

Msc in Data Science

Course: Programming in Data Science

Title: Twitter sentiment analysis for three F1 drivers almost one week before the launching of season 2023

Professor: Dr. Berberidis

Students: Georgios Diamantis, Christos Galanis

**Introduction**

Formula1 is a highly competitive sport that needs a combination of technical expertise, strategic and analytical thinking and obviously driving skill. Also, it is a highly popular sport, with millions of supporters around the globe tuning in to watch the races live or on television. Furthermore, the sport is known for its high speeds, advanced technology, and high competition, making it one of the most exciting and challenging forms of motorsport. That is why, we selected to perform sentiment analysis in this sport.

In the current paper, we selected three F1 drivers specifically Max Verstappen, Lewis Hamilton, and Charles Leclerc for extracting the sentiment of the public as the new season 2023 is about to start.

This way, we can achieve the following:

* **Fan Involvement:** there are millions of passionate fans that watch F1 races, and they are highly engaged in social media, and express often strong opinions about drivers. Sentiment analysis can help teams and sponsors comprehend how fans feel about specific drivers, which can help them adjust their sponsorship and marketing strategies accordingly.
* **Reputation** **Metrics:** F1 races are highly competitive, and this can lead drivers often under intense scrutiny from fans, media, and competitors. Extracting the sentiment of the public, drivers and teams can manage their reputations. This can help them identify potential issues early and take steps to address them before they become more significant problems.
* **Driver** **Performance:** In addition to reputation metrics, sentiment analysis can also provide insights into driver performance. By analyzing sentiment of people about drivers before and after races, teams can gain better insights and understand of how drivers are perceived by fans and the media, which can lead them to make informed decisions about driver selection, contract negotiations, and other team-related issues.

All in all, sentiment analysis can be an important tool for teams, drivers, and sponsors in F1 to gain insights into fan involvement, reputation metrics and driver performance. Additionally, it can support them to make data-driven decisions and stay ahead of the competition in this highly competitive sport.

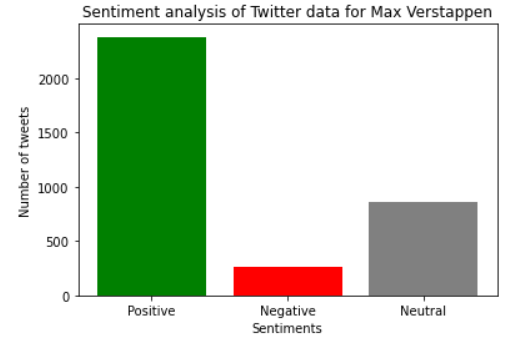
**Methodology**

Below, the outline for the methodology section is described:

* Objectives: Extract the sentiment of public to get a sense of how they feel about the drivers and their team. The drivers selected are Max Verstappen of Oracle Red Bull Racing Team, Lewis Hamilton of Mercedes-AMG Petronas Formula One Team, and Charles Leclerc of Scuderia Ferrari Team. Specifically, we used the twitter social media platform to get the sense of public during a period e.g., roughly a week prior to the upcoming F1 season. Also, a comparison of F1 drivers is being made.
* Data Collection: the whole analysis was done in python programming language. We used Twitter’s API to connect to twitter and extract tweets for each F1 driver. To accomplish this, we utilize the tweepy python library. For each F1 driver 3500 tweets were extracted including the full text of each tweet and the language was mandatory to be in English.
* Data Preprocessing: Once the tweets were collected, we removed their text URLs and Mentions. Tweets that only contained URLs and Mentions were not included in the analysis.
* Data Storage: As data repository MySQL database was used. With python we created all the necessary commands to run in MySQL. To begin with, we created a new database named twitter\_sentiment\_analysis, then 3 tables were created tweets\_max, tweets\_lewis and tweets\_lec. Each table includes the following information: id, username, followers, tweets\_plus\_retweets, text, date, location, hashtags, and sentiment. For manipulating the data inside the database MySQL Workbench was utilized.
* Sentiment Analysis: Textblob which returns each tweet polarity and subjectivity was used to categorize each tweet’s sentiment. There is a library called TextBlob in python that does this exactly procedure. And we store this information on the corresponding attribute in each table respectively.
* Results: Using pandas, matplotlib, and seaborn libraries from python we were able to make plots to visualize the findings. More specifically, a count plot in order to get the distribution of positive, negative and neutral tweets, a scatter plot to view how the tweets are placed in terms of polarity and subjective in a 2-d plot. Then, we created a line graph to see how positive and negative tweets behave over roughly a week. After that, we weighted each tweet based on its account’s followers and tweets plus retweets numbers. Finally, we make some comparisons between the F1 drivers sentiment results.

**Results – Findings**

First, we created a sentiment distribution using bar plots for each F1 driver as shown below:

**** **Chart, bar chart

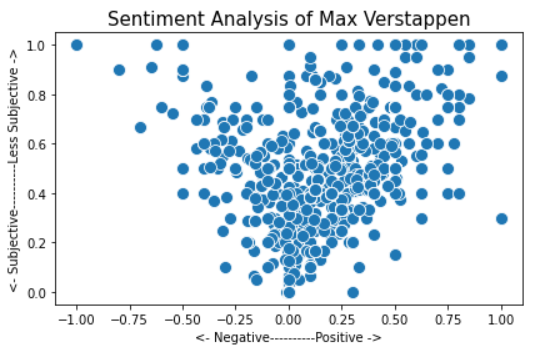
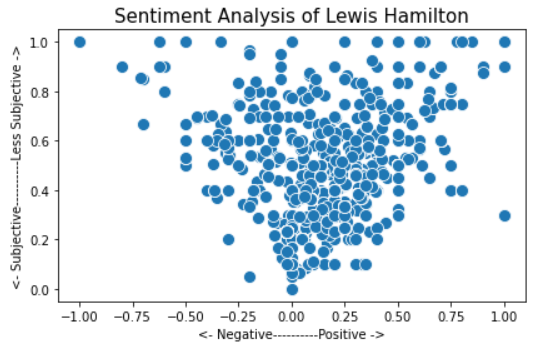
Description automatically generated**

**Chart, bar chart

Description automatically generated**

For the first diagram which corresponds to Max Verstappen sentiment analysis we can observe that the positive tweets are almost 2500 out of 3500 total tweets. However, it is obvious that, for Lewis Hamilton and Charles Leclerc neutral tweets are the majority of their dataset.

All three drivers have approximately 250 negative tweets.

Next, we created scatter plots to visualize each tweet polarity and subjectivity.

Chart, scatter chart

Description automatically generated

As we can examine from a general view, all plots seem to be almost similar. In other words, the majority of tweets for all drivers are located in the center and slightly right of the diagram which means that the tweets are roughly positive and neither subjective or objective. Also, we can see there are no tweets placed on the negative and subjective side of the plot.

After that, the sentiment of each driver is plotted over a period of around a week.

Chart, line chart

Description automatically generated Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

First of all, the common result of all three plots is that the positives tweets are more than the negatives. There are some spikes during the week. Specifically, for Max Verstappen on 17.02.2034 there were approximate 550 positive tweets from the 3500 total. Lewis Hamilton positive tweets started from 300 on 16.02. and decreased a lot especially during the days between 22.02. and 24.02. This can be an indication that the public showed its disappointment on him since he did not perform well on the Pre-Season Testing during 23-25 February. For Charles Leclerc we observe the minimum point of number of positive tweets to be on 22.02..

Finally, taking into consideration the account’s followers, tweets, and retweets we can weigh each tweet and plot it again during the specific period.

Chart, line chart

Description automatically generated Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

For all drivers, the negative tweets are not being created by accounts with high influence. On the other hand, we can see that for Max Verstappen there are some days that the tweets being made are from accounts with high influence. Whereas, for Lewis Hamilton only on 21.02. had positive tweets made by accounts with high influence and this idea is noted for Charles Leclerc during 17.02. of February.

**Final Overview**

From the whole analysis, Max Verstappen dominates in comparison to Lewis Hamilton and Charles Leclerc since it has more positive tweets, the positive tweets do not fluctuate a lot during the specified period and the tweets are posted from accounts with higher influential score. These findings reflect the performance of Max Verstappen during the last season which became F1 Champion.

**Limitations**

It is important to be noticed that since Twitter switched its API policy restricting functionalities, non-standard libraries such as tweepy are becoming more unstable. This results in limiting users of extracting large number of tweets and the search period capabilities. To overcome this issue, our team performed multiple extractions of 500 tweets for each day during the specified period.