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

November 24, 2024

Contents

1 Introduction

In the dynamic world of motorsport, fan engagement and sentiment analysis play a vital role in understanding audience perception and preferences. This project explores the wealth of insights hidden within fan comments on online platforms, focusing on the 2024 Formula One season's first 20 races. By analyzing 100,000 scraped comments, this project aims to uncover trends, sentiments, and public opinions related to drivers, teams, and the races themselves.

The objective 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

a) 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.
- **Gaps Identified:** Limited focus on sport-specific terminologies and deeper sentiment dimensions such as surprise or fear.

b) 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.
- **Gaps Identified:** Many studies are limited to quantitative metrics such as the volume of comments without delving into the qualitative sentiment nuances.

3 Methodology

This project employs a comprehensive approach to analyze fan sentiment surrounding Formula 1 races in the 2024 season. Data was collected by scraping over 100,000 comments from YouTube videos related to 20 races. Preprocessing techniques were applied to clean and filter the data, ensuring relevance to racers and teams. Sentiment analysis was conducted using the NRC Lexicon, categorizing emotions into five distinct classes: joy, sadness, anger, surprise, and neutral. The analysis focused on identifying trends in fan perceptions, comparing team and driver reception, and uncovering insights into audience engagement. Results were visualized using Python libraries and Tableau to deliver an intuitive understanding of the findings.

3.1 Data Collection

- **Source:** Data was collected from YouTube comments on videos related to Formula 1 races in the 2024 season.
- **Volume:** Over 100,000 comments were scraped from 20 videos, covering insights from fans across all major teams and drivers.
- **Tools:** The scraping process was performed using Python libraries like BeautifulSoup and pandas to extract and structure the data.

3.2 Data Preprocessing

- **Cleaning:** The data was cleaned to remove noise such as Stop words (”the”, ”and”, ”but”) using nltk. Non-alphanumeric characters and special symbols using regular expressions.
- **Filtering:** Relevant comments were identified by searching for mentions of racers and teams using a predefined list of keywords.
- **Handling Missing Data:** Rows with empty or irrelevant fields were excluded, and a small percentage of incomplete comments were filled manually for accuracy.
- **Normalization:** Text was converted to lowercase for uniformity and to avoid duplication during analysis.

3.3 Analysis Techniques

- **Tools:** The NRC Lexicon was used to categorize emotions into five distinct categories: Joy, Sadness, Anger, Surprise, Neutral
- **Process:** Comments were tokenized to analyze individual words. Words were mapped to their respective emotional scores using NRC Lexicon. Each comment was classified based on the predominant emotion.

4 Results

Present the results of the analysis. Use tables, figures, and charts to support the findings.

5 Discussion

5.1 Results Interpretation

The analysis of YouTube comments revealed several key insights into fan sentiment regarding Formula 1 racers and teams:

- **Driver Perception:** Among the racers, Max Verstappen received overwhelmingly positive sentiments, with ”joy” being the predominant emotion in their mentions. Conversely, Carlos Sainz faced a higher proportion of negative sentiments, predominantly ”anger” and ”sadness,” often tied to race incidents.
- **Team Sentiment Trends:** Red Bull, McLaren, and Mercedes emerged as the teams with the most positive mentions, reflecting strong fan appreciation for their performance. On the other hand, Ferrari showed a mix of ”sadness” and ”anger,” often linked to discussions of strategic missteps.
- **Emotional Distribution:** Across all races, ”joy” was the dominant emotion, followed by ”neutral” and ”surprise,” indicating fans’ overall positive engagement with the season. However, specific races, such as Race A, showed spikes in ”anger” and ”sadness,” likely tied to controversies or crashes.

5.2 Comparison with Literature

The results align with existing research in sports sentiment analysis:

- **Fan Loyalty and Performance:** Similar to findings by stronger positive sentiments for consistently performing drivers and teams, reflecting a correlation between performance and fan perception.
- **Emotion Triggers:** Studies like highlight that fanuring controversial events or race incidents. This was confirmed in our data, where specific incidents were tied to heightened ”anger” and ”sadness.”

However, some **discrepancies** were observed:

- **Intensity of Negative Sentiment:** Literature often suggests that fans tend to express polarized emotions during poor performances. In our analysis, ”neutral” sentiment was higher than expected, which might reflect the platform (YouTube) and its diverse audience base.
- **Team Reception Variability:** While prior studies suggested Ferrari’s enduring popularity offsets negative sentiments, our findings showed a strong association between fan frustration and strategic decisions, suggesting a shift in perceptions.

5.3 Implications

- **For Teams:** Insights into team-specific sentiment can inform public relations strategies. Teams like Ferrari may need to address fan frustrations through transparent communication or enhanced strategies.
- **For the sport:** Understanding fan sentiment at a granular level helps in creating narratives around racers and races to enhance engagement.
- **Platform Consideration:** The use of YouTube as a data source highlights how online platforms shape and reflect audience sentiment, offering valuable insights for digital engagement strategies.

By comparing our findings with literature and exploring discrepancies, this project emphasizes the evolving nature of fan engagement and its potential for influencing team dynamics and sport narratives.

6 Conclusion

6.1 Key Findings

- **Driver Sentiment Trends:** The analysis highlighted 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, races with contentious incidents or surprises 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, offering a proxy for fan loyalty and enthusiasm.

6.2 Limitations

- **Data Source Bias:** The data was exclusively sourced from YouTube comments, which may not fully represent the broader Formula 1 fanbase. Fan sentiment on other platforms like Twitter or Reddit may vary significantly.
- **Sentiment Granularity:** While NRC Lexicon effectively categorized emotions, some nuanced sentiments or sarcasm in comments may have been misinterpreted.
- **Temporal Context:** Comments were analyzed in aggregate, without factoring in the time they were posted relative to the race, which might influence sentiment trends.
- **Language Constraints:** Comments in languages other than English were excluded, potentially missing diverse perspectives.

6.3 Areas for Future Research

- **Multilingual Analysis:** Expanding the analysis to include non-English comments would provide a more comprehensive understanding of global fan sentiment.
- **Platform Comparison:** Including data from other platforms like Twitter, Reddit, or Instagram could validate findings and highlight differences in sentiment across fan communities.
- **Temporal Analysis:** A time-series analysis of sentiment before, during, and after races could provide deeper insights into the impact of race events on fan emotions.
- **Sentiment and Performance Correlation:** Future studies could quantitatively correlate sentiment trends with team and driver performance metrics to identify patterns more explicitly.
- **Advanced NLP Techniques:** Leveraging transformer-based models like BERT or GPT for sentiment analysis could improve the accuracy and granularity of emotional insights.

These findings lay the foundation for understanding fan sentiment in Formula 1 and open avenues for richer, more targeted research on audience behavior and its implications for the sport.

7 References

A Appendix A: Code

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

```python
import pandas as pd
# Load data
df = pd.read_csv('data.csv')
# Preprocess data
df = df.dropna()
```

Listing 1: Example Python Code

B Appendix B: Additional Figures

Include any additional figures or tables that support the analysis.