

Examining the Impact of Max Verstappen's First World Championship on Dutch Twitter

Joost Oving - S5251370

Abstract

This scientific research paper investigates the frequency of the words "formule 1" and "f1" on Twitter around the time Max Verstappen became world champion of formula 1 for the first time. We hypothesized that the frequency of mentions of "formule 1" and "f1" would be higher around the time of Verstappen's championship. We collected and analyzed data from a one-month period between 1 December and 31 December 2021. Max Verstappen became world champion for the first time on 12 December. The results indicate a significant difference in frequency, with a notable increase leading up to and on the day of Verstappen's victory. However, the study has limitations, including the lack of pre-existing trends, and the limited coverage of Twitter data. Future research could consider larger datasets and a pre-existing trend analysis. Nonetheless, the study provides insights into the power of significant events and icons in generating national pride, and increasing engagement on social media platforms.

1 Introduction

In recent years, the Formula 1 racing scene has witnessed a remarkable rise in popularity, captivating audiences worldwide. This rise could be caused by the rise of a new star within the sport: Max Verstappen. With his extraordinary driving and his ability to not back down for anyone Max Verstappen not only captured the hearts of Formula 1 enthusiasts but also sparked a wave of national pride within the Netherlands. In this study, we aim to investigate the frequency of 'formule 1' on Dutch Twitter before and after the first histor-

ical formula 1 championship of Max Verstappen. We questioned ourselves: Did Max Verstappen's first world championship in formula 1 change the frequency of 'formule 1' on Dutch Twitter? We hypothesize that the frequency of 'formule 1' will be high around the time Max Verstappen became world champion, reflecting the reach of his influence in the Netherlands. By investigating the frequency of 'Formule 1' tweets during the month of Max Verstappen's first World Championship, we provide insights into how significant sporting events can shape public sentiment and generate heightened engagement on social media platforms like Twitter. This research also sheds light on the power of sport icons to influence public narratives and cause a sense of national pride. Analyzing the frequency of 'Formule 1' tweets allows us to examine the extent to which Verstappen's success impacts online conversations, which can serve as a scale for public engagement when a national wide event occurs.

2 Related Work

The study by Rajput et al. (2020) presents a statistical analysis of Twitter messages related to the COVID-19 pandemic. Their research is somewhat similar to our research, focuses on an event, they also use word frequency analysis as well for their research. With their word frequency analysis, the authors examine the patterns and trends in the words used on Twitter during the pandemic. Their results suggest that certain patterns in usage of words are related to their event which is the COVID-19 pandemic. Their conclusion strengthens the hypothesis of our project, namely the word frequency analysis.

Another study we found discusses the development of a system for detecting local events

using Twitter data. The study by Lee and Sumiya (2010) focuses on utilizing blogging sites like Twitter, which enable people to share updates from any location, to indirectly monitor crowd behaviors and identify local events. This study relates to our study in terms of a significant local event. Their research method of using Twitter for building the system also strengthens our hypothesis. Their study method suggests that Twitter is a good source for getting information for relevance on a local event.

3 Data

This research utilizes a by the Rijksuniversiteit Groningen provided Twitter corpus called, the Dutch Tweet Corpus. The corpus covers a span of multiple years and keeps getting updated. For this research, we only used the month December of the year 2021, since Max Verstappen became world champion for the first time in this month. The mainly used feature within this corpus is tweet2tab function, with this we could extract the tweets and their text.

Data Collection The data was collected by the web scraper utilizing the Twitter API, to gather tweets, retweets and mentions containing the word "formule 1" and "f1"¹. "formule 1" and "f1" were collected so that every version of the word matches the word we sought uppercase or not². By gathering the data this way, we made sure that we had all possible matches for "formule 1" and "f1".

Data Processing The Twitter data of "Formule 1" and "f1" was collected around the time of Max Verstappen's first world championship in formula 1. For gathering the data we used a bash script in the terminal of the Twitter API. This however has some limitations, the Twitter API only allows access to a limited amount of tweets, which may not capture the entire base of the "formule 1" discussion online. Also we only focused on the month where Max Verstappen became world champion, because of that our research does not any pre-existing trends of 'formule 1' tweet frequencies in consideration.

¹Formule 1 is the Dutch word for formula 1, that is why formule 1 is used as match.

²This was done in the corpus with the -i grep command, this makes the tagged word incasesensitive so it matches all tags

Pre-Processing We processed the data by³ .:

- Writing a bash script which is run in the terminal of the Twitter API. The bash script searches for either "formule 1" or "f1" and prints the results of the word count per day.
- Tokenization, using -iw we made sure so that all characters match.
- Optimizing the code so it searches for the whole month but still prints output per day.
- Downloaded and saved the data

Table 1 provides an example of the output of the data.

Date	"formule 1" word count
1	15857
5	23345
12	104150
16	14357

Table 1: Example of the collected and analyzed tweet data.

4 Analysis and Results

In this section we discuss the method used in this study, the frequency of "formule 1" around the period, Max Verstappen became world champion of the formula 1 for the first time. The bash script loops through each day of the month and prints the word count for each day. For the time period we took between 1 December and 31 December 2021 these are the results:

Results Table 2 shows the word count of "formule 1" per day in December 2021.

Table 3 shows the word count of "f1" per day in December 2021. Based on our results, there is a significant difference between the frequency of "formule 1" and "f1" around Max Verstappen his first world championship.

Based on these results, it seems that our hypothesis is supported. Our hypothesis stated that the frequency of "formule 1" and "f1" would be higher around Max Verstappen his first world championship. Viewing the results we can see that

³The code for full reproducibility can be found here: <https://github.com/JoostOving/research-methods-resit>

Date	"formule 1" word count
Day 01	15857 words
Day 02	11966 words
Day 03	7740 words
Day 04	5070 words
Day 05	23345 words
Day 06	11561 words
Day 07	5828 words
Day 08	28693 words
Day 09	12746 words
Day 10	14917 words
Day 11	20104 words
Day 12	104150 words
Day 13	33474 words
Day 14	10841 words
Day 15	23879 words
Day 16	14357 words
Day 17	10345 words
Day 18	3923 words
Day 19	3514 words
Day 20	4390 words
Day 21	3380 words
Day 22	6838 words
Day 23	7170 words
Day 24	3373 words
Day 25	2344 words
Day 26	6908 words
Day 27	4482 words
Day 28	2035 words
Day 29	1548 words
Day 30	2269 words
Day 31	2360 words

Table 2: Word count of "formule 1" on Twitter in December 2021

Date	"f1" word count
Day 01	55201 words
Day 02	60198 words
Day 03	36331 words
Day 04	59671 words
Day 05	194719 words
Day 06	68574 words
Day 07	25400 words
Day 08	86882 words
Day 09	59241 words
Day 10	63622 words
Day 11	112651 words
Day 12	529008 words
Day 13	142044 words
Day 14	50594 words
Day 15	59431 words
Day 16	44248 words
Day 17	35385 words
Day 18	20161 words
Day 19	14473 words
Day 20	12485 words
Day 21	17911 words
Day 22	15234 words
Day 23	14129 words
Day 24	8625 words
Day 25	6020 words
Day 26	6713 words
Day 27	11761 words
Day 28	6060 words
Day 29	6546 words
Day 30	8365 words
Day 31	10623 words

Table 3: Word count of "f1" on Twitter in December 2021

the difference is significant and we can conclude that our hypothesis is supported.

If we look into these results we can see that the frequency of "formule 1" and "f1" did spark up in the beginning of the month. We can also see that on the day (12 December) Max Verstappen became world champion the frequency is the highest in the whole month. The word frequency in the beginning of the month is the highest which probably is due to the formula 1 season still ongoing. For example with both "formule 1" and "f1" we can see that there also was a spike in frequency on 5 December. This is due to the second to last formula 1 race happening that day. We also see that "f1" is mentioned way more than "formule 1". We are not sure why this is the case but it could be that there is another meaning for "f1".

To further verify these results, a hypothesis test could be performed to determine if the observed difference is significant. A hypothesis test is essential to provide clear evidence and to confirm that the results are reliable. We chose the "formule 1" word for the test since its more reliable due to the fact that the word "formule 1" is directly linked to the sport.

We first did a simple t-test and after that calculated the p-value.

- Word Count on Day 12: 104150
 - Average Word Count (excluding Day 12): 9958.4333
 - Sample Standard Deviation: 18513.75
 - Number of Observations: 31
- $$t = (\text{mean-day12} - \text{mean-all}) / (s / \sqrt{n})$$
- $$t = (104150 - 9958.4333) / (18513.75 / \sqrt{31})$$
- $$t = 28.4081$$

$$p\text{-value} = .00001$$

The p-value is less than the significance level of 0.05 which means that the null hypothesis can be rejected. The p-value indicates that the observed difference on Day 12 is highly unlikely to have occurred by chance alone, providing strong support for the hypothesis that Max Verstappen's

first world championship influenced the mentions of "formule 1" and "f1".

5 Conclusion

Based on the analysis conducted in this study, we aimed to investigate the impact of Max Verstappen's first world championship on the frequency of mentions of "formule 1" and "f1" on Dutch Twitter. Our hypothesis suggested that the frequency of these mentions would be higher around the time of Max Verstappen's championship. The results indicated that our hypothesis was supported, as the observed difference in word frequency was statistically significant. The data showed an increase in mentions of "formule 1" and "f1" leading up to and on the day of Max Verstappen's victory. The higher frequency at the beginning of the month may be attributed to ongoing Formula 1 races, while the substantial spike on the championship day demonstrates the impact of Max Verstappen's success on public engagement. The study did however have some limitations, the study did not look in months and years before for averages of "formule 1" and "f1" in Dutch tweets. If the study did do this we could say with more clarity it was really due Max Verstappen and not just because of formula 1 being popular in the Netherlands. Future studies can maybe elaborate on this by using a larger dataset covering multiple months and years. The data collection process also relied on a web scraper utilizing the Twitter API, which has its limitations. The Twitter API only provides access to a limited number of tweets, which does not cover the entirety of "formule 1" discussions on Twitter. These are multiple reasons why the results of this study should be approached with caution. These findings highlight the power of significant events and icons in generating national pride, and increasing engagement on social media platforms like Twitter. All in all this study contains useful insights into the frequency of "formule 1" and "f1" on Twitter, however there is huge potential for further research in this research area.

References

Lee, R. and K. Sumiya (2010). Measuring geographical regularities of crowd behaviors for twitter-based geo-social event detection. In *Proceedings of the 2nd ACM SIGSPATIAL international workshop on location based social networks*, pp. 1–10.

Rajput, N. K., B. A. Grover, and V. K. Rathi (2020). Word frequency and sentiment analysis of twitter messages during coronavirus pandemic.