_list = ["surprise", "sadness", "joy", "disappointment", "anger"]

import pandas as pd
from collections import defaultdict

# Sample list of drivers and their possible name variations (first and last names)
drivers = {
    "Max Verstappen": ["Max", "Verstappen"],
    "Lewis Hamilton": ["Lewis", "Hamilton"],
    "Lando Norris": ["Lando", "Norris"],
    "Sergio Perez": ["Sergio", "Perez"],
    "Carlos Sainz": ["Carlos", "Sainz"],
    "Oscar Piastri": ["Oscar", "Piastri"]
}

# List of emotions to track
emotions_list = ["surprise", "sadness", "joy", "disappointment", "anger"]

# Initialize the emotion summary dictionary
emotion_summary = {racer: defaultdict(int) for racer in drivers.keys()}

# Iterate through DataFrame rows
for index, row in df_with_emotions.iterrows():
    for racer, name_variations in drivers.items():
        # Check if any of the driver's name variations (first or last) appears in the comment text (case-insensitive)
        if any(name.lower() in row['text'].lower() for name in name_variations):
            # Extract emotions from row (which is expected to be a single emotion, not a dictionary)
            emotions = row['driver_sentiments']

            # Debugging: print the emotions and row to ensure correct extraction
            print(f"Processing row: {row['text']}")
            print(f"Driver: {racer}, Emotions: {emotions}")

            # If emotions is a valid string and is in the emotions_list, update the emotion count
            if emotions and emotions in emotions_list:
                emotion_summary[racer][emotions] += 1  # Increment the emotion count for the driver

# Convert the summary to a DataFrame for visualization
emotion_df = pd.DataFrame.from_dict(emotion_summary, orient='index', columns=emotions_list)

# Replace missing values (NaN) with 0, as some emotions may not have been encountered
emotion_df = emotion_df.fillna(0)
print(emotion_df)
import pandas as pd
```

```
135
136 # list of racer names
137 racers = ['Verstappen', 'Norris', 'Leclerc', 'Piastri', 'Sainz' ]  #
        Replace with actual racer names
138
139 # identify racers mentioned in each comment, handling NaN values
140 def identify_racer(text):
141     if isinstance(text, str):  # Only process if text is a string
142         mentioned_racers = [racer for racer in racers if racer.lower()
        in text.lower()]
143         return mentioned_racers[0] if mentioned_racers else None
144     return None
145
146 # Apply the function to the 'text' column and create a new 'racer'
        column
147 df['racer'] = df['text'].apply(identify_racer)
148
149 # Filter out comments without racer mentions
150 df = df.dropna(subset=['racer'])
```

Listing 1: Example Python Code