Contents

1. Introduction

The second screen phenomenon explains the engagement on multi-screen by TV viewers. While Meulenaere et al. (2015) argue that the multiscreen can post a challenge to TV viewers' attention, Anderson (2018) states that the second screen can retain users' participation and foster their online discussion. Using machine learning to analyse 13,524 live tweets, we aim to improve understanding among Supercars fans and provide suggestions to Supercars' management. This report first lists the questions that can and cannot be answered by the data. Second, the exploratory data analysis (EDA) provides a glance at the quantitative data. Third, data pre-processing including tokenisation and lemmatisation is conducted before analysing. Fourth, regression, sentiment analysis and topic modelling have been used to analyse the tweets. Lastly, recommendations are provided based on the model results.

1.1 Questions that can't be answered

Supercars should understand their audience firstly to implement an effective second screen strategy (Nizam, 2020). There are several questions we are interested to know. First, what are the demographic characteristics of tweet users who regularly engage with Supercars events? However, since the survey data lacks variables relating to demographic information, the question cannot be resolved.

Second, is there any tweet from the official accounts, such as Supercars or the team's Twitter accounts? What are the topics, and how do fans respond to these topics? The details about who send the tweets and how many tweets they send can be accessed by analysing the dataset, but it is difficult to distinguish between fans' accounts and official accounts due to the absence of an official account list.

Third, if advertisers want to work with a team or player, whom can advertisers reach out to via Twitter? This is a network question, but the data does not reveal any detail about the connections between players' or teams' accounts.

1.2 Questions that can be answered

Regression analysis, sentiment analysis, and topic modelling were used to address the following two questions. First, among TV channels, tweet content, weekday and start time, which factors have a great impact on TV viewers? Second, how does Supercars improve tweet engagement based on fans' attitude?

2. Data Pre-processing

Text Pre-processing

Before analysis, tweet texts should be cleaned. Tweet texts are firstly transformed to lower case, then applied by some text pre-processing techniques like tokenization and lemmatization. Punctuations, stop words, "#" and "@" before text and texts in square brackets e.g. [sdf.12] are removed. Since hashtags are removed here, to analyse which hashtag topics are popular, texts after hashtags are extracted separately and visualized in Figure 11 in Appendix.

Feature Engineering

To analyse tweet data combined with TV viewership, tweet data are aggregated every 15 minutes and joined with TV data. It is assumed that any tweet starts to impact after 15 minutes posted. After joining, null values are dropped. Noticed that the commas in the TV channel column are used to split the cell into multiple rows.

To analyse whether including web links or videos would impact TV viewership, web links are substituted by a string "url" and two dummy variables for including web links or videos are generated respectively. Dummies for TV channels and cities are also generated for regression analysis. Besides, start time is encoded from 7:15, 7:30, 7:45… to 0, 1, 2… Total engagement is calculated as the sum of retweets, likes and comments (Anderson & Xu, 2019).

3. Analysis

3.1 EDA

3.1.1 Bigram

Research (Chen, Zafar, Galperin-Aizenberg, & Cook, 2018) shows that the best-performing NLP model is the combination of tokenized unigrams and bigram. Increasing N-gram length makes little to no improvement for most machine learning algorithms. Hence, we utilise bigram to help to understand the dataset which can be shown in a bar chart. In addition, the same result can also be seen using a word cloud (Figure 12 in Appendix).

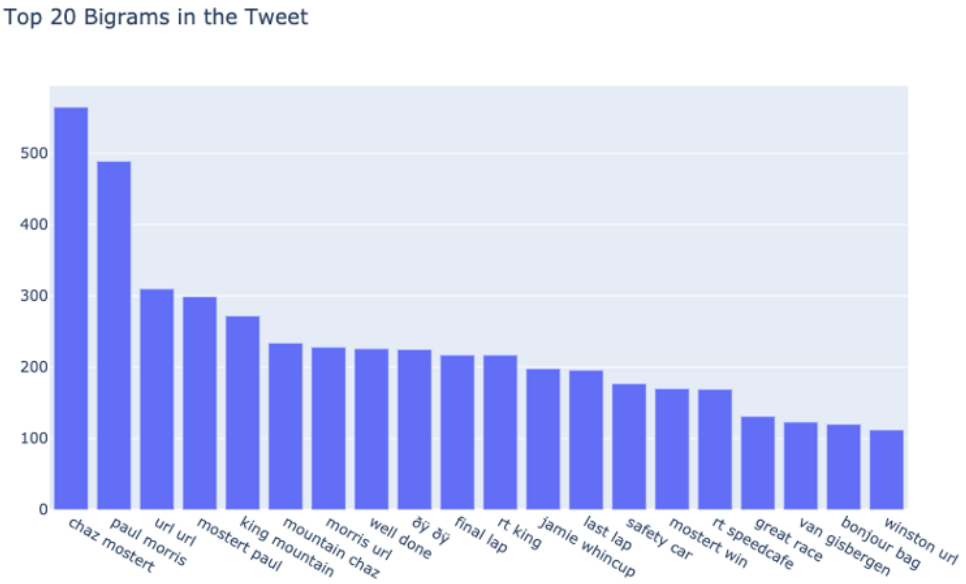

Figure 1: Bigram

There are several findings according to the bigram (Figure 1). First, the top two keywords are mentioned significantly higher than others, which are “chaz mostert” and “paul morris”. The two racing drivers seem to be influential key opinion leaders who draw the most attention online. Second, “final lap” and “last lap” are both on the top keyword lists. Third, surprisingly, “safety car” has high frequency indicates that fans value the safety of the races and Supercars should prioritise the safety measures for players during races.

3.1.2 Unigram

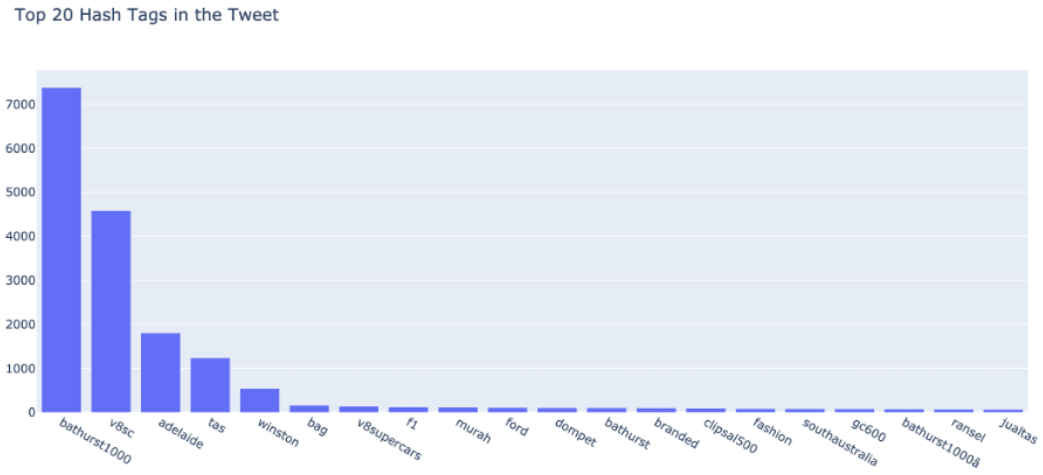

Figure 2: Unigram

We apply Unigram to find out hashtags’ trend (Figure 2). Top hashtags include the major Supercars annual events such as #bathurst1000 for Bathurst 1000, #v8sc for V8 Supercars, #adelaide and #clipsal

#southaustralia for Adelaide 500, #tas for Tasmania SuperSprint, #f1 for Formula One World Championship and #gc600 for Gold Coast 600.

Interestingly, there's no Supercars player name #winston, therefore, we looked up the actual tweets and find out most of them are about the scandal of Jamie Winston, a US football quarterback.

Another unusual thing is that some words do not seem to be related to Supercars, such as #bag, #murah, #branded #fashion #dompet (means wallet in Indonesian) #ransel (means backpack in Indonesian). After research, it turns out that "Jual tas branded murah" means "Selling cheap branded bags" in Indonesian and those tweets are ads (Figure 3). It is surprising how many of the ads out there that they even go on the list of the top hashtags.

Figure 3: Ad Sources

Although EDA provides crucial information in terms of which keywords and hashtags should be paying more attention to, it doesn't tell us whether these keywords are discussed in a positive or negative attitude and which topics are they associated with. Therefore, the following will be using machine learning models to disclose and more in-depth analysis.

3.2 Among TV channels, tweet content, weekday and start time, which factors have a great impact on TV viewers?

This section aims to provide suggestions on how to improve TV viewership for Supercars by investigating the relationship between the number of TV viewers and other factors through regression.

A heatmap is generated for examining the multicollinearity. According to the heatmap in Figure 14 in Appendix, since cities are highly correlated with local TV channels and the number of tweets is highly correlated with other tweet relevant variables (player, url, total engagement), cities and number of tweets are removed from the regression analysis.

	coef	P>|t|
const	0.0715	0.739
weekday	0.3960	0.000
number_of_player_mentioned	-0.0211	0.000
number_of_team_mentioned	-0.0117	0.298
number_of_video	0.0525	0.001
number_of_url	0.0138	0.000
total_engagement	0.0004	0.028
tv_channel_7mate Ade	-0.7453	0.000
tv_channel_7mate Bri	-0.3139	0.103
tv_channel_7mate Mel	0.1979	0.196
tv_channel_7mate Per	-1.0169	0.000
tv_channel_7mate Syd	-0.3976	0.044
tv_channel_7mate Tasmania	-1.7646	0.000
tv_channel_ATN7	0.8507	0.000
tv_channel_BTQ7	0.7747	0.000
tv_channel_HSV7	0.9535	0.000
tv_channel_SAS7	-0.0645	0.676
tv_channel_Southern Cross Tasmania	-0.7521	0.000
tv_channel_TVW7	-0.2736	0.077
start_time_new	0.0328	0.000

Figure 4: OLS Result

According to Figure 4, tweet relevant variables (player, team, video, url, total engagement) have a relatively smaller impact on TV viewership than TV channels. Among all tweet relevant variables, "number of video" has the largest impact on TV viewership. The three TV channels with the largest positive coefficients are HSV7, ATN7 and BTQ7; while the three TV channels with the smallest negative coefficients are 7mate Tasmania, 7mate Per and Southern Cross Tasmania.

Recommendations:

There are four recommendations for Supercars. Firstly, since TV channels have a greater impact on TV viewership than tweet relevant variables, Supercars could focus more on the collaboration with TV channels like HSV7, ATN7 and BTQ7. One of the reasons why these TV channels have a positive impact on TV viewership might because of their shooting angles where Supercars could put important ads or notifications accordingly.

Secondly, the reason why the TV channels in Tasmania have a negative impact might be lack of marketing and advertising, thus Supercars could add more ads about the events in TV channels like 7mate Tasmania and Southern Cross Tasmania and advertise in some traditional channels like newspapers and magazines. However, since the potential value of the fans in Tasmania is unknown from the data, Supercars might need to do more research before investing.

Thirdly, since "number of video" has a positive impact on TV viewership, Supercars could include more videos in the tweets (Zote, 2020).

Fourthly, since the weekday and start time have a positive impact on TV viewership, Supercars could schedule important events in the afternoon and at the end of the week.

3.3 How does Supercars improve tweet engagement based on fans' attitude?

Brands can use live tweeting to increase brand exposure, inspire more people to engage in events and draw fans' attention during the events (Tran, 2020). Furthermore, brand exposure can be increased by improving user engagement including retweets, shares, and comments (Barker, 2019). As a result, Supercars can benefit from improving user engagement on Twitter which can lead to a higher reach of audiences during Supercars events.

To identify effective strategies for improving engagement, Anderson and Xu (2019) claim that positive tweets are more likely to inspire user engagement than negative or neutral tweets. Therefore, understanding what positive tweets are talking about is vital for Supercars to strategically share topics relevant to these positive contents to increase user interaction at following events. Although negative tweets have a lower impact on user engagement improving, Supercars can use negative contents to learn about brands' disadvantages and strengthen them.

Overall, this question aims to figure out the most positive and negative topics to help Supercars to increase user engagement and enhance future success. To answer the question, sentiment analysis is first applied to the data which divided all tweets into positive, negative and neutral categories. Topic modelling is then applied to both positive and negative tweets to further evaluate the tweets.

3.3.1 The positive tweets

All positive tweets were grouped into 10 topics by topic modelling. The top three discussed topics result from the topic map are shown in Figure 5, 6 and 7. According to Anderson and Xu (2019), the idea of engagement on Twitter can be identified as the sum of retweets, likes, and comments. Therefore, "total engagement" was created by adding the sum of the number of retweets, likes, and comments for each tweet.

Topic 1, 2 and 3 were the most discussed amongst all tweets during Supercars events, which occupied the proportions of 15.5%, 13.8% and 11.4%, separately.

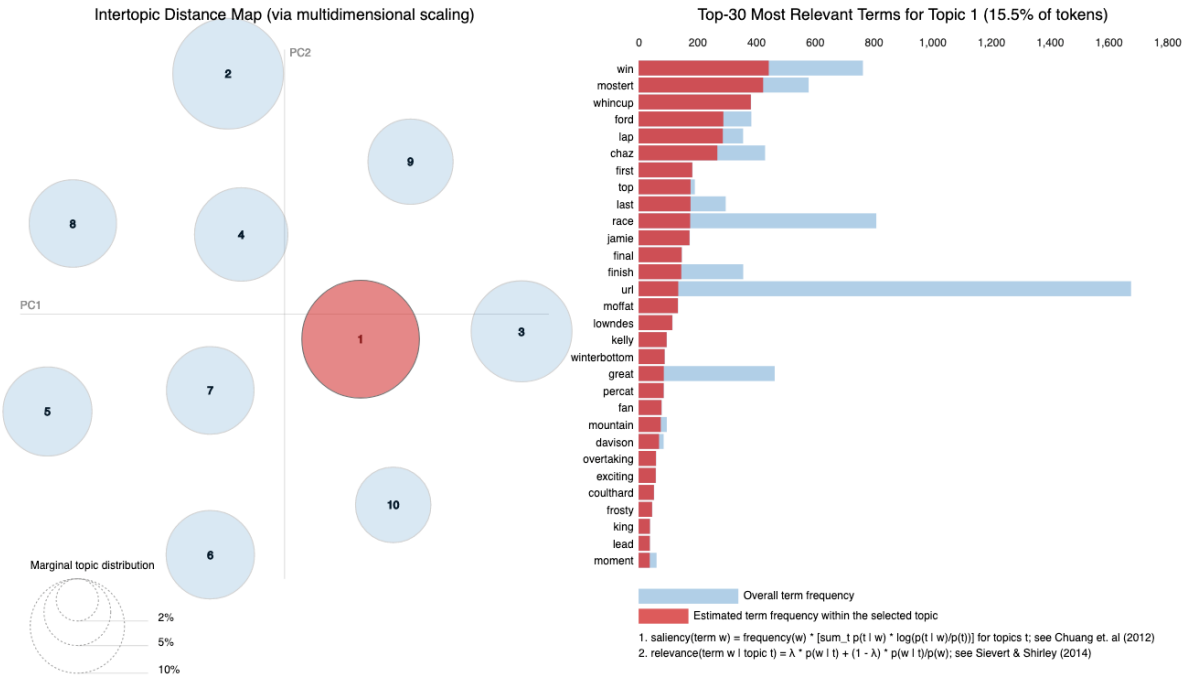

Figure 5: Positive Topic 1

According to Figure 5, as tweets included the topic-specific words like win, Mostert, Whincup, ford, first, last, top, these tweets are most likely from Topic 1, which is about Chaz Mostert's victory after passing Jamie Whincup on the last lap and discussion of top drivers in Bathurst 1000. This topic shows the top drivers in Bathurst 1000 are Mostert, Whincup, Winterbottom, Moffat, Percat, division, Kelly and Coulthard. Besides, fans are excited about Mostert's win.

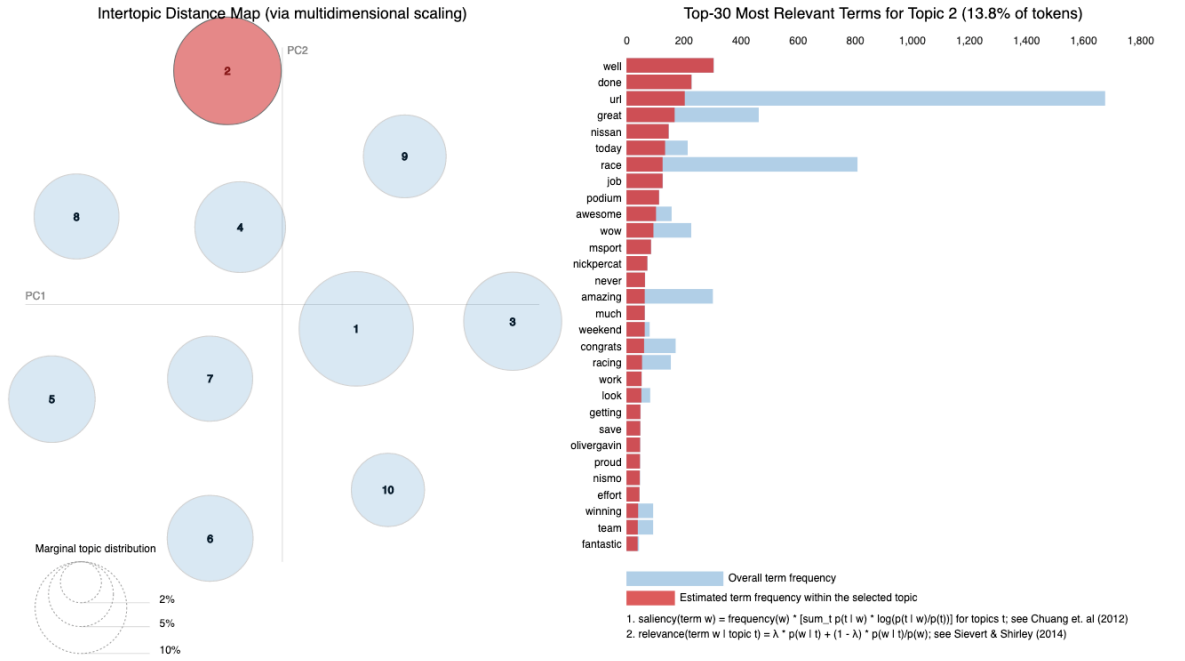

Figure 6: Positive Topic 2

Figure 6 shows the most relevant terms for Topic 2, which is about congratulation of Nissan team get the second place, and Nick Percat and Oliver Gavin in their third place.

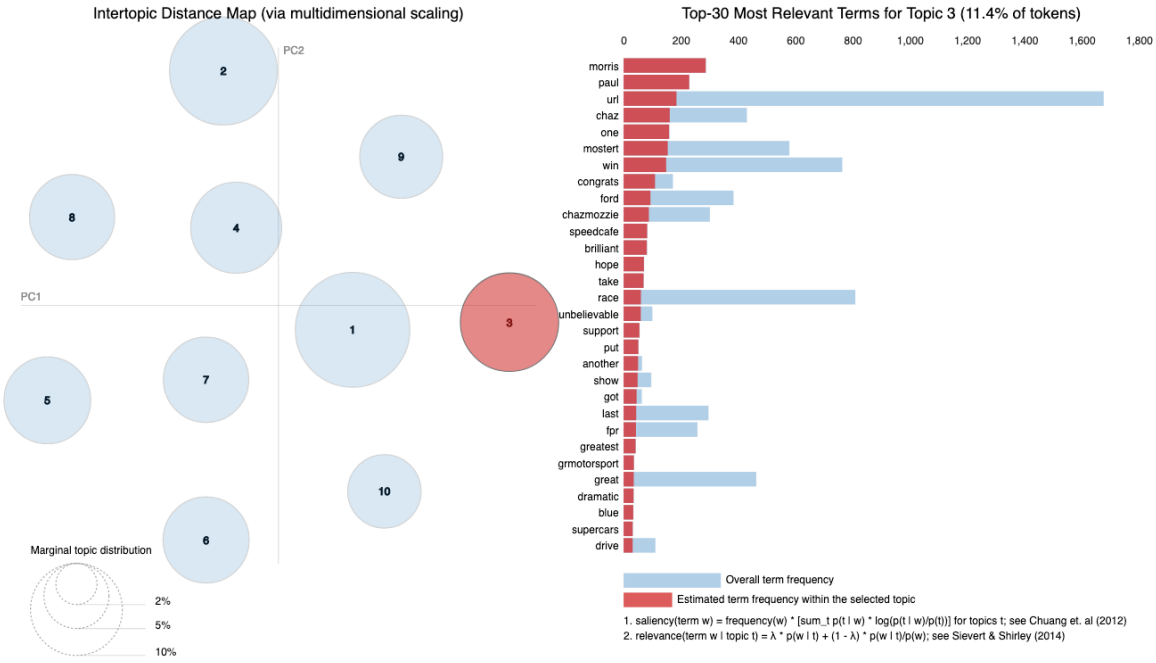

Figure 7: Positive Topic 3

In Figure 7, the Most relevant terms for Topic 3 are Chaz, Mostert, paul, morris, congrats, speedcafe and win. This topic is for congratulating Chaz Mostert and Paul Morris on their wins.

Table 1: Top 3 Positive Topics

	Average sentiment scores	Total engagement	Number of tweets
Topic 1	0.575	804	783
Topic 2	0.576	332	816
Topic 3	0.581	281	626

It is fascinating to note that all these three positive topics are related to the match outcome and congratulation for the winners. Besides, by observing Table 1, the average sentiment scores for all three topics are almost identical, but Topic 1 has around 2.5 times the amount of the other two for total engagement. Therefore, users are more interested in Topic 1 that is related to Chaz Mostert.

Recommendations:

Firstly, Supercars need to refresh the match process and results timely to increase user engagement, because it is what audiences are most interested in. Hashtags, which can make the tweets rank better in search results, need to be used when updating the race outcome (Lozano, 2016).

Secondly, Supercars can launch campaigns on Twitters before the event to increase event awareness and attract more audiences. To do so, Supercars can create a topic for Twitter users to forecast match outcomes. To encourage users to actively engage in the topic, Supercars can award winners' autographs to those who accurately forecast the result.

Thirdly, the winners of previous events can serve as spokespersons for Supercars, like Chaz Mostert and Paul Morris. Supercars can post their pictures, videos or even relevant driver GIF in the official account during the event to boost user engagement (Zote, 2020).

3.3.2 The negative tweets

On the opposite, the negative tweet should provide different topics which could draw different suggestions. Therefore, we may use the same process as positive tweets used, and the graphs of negative topics can be presented as follows.

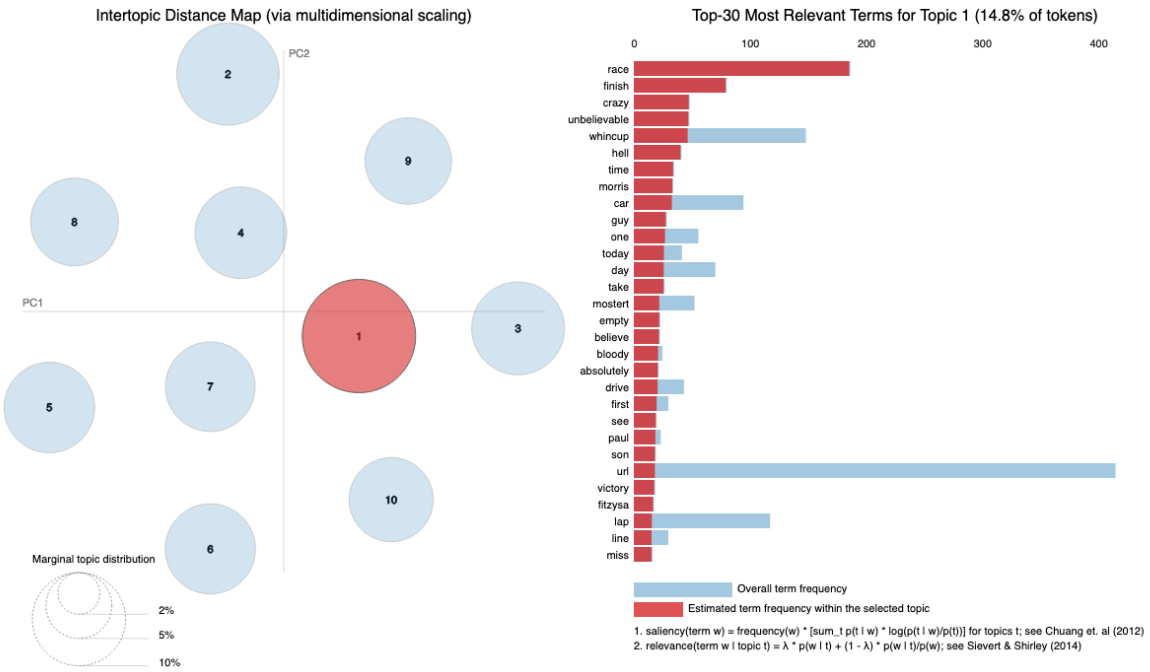

Figure 8: Negative Topic 1

The most important negative topic (Figure 8) mainly focuses on the crashing of Jamie Whincup that happened at the beginning of the race (Sporting News, 2020). People show their regrets to him since they should look forward to the excellent performance of this legend Supercars player. The engagement score and the sentiment score in table 3 can reveal the disappointment of fans.

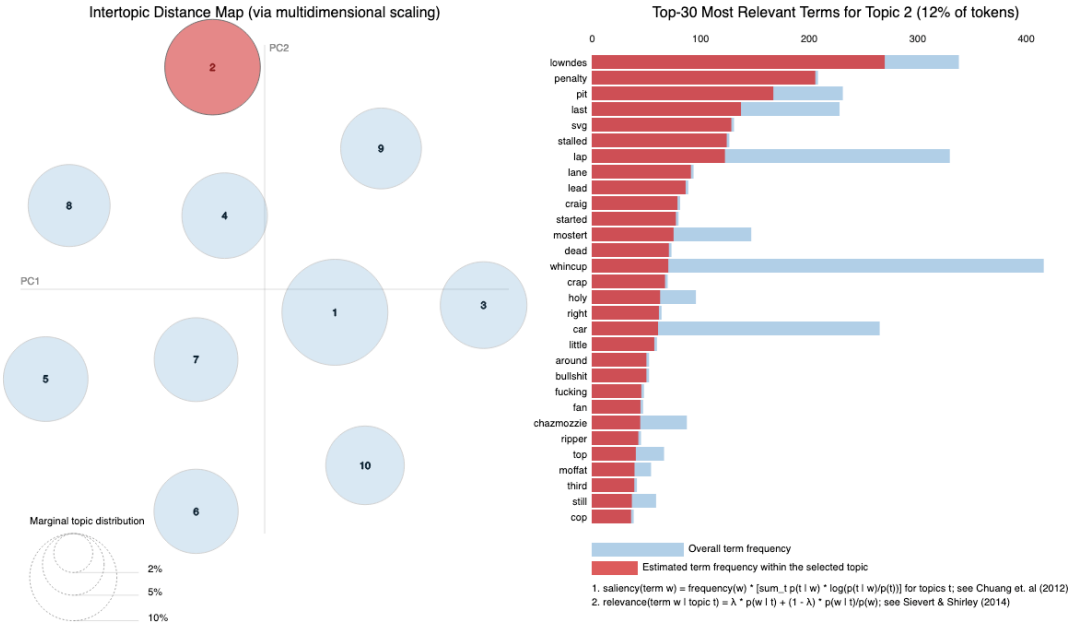

Figure 9: Negative Topic 2

In the negative Topic 2 results (Figure 9), we have known that people also discuss another famous player Craig Lowndes and the penalty he confronted. Similarly, as Jamie Whincup, Lowndes does not win the race since the factor outsides competition. Since the found guilty that breaching the FIA's Obligation of Fairness regulations, the team were decreased 300 championship points until the end of 2021 (AAP, 2019). People said many disappointed words on this topic to shows their pit feeling for Lowndes and unsatisfactory with the race.

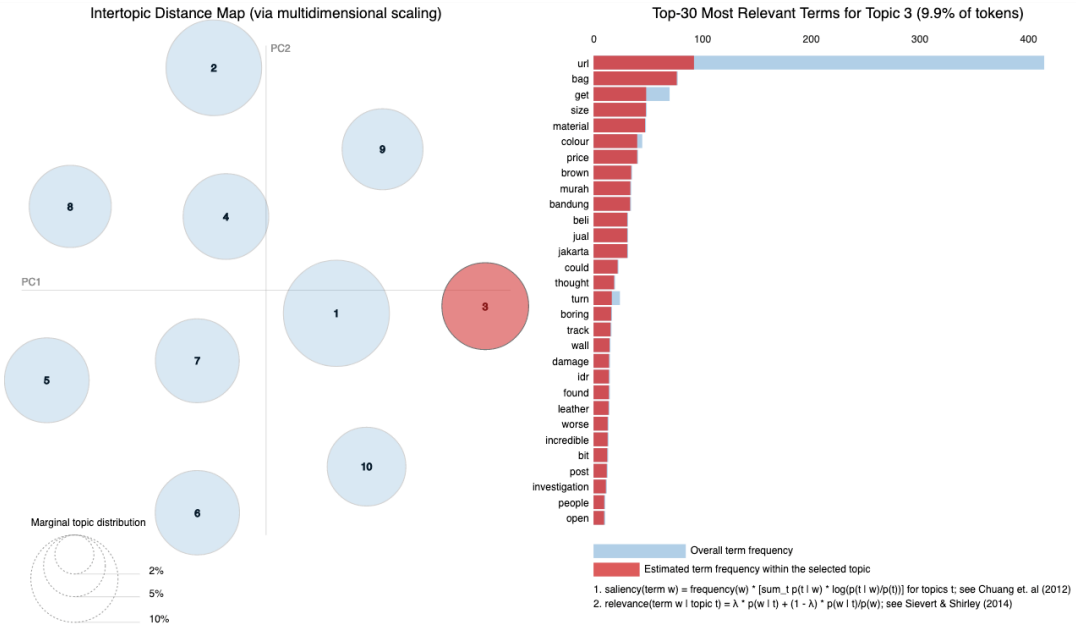

Figure 10: Negative Topic 3

Topic 3 in Figure 10 is an uncommon one because it is all about advertising. The main word in this topic provides information about goods, such as material, size, colour, the purchase website, and prices. Based on the sentiment analysis, we can know that advertisements lead to the negative impression of Supercars tweets. It is reasonable that people do not like a commercial advertisement when they care more about races and players.

Table 2: Top 3 Negative Topics

	Average sentiment scores	Total engagement	Number of tweets
Topic 1	-0.486	31	345
Topic 2	-0.426	19	257
Topic 3	-0.468	3	128

Based on the result in Table 2, the number of negative tweets and the total engagement has decreased if we consider them in the positive tweets. It can be interpreted as a recommendation to improve Supercars' management performance.

Recommendations:

Firstly, Supercars' social media management team may find the appropriate method to tackle the commercial advertisements mixed in their Twitter discussion. It can only decrease the discussion quality but neither improve users' engagement non-TV viewers. Applying an official news channel within Twitter may give them more management abilities for their news discussion.

Second, the management team should also focus more on racing safety. Based on Topic 1, we may know that car crushing, especially for famous players, would be an important topic and increase people engagement. However, the discussion of crushing is negative. People deliver their pit, disappointed feeling on these topics. Even though it can attract more people in the short term, people may feel depressed and reduce their passion in the long term.

Compared with positive tweets, people's engagement has decreased, which means their emotional impulses are not as large as seeing the positive tweets. Therefore, the final suggestion could be that, to increase people engagement and to improve people's impression of race, Supercars may share more positive tweets, such as exciting results, which could influence more people. Because the spontaneous engagement of people would attract more audience and help Supercars' races develop better in the future.

4. Conclusion

Overall, our research suggested Supercars post tweets with videos during the game to increase TV rating. And Supercars should leverage the champions who can draw positive engagement and try to improve safety and reduce obvious ad can bring the negative buzz.

There are several limitations to this report. First, the data preparation process designed for English texts only. Therefore, non-English tweets are not considered in the analysis, thus the result only limits to a certain group of fans and may not be representable to the whole population (Xue et al., 2020). Second, our dataset only includes tweets that contain #vasc. Third, our result shows that Event 11 has the highest number of tweets and total engagements. In other words, tweets about Event 11 dominate the whole dataset. Further details of Event 11 might be suggested to be taken into consideration to exam whether the nature of Event 11 has masked some important information about Supercars.

5. References

Anderson, B. (2018). *Winning over Fans: How Sports Teams Use Live-Tweeting to Maximize Engagement*. Retrieved from https://www.elon.edu/u/academics/communications/journal/wp-content/uploads/sites/153/2018/05/06_Anderson_Livetweeting.pdf

Barker, S. (2019). *Why twitter engagement is essential for brands today*. Retrieved from https://blog.markgrowth.com/why-twitter-engagement-is-essential-for-brands-today-f89c9d61514e

Chen, P.-H., Zafar, H., Galperin-Aizenberg, M., & Cook, T. (2018). Integrating Natural Language Processing and Machine Learning Algorithms to Categorize Oncologic Response in Radiology Reports. *Journal of Digital Imaging*, 31(2), 178–184. https://doi.org/10.1007/s10278-017-0027-x

Lozano, D. (2016). *5 simple ways to improve the performance of your tweet*s. Retrieved from https://www.socialmediatoday.com/social-business/5-simple-ways-improve-performance-your-tweets

Meulenaere, J. D., Bleumers, L., & Broeck, W. V. den. (2015). An Audience Perspective on the 2nd Screen Phenomenon. *The Journal of Media Innovations*, 2(2), 6–22. https://doi.org/10.5617/jmi.v2i2.909

Nizam, A. (2020). *A complete guide to second screen media and its impact on marketing*. Retrieved from https://www.lemonlight.com/blog/a-complete-guide-to-second-screen-media-and-its-impact-on-marketing/

Xue, J., Chen, J., Chen, C., Zheng, C., Li, S., & Zhu, T. (2020). Public discourse and sentiment during the COVID 19 pandemic: Using Latent Dirichlet Allocation for topic modeling on Twitter. *PLoS ONE*, 15(9), e0239441–e0239441. https://doi.org/10.1371/journal.pone.0239441

Tran, T. (2020). *How to live tweet an event: Tips, best practices, and examples*. Retrieved from

https://blog.hootsuite.com/how-to-live-tweet/

Zote, J. (2020). *5 strategies to amplify your Twitter engagement*. Retrieved from https://sproutsocial.com/insights/twitter-engagement/

6. Appendix

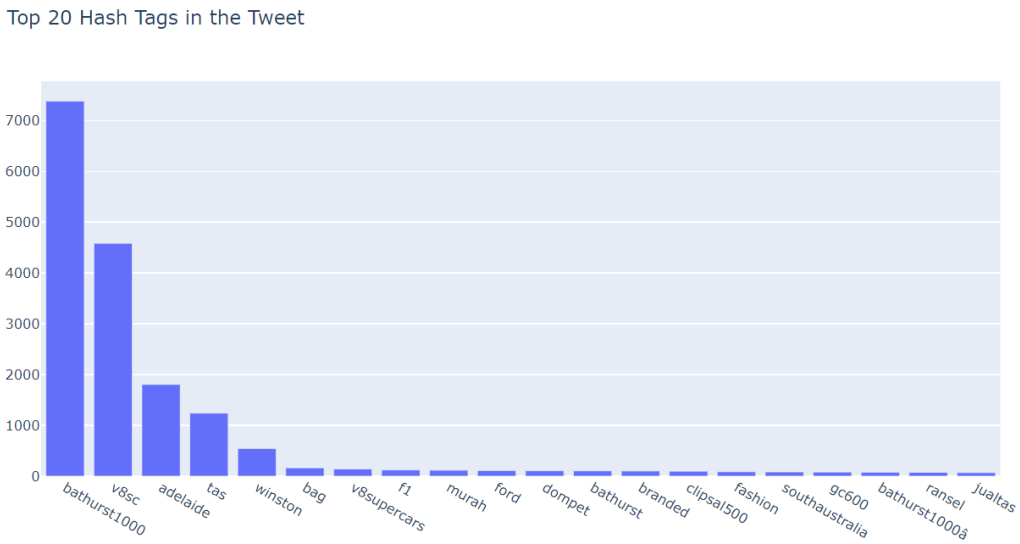

Figure 11: Top 20 Hash Tags in the Tweet

Figure 12: Word Cloud

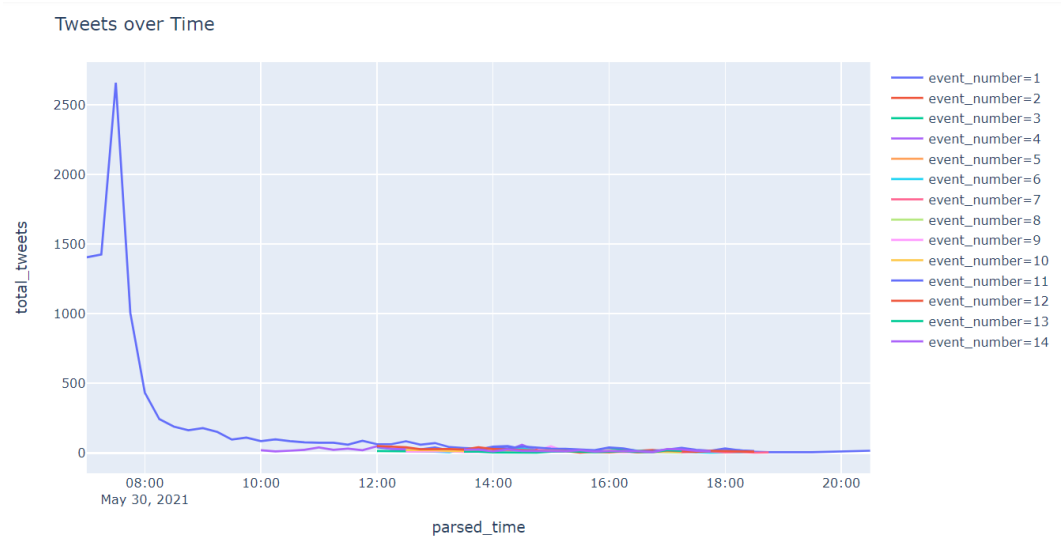

Figure 13: Tweets over Time

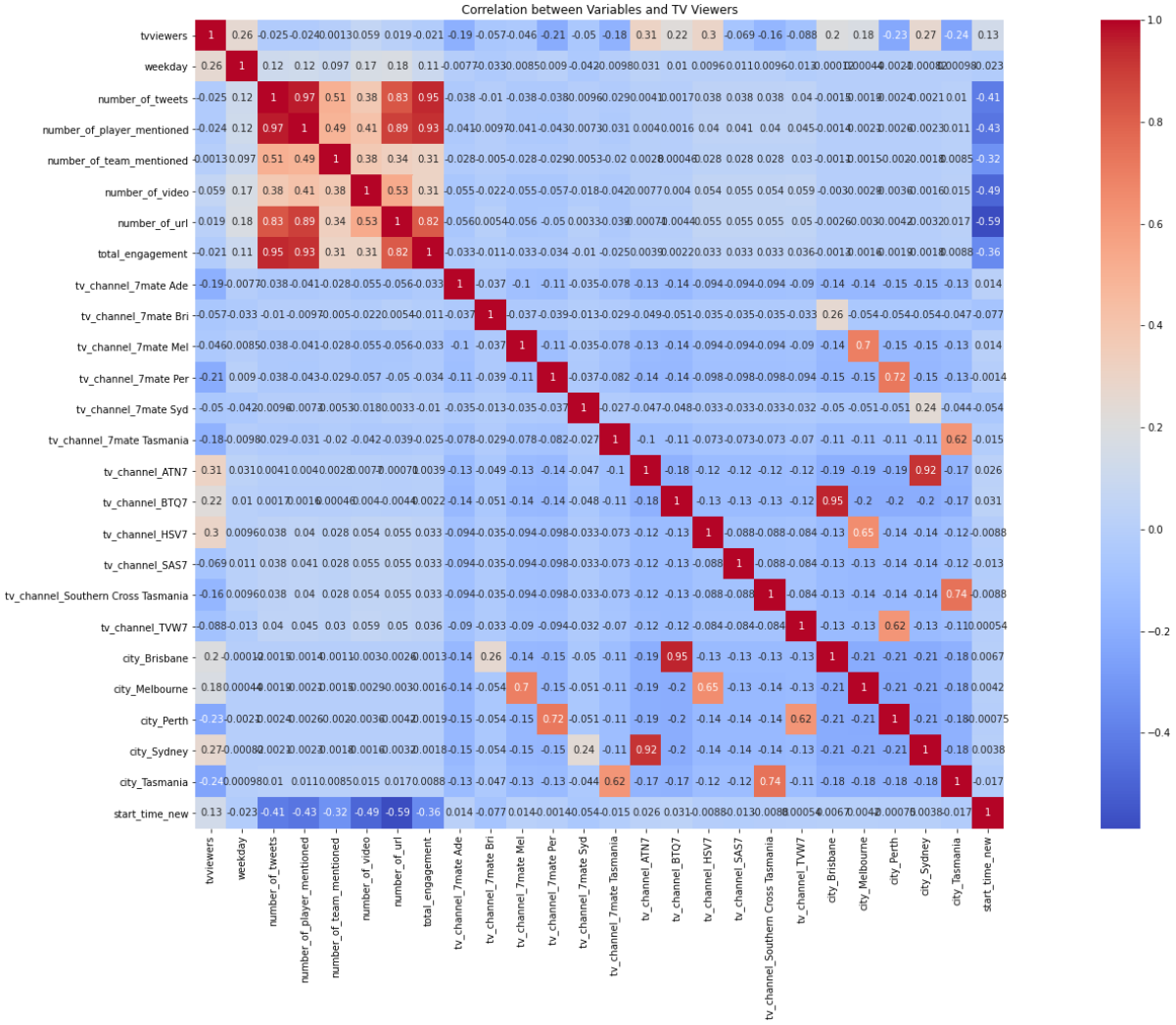

Figure 14: Heatmap

OLS Regression Results

Dep. Variable:	tvviewers	R-squared:	0.580
Model:	OLS	Adj. R-squared:	0.577
Method:	Least Squares	F-statistic:	202.2
Date:	Sat, 22 May 2021	Prob (F-statistic):	0.00
Time:	03:28:59	Log-Likelihood:	-3266.7
No. Observations:	2802	AIC:	6573.
Df Residuals:	2782	BIC:	6692.
Df Model:	19		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0715	0.215	0.333	0.739	-0.349	0.492
weekday	0.3960	0.025	15.706	0.000	0.347	0.445
number_of_player_mentioned	-0.0211	0.004	-5.956	0.000	-0.028	-0.014
number_of_team_mentioned	-0.0117	0.011	-1.040	0.298	-0.034	0.010
number_of_video	0.0525	0.016	3.186	0.001	0.020	0.085
number_of_url	0.0138	0.002	6.140	0.000	0.009	0.018
total_engagement	0.0004	0.000	2.199	0.028	4.49e-05	0.001
tv_channel_7mate Ade	-0.7453	0.153	-4.870	0.000	-1.045	-0.445
tv_channel_7mate Bri	-0.3139	0.192	-1.631	0.103	-0.691	0.063
tv_channel_7mate Mel	0.1979	0.153	1.293	0.196	-0.102	0.498
tv_channel_7mate Per	-1.0169	0.152	-6.669	0.000	-1.316	-0.718
tv_channel_7mate Syd	-0.3976	0.197	-2.017	0.044	-0.784	-0.011
tv_channel_7mate Tasmania	-1.7646	0.158	-11.203	0.000	-2.073	-1.456
tv_channel_ATN7	0.8507	0.150	5.667	0.000	0.556	1.145
tv_channel_BTQ7	0.7747	0.150	5.170	0.000	0.481	1.069
tv_channel_HSV7	0.9535	0.154	6.182	0.000	0.651	1.256
tv_channel_SAS7	-0.0645	0.154	-0.418	0.676	-0.367	0.238
tv_channel_Southern Cross Tasmania	-0.7521	0.154	-4.876	0.000	-1.055	-0.450
tv_channel_TVW7	-0.2736	0.155	-1.767	0.077	-0.577	0.030
start_time_new	0.0328	0.003	12.713	0.000	0.028	0.038

Omnibus:	3854.229	Durbin-Watson:	1.955
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1356982.363
Skew:	-7.698	Prob(JB):	0.00
Kurtosis:	109.705	Cond. No.	8.75e+03

Figure 15: OLS Result