F1 DRIVERS' CHAMPIONSHIP ANALYSIS

Elena Maggiore, Giovanni Noé

Abstract

This project analyses discussions within the Formula 1 community on Reddit during the 2025 season, with the aim of understanding how fans reacted to the Drivers' Championship outcome and how opinions and interactions are structured across different fan groups. Posts and comments are collected from both general Formula 1 subreddits and team-specific subreddits, and analysed through social network analysis, community detection, and sentiment analysis techniques. After constructing a user interaction graph, several community detection algorithms are applied and compared, using subreddit affiliation as a reference point rather than as a strict objective. Finally, sentiment analysis based on AFINN, VADER, and BERT is performed to investigate how different communities emotionally responded to the evolution and final outcome of the championship. The results highlight the presence of latent communities that partially reflect team allegiances but also go beyond subreddit boundaries, as well as clear differences in sentiment trends linked to key events of the 2025 Formula 1 season.

Introduction

This project aims to analyse the Formula 1 community on the platform Reddit to discover its users' opinion on driver Lando Norris being crowned World Champion.

Both the general Formula 1 subreddit and subreddits of the most important racing teams are be explored, trying to highlight eventual similarities or differences. A social network analysis is conducted through centrality evaluation of nodes and consequential community detection using the greedy modularity-based algorithm, the Louvain algorithm, and the fluid algorithm, while Afinn, VADER and BERT complete the social sentiment analysis.

The expectation is to see the general community split into smaller ones, which should resemble the behaviour of specific racing teams' subreddits.

Data Acquisition

Reddit comments and posts are acquired through the official Reddit API [1]. The following six subreddits are selected for the analysis:

- *Formula 1*, 1.1 million weekly visitors, where posts are ranked by relevance and include news, highlights, and humorous content;
- *Formula 1 Discussions*, 149647 weekly visitors, which is more focused on opinions and direct interactions;
- *McLaren*, 16839 weekly visitors;
- *RedBull*, 51799 weekly visitors;
- *Ferrari*, 45368 weekly visitors;
- *Mercedes*, 13737 weekly visitors.

The first two subreddits appeal to a broad audience, ranging from casual viewers to highly engaged fans, while the remaining four represent team-specific communities associated with the largest fan bases on Reddit.

After selecting the subreddits, a series of filters is applied to identify relevant posts:

1. **Keywords**: only posts that contain at least one of "*Lando*", "*Norris*", "*WDC*", "*Champion*" (either in the title or in the body and ignoring case) are kept.
2. **Temporal horizon**: after parsing the post creation date from timestamp to UTC, only posts published from March 7, 2025, to December 15, 2025 (one week before to one week after the Formula 1 Championship Season [2]) are kept.
3. **Comments**: only posts with at least one comment are kept.

Resulting posts are collected into dataframe *posts_df*, consisting of:

Formula1	F1Discussions	McLaren	RedBull	Ferrari	Mercedes	Total
40	276	401	117	54	41	929

Once the extraction is completed, posts' comments are retrieved through a function and imported into dataframe *comments_df*. After polishing, results are the following:

Formula1	F1Discussions	McLaren	RedBull	Ferrari	Mercedes	Total
2574	2333	5811	1460	143	97	13471

This completes the data acquisition step of the project.

Exploratory Data Analysis

Before proceeding with community detection and sentiment analysis, the collected data is explored to assess its structure and suitability.

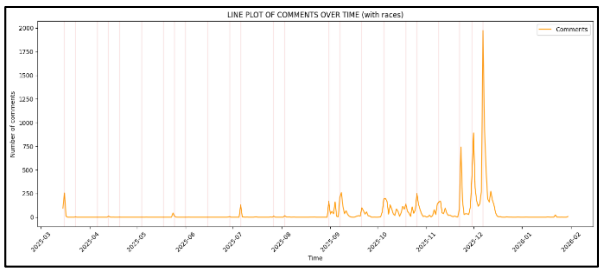

Figure 1: comments over time, considering races.

This graph shows the frequency of comments during the Formula 1 Season, highlighting race weekends [2]. Comment activity clearly increases during race weekends and rises overall during the final third of the season, with the highest peak occurring in correspondence with the final race.

By analysing authors and commenters across subreddits, it emerges that 19 users authored posts in more than one subreddit, while 670 users commented on posts belonging to multiple subreddits, indicating that the subreddits are not isolated environments and are expected to be linked when performing community detection.

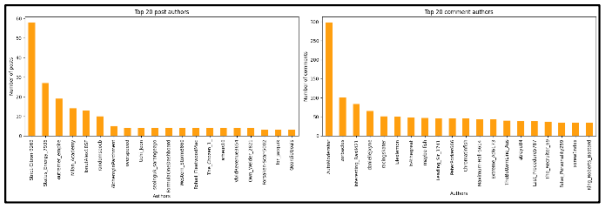

Figure 2: distribution of posts per author and distribution of comments per author.

These distributions describe user-level activity and reveal that a small fraction of users is responsible for a large share of content and interactions. Such highly active users are expected to play a central role in the interaction network by connecting different parts of the graph and different subreddits.

To further compare the subreddits, word clouds are generated from all comments combined and from each subreddit separately.

Figure 3: word clouds.

While common terms such as *Lando Norris*, *Max Verstappen*, *Oscar Piastri*, *car*, and *driver* appear across all subreddits, each community also exhibits distinctive vocabulary reflecting its specific focus and interests.

Taken together, these observations point to a strong inter-subreddit interaction component while still preserving meaningful differences between communities, justifying the subsequent network-based analysis.

Community Detection

All comments are merged into a single dataset and used to build a user interaction network. The objective of community detection is to identify cohesive and interpretable latent communities that represent meaningful patterns of interaction beyond explicit subreddit boundaries.

The analysis consists of three main steps:

1. **Graph building**, where users' relationship are modelled into a graph. It can be visualized to serve the making of assumptions on communities.
2. **Social Network Analysis measures**, where metrics related to connection, distribution and segmentation are computed for the graph, exploring its structure.

3. **Community detection algorithms**, where three different community detection algorithms are applied on the graph, comparing the results and choosing the best model.

Graph Building

A **graph** $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E})$ is made of **vertices** $\boldsymbol{V}$ (**nodes**) and **edges** $\boldsymbol{E}$. Specifically:

- The **nodes** of this graph are users.
- The **edges** of this graph are comments, connecting the author of the comment to the author of the commented post.

This way, the relationships captured by the graph are

$$A \longrightarrow B = User\ A\ commented\ on\ User\ B's\ post$$

This definition results in a directed graph, as interactions have a clear direction from commenter to post author. Prior to graph construction, users are filtered: only users who commented at least three times or received at least three comments on their posts are retained. The purpose of this filtering step is to remove very weakly connected nodes and reduce graph sparsity.

The graph is built using the *NetworkX* library [3]. The final graph consists of 882 nodes and 3,926 edges. Degree centrality and betweenness centrality are computed to determine node importance and to scale node sizes for visualisation, while edge colours reflect the subreddit of the target post.

The following results are obtained:

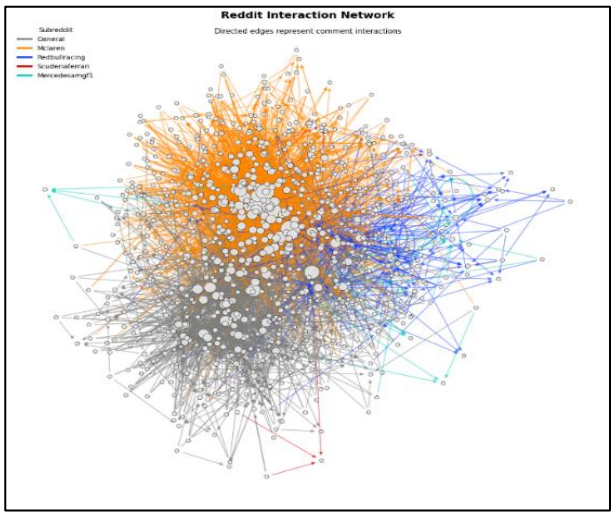

Figure 4: interaction network graph.

The network exhibits a highly heterogeneous topology characterized by a dense core and a weakly connected periphery. The central region contains a large concentration of nodes with high edge density, suggesting the presence of highly active users who engage frequently in discussions and receive a substantial number of replies. Three clusters (McLaren, RedBull, general) are visibly concentrated, suggesting that users predominantly interact with the same subreddit. However, the substantial overlap indicates significant cross-subreddit engagement, implying that certain users act as connectors between communities.

After visualising the graph, more social network analysis measures can be calculated.

Social Network Analysis Measures

Social network analysis metrics are computed to characterise the structure of the network. These metrics can be grouped into:

- **Connection metrics** measure how entities connect with each other; an example is the clustering coefficient.
- **Distribution metrics** analyse how information flows through a network; in this case, centrality metrics and density will be used.
- **Segmentation metrics** measure how much a network is clustered; again, a good segmentation metric is the clustering coefficient.

First, an undirected version of the graph must be built to calculate:

- **Number of connected components**, equal to 1.
- **Node connectivity**, equal to 3.
- **Edge connectivity**, equal to 3.

In particular, these values of node and edge connectivity are expected due to the filtering applied on the dataframe.

The 20 nodes with the highest in-degree are then analysed in more detail by comparing normalised in-degree, betweenness, and closeness centrality.

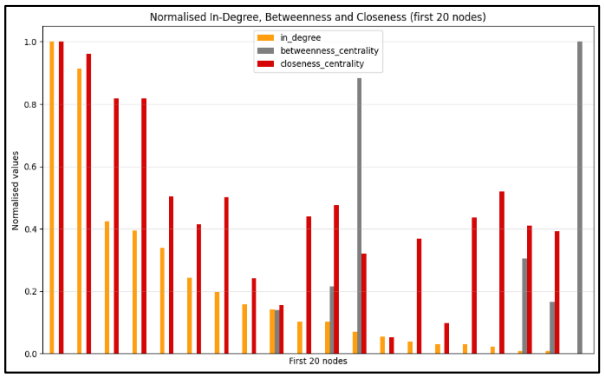

Figure 5: normalised In-Degree, Betweenness and Closeness (first 20 nodes)

In-degree and closeness show similar decreasing trends, while betweenness does not follow the same pattern, suggesting that highly visible users are not necessarily the main intermediaries between different parts of the network.

Additional global metrics are computed:

- **Average clustering coefficient**: 0.028;
- **Global clustering coefficient**: 0.023;
- **Density**: 0.005;
- **Assortativity**: −0.108.

The low clustering coefficients and density indicate a sparse network with limited local clustering, which is typical of large online discussion networks. The negative assortativity suggests that highly connected users tend to interact with less connected users.

Community Detection Algorithms

Community detection is performed using three algorithms: the greedy modularity-based approach [4], the Louvain algorithm [5], and the fluid algorithm [6]. Subreddit affiliation is used as a reference point to evaluate the interpretability of the detected communities, rather than as a strict objective. The expectation is that meaningful latent communities may partially reflect team allegiances while also capturing more complex interaction patterns.

All algorithms are constrained to return five communities, corresponding to the main fan bases, in order to enable comparison.

Greedy modularity-based approach

The first algorithm considered is the greedy modularity-based approach [4], which iteratively merges communities so as to maximise modularity at each step until no further improvement is possible. The following table describes the resulting communities and the percentage of users from different subreddits for each of them:

Community	Size	McLaren	Mercedes	RedBull	General	Ferrari
0	258	**73.485**	1.010	5.050	19.949	0.505
1	250	19.898	2.551	9.439	**67.347**	0.765
2	159	**69.874**	2.092	5.439	22.176	0.418
3	149	29.268	6.272	**39.024**	23.693	1.742
4	66	36.283	0.885	6.195	**56.637**	0.000

The resulting community composition shows that McLaren users are mainly split between two communities, while general fans also appear divided across two groups. RedBull users are largely concentrated in a single community, whereas Mercedes users are spread across several communities but with a stronger presence in the Red Bull–dominated one. Ferrari users, due to their relatively small number, are diluted across multiple communities, preventing strong conclusions.

Louvain algorithm

The Louvain algorithm [5] is also based on modularity maximisation but is optimised for detecting non-overlapping communities efficiently.

Once again, the output of the algorithm is stored in the following table:

Community	Size	McLaren	Mercedes	RedBull	General	Ferrari
0	258	**65.526**	1.467	8.068	24.450	0.489
1	266	22.426	3.890	3.890	**62.243**	0.229
2	6	**77.778**	11.111	11.111	11.111	0.000
3	121	23.585	4.717	4.716	**34.434**	0.472
4	231	**66.111**	1.111	1.111	22.778	1.944

Results show one very small community composed of only six users, mostly associated with McLaren, while general fans are primarily concentrated in a single community. McLaren users are more fragmented across multiple communities, and RedBull users are mainly concentrated in one community, though less distinctly than with the greedy approach.

Fluid algorithm

The fluid algorithm [6] models communities as entities that can expand or contract through

interactions and requires the number of communities as a parameter.

This table contains the proportion of subreddit users in the resulting communities.

Community	Size	McLaren	Mercedes	RedBull	General	Ferrari
0	176	25.083	4.290	14.521	**55.776**	0.330
1	177	**65.371**	2.120	10.601	19.788	2.120
2	177	**63.869**	1.460	12.774	21.533	0.365
3	176	**54.909**	1.818	15.273	28.000	0.000
4	176	25.342	3.425	13.014	**57.192**	1.027

The resulting communities are well balanced in size; however, team-based fan groups are highly fragmented. McLaren and general fans are spread across several communities, and RedBull users no longer form a clearly identifiable group, making interpretation more difficult.

Results

A quantitative comparison of the algorithms is first performed using modularity, which measures how well a network is partitioned into communities by comparing the density of intra-community edges to inter-community edges.

The obtained values are:

- Greedy modularity-based approach: 0.39;
- Louvain algorithm: 0.40;
- Fluid algorithm: 0.31.

Excluding the Fluid algorithm due to its lower modularity and interpretability, the greedy modularity-based approach is selected over Louvain. Although the Louvain algorithm achieves slightly higher modularity, the greedy approach yields more coherent Red Bull and Mercedes communities and a more interpretable structure overall.

Under this solution, McLaren users are mainly divided into two communities, which can be interpreted as reflecting internal divisions within the McLaren fan base during the 2025 season, when both Lando Norris and Oscar Piastri were championship contenders [7].

Due to the fact that communities can be traced back to subreddit, from this point on they are referred to as follows:

0. McLaren fans - type 1;
1. General fans with a Mercedes and Ferrari component;
2. McLaren fans - type 2;
3. RedBull fans;
4. General fans.

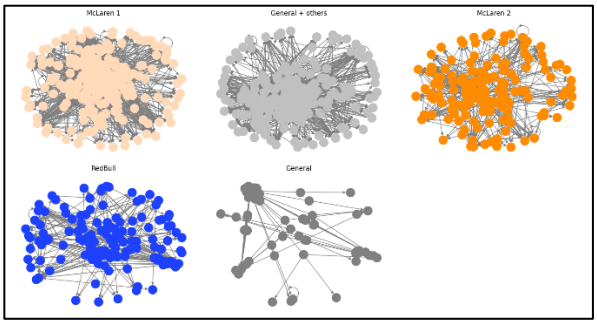

Figure 6: communities that have been found.

The persistence of subreddit influence alongside meaningful overlap is considered a positive outcome, as it suggests that community detection captures interaction patterns beyond simple subreddit membership.

Sentiment Analysis

Sentiment analysis is performed on comments to assess how different communities emotionally responded to the championship. Rather than focusing on a single aggregate value, sentiment distributions and temporal trends are analysed.

Three approaches are used:

- **AFINN** [8] is one of the simplest yet most popular lexicon-based approaches. It ranks each word with a score that ranges from -5 (extremely negative) to $+5$ (extremely positive), while 0 represents neutrality. The algorithm then sums algebraically the scores to rank an entire comment.
- **VADER** [9] is another sentiment scoring technique specific for short text such as comments or tweets. This algorithm scores short texts based on sentiment, where the score can range from -1 (extremely negative) to $+1$ (extremely positive).
- **BERT** [10] is based on a pre-trained model (the *cardiffnlp/twitter-roberta-base-sentiment* model is chosen) that assigns probabilities to negative, neutral, and

positive sentiment classes, which are aggregated into a single score.

First, the average sentiment scores for each community are plotted. For AFINN, results are the following:

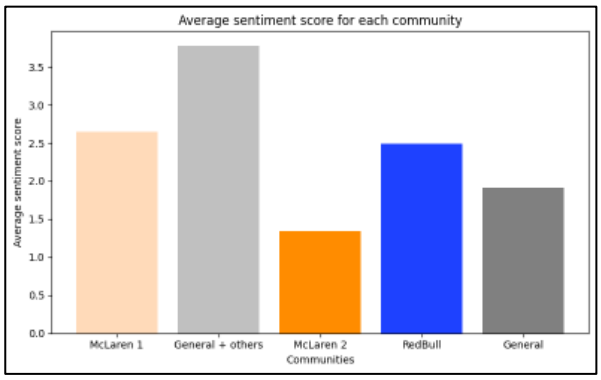

Figure 7: average sentiment score for each community (AFINN).

Results from VADER scores and BERT scores show similar trends between communities (strong difference between the McLaren communities, general and others being the most positive group), but VADER scores appear shrunk, and BERT deems the general sentiment much more negative even though relationships between communities stay the same.

The most striking result is the contrast between the two McLaren communities. This difference becomes particularly clear when focusing on the final two months of the season.

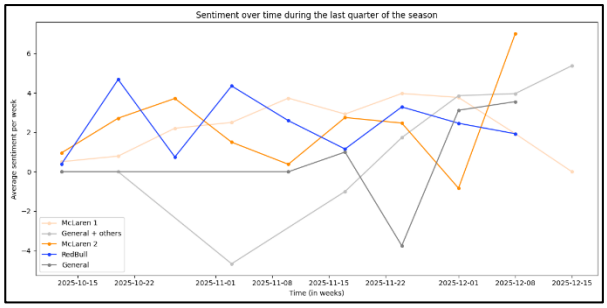

Figure 8: sentiment over time during the last quarter of the season (AFINN).

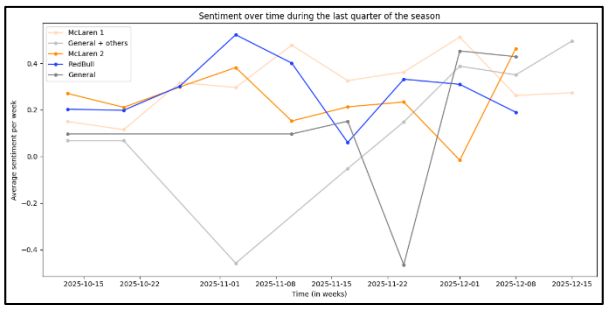

Figure 9: sentiment over time during the last quarter of the season (VADER).

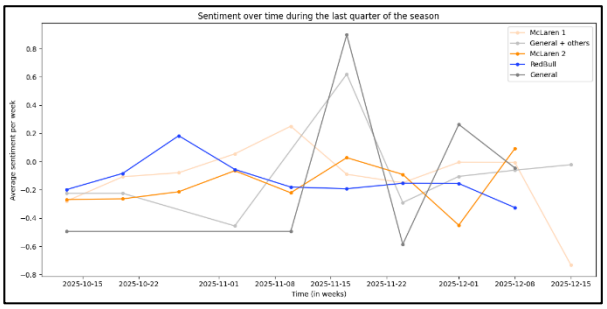

Figure 10: sentiment over time during the last quarter of the season (BERT).

One McLaren community shows increasing positivity toward the end of the season, consistent with Norris's comeback and championship win, while the other shows declining sentiment, consistent with Piastri' loss of the title after being the season favourite [7].

Red Bull fans display high sentiment during periods of strong performance and decline following the loss of the championship, while the general fan community maintains a largely positive or neutral sentiment throughout the season, reflecting engagement with a competitive and eventful championship rather than allegiance to a single outcome.

Conclusion

This project analysed the Reddit Formula 1 ecosystem to study fan interactions and sentiment during the 2025 Drivers' Championship. Social network analysis revealed strong inter-subreddit interaction alongside the presence of meaningful latent communities. Community detection highlighted internal divisions within major fan bases, particularly McLaren, reflecting the rivalry between Norris and Piastri.

Sentiment analysis showed that emotional responses closely followed the evolution of the championship, with clear differences across communities and over time. While community detection captured the structural division of fans, sentiment analysis provided insight into how these groups reacted to key events and the final outcome.

Overall, the results confirm that even in highly overlapping online environments, community detection and sentiment analysis together offer a

powerful framework for understanding both the structure of fan communities and their collective reactions to real-world events.

References

[1] Reddit, “Reddit Data API Wiki,” [Online]. Available at https://support.reddithelp.com/hc/en-us/articles/16160319875092-Reddit-Data-API-Wiki

[2] Fédération Internationale de l'Automobile, “Formula 1 2025 Season Race Calendar,” [Online]. Available at https://www.formula1.com/en/racing/2025

[3] NetworkX, “Python library NetworkX documentation,” [Online]. Available at https://networkx.org/documentation/stable/reference/index.html

[4] NetworkX, “Greedy Modularity Communities,” [Online]. Available at https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.community.modularity_max.greedy_modularity_communities.html

[5] Vincent Blonde et co., “Fast Unfolding of Communities in large networks”

[6] Ferràn Pares et co., “Fluid Communities: A Competitive, Scalable and Diverse Community Detection Algorithm”

[7] M. Wessel, “Formula 1 POINTS,” [Online]. Available at https://www.formula1points.com/season

[8] H. Lohiya, “Sentiment Analysis with AFINN lexicon,” [Online]. Available at https://himanshulohiya.medium.com/sentiment-analysis-with-afinn-lexicon-930533dfe75b

[9] C. Utto, “VaderSentiment documentation,” [Online]. Available at https://vadersentiment.readthedocs.io/en/latest/

[10] A. Rodriguez, “Sentiment Analysis with BERT: a comprehensive guide,” [Online]. Available at https://medium.com/@alexrodriguesj/sentiment-analysis-with-bert-a-comprehensive-guide-6d4d091eb6bb