Social Media and Network Analytics (COSC 2671)

"#IStoodWithDan: Tracking the Twitter Sentiment of the Victorian Premier from January to August 2022"

Arlo Rostirolla (s3872916)

Introduction

The coronavirus pandemic, and the resultant policies implemented by governments to mitigate the crisis, have invoked highly polarised views from the populace. The Victorian government, specifically the premier, has arguably received the most attention compared to other Australian state premiers. On Twitter, where much of this political discussion takes place, groups who support or oppose the premier organize themselves via hashtags, such as '#IStandWithDan' or '#SackDanAndrews".

It is common in this environment to witness the same tweet posted by multiple different users, suggesting a level of automated activity. Indeed, this was the main finding by a previous study on the same topic [1], who found single users may also repeat the same post over multiple twitter threads. This co-ordinated activity must be distinguished from organic activity; what ordinary Australian twitter users think of the Victorian premier.

The purpose of this report is to determine the average sentiment regarding the Victorian premier. The first aim is to quantify the level of co-ordinated and/or automated activity, and separate this data from organic tweets for further analysis. The second aim will be to determine sentiment for the premier over the period from 1st of January 2022 to 22nd August 2022, and model topics commonly discussed in these tweets.

Methods

Data Scraping

Multiple methods were used to scrape data from twitter due to the limitations of a regular twitter developer account. The python library Twint was initially used to scrape tweets, however it only returned small amounts of tweets at a time. A for loop was run to iterate over every day of 2022 from January to August. An attempt was made to use the Tweepy streaming API, however this method was exchanged for the Twarc API, which automatically returns all data fields and is more suited for research uses. Previous analyses on this very subject have obtained data via Twarc [1]. The final uncleaned dataset was comprised of 69,317 tweets. Of these, 7898 were historical tweets from January to July 2022, scraped

using the Twint library; 9258 were collected via the REST API; and the remaining 52,161 were streamed over the period of the 6th to the 22nd of August.

Data Cleaning

Data from all scraping methods were merged into a single data frame, and 5,388 duplicates caused by overlapping scraping methods were removed. 31,350 retweets were isolated to their own dataframe for separate analysis. Redundant fields were dropped, and a hamming distance was calculated for every pair of tweets in the dataset. Any tweet pairs with a hamming distance of less than 10 were assumed to be inorganic and isolated into their own dataframe. Initially the hamming distance was set to 0, however it was noticed in the data that some inorganic tweets used the same tweet, with slight changes in vocabulary and structure, presumably to evade detection. This led to 16 repeated tweets being quarantined into a duplicates dataframe.

15,974 tweets that did not contain mention of the 22 keywords related to Dan Andrews in the actual tweet body, or that mentioned multiple WEF associated world leaders targeted by conspiracy theories, were removed. Remaining were 16,589 tweets, which could be considered organic. Regular expressions were used to remove username tags, URLs and any non-English characters from the data. Any emojis contained in the tweet were 'demojized', or converted into their English description to retain their semantic value for analysis. Irrelevant or sparse columns such as 'photos' and 'attachments.poll.end_datetime' were dropped from the dataframe.

Data Analysis

Pandas was used for basic data analyses. Top K words and hashtags were visualised using Matplotlib. NLTK was used to tokenize the data, and find concordances, bigrams, trigrams and quadgrams. PyLADavis and wordcloud were used for Latent Discriminant Analysis (LDA). UCTopic, a relatively new method called unsupervised contrastive learning [1] was used for phrase mining.

Results

Data analysis with pandas and matplotlib (Figure 1) revealed that sackdanandrews was the most common word with 3844 occurrences (considering words to be a series of characters without space between them). Labor was second with 1336 occurrences, and 'istandwithdan' was third with 1063. Ignoring descriptive hashtags commonly used by news organizations (#springst, #auspol etc), the most common organic hashtag was #sackdanandrews with 103 occurrences, and the second was #dictatordan with 58. The supportive #IStandWithDan hashtag had twenty occurrences.

Figure 1.

Top 20 words and hashtags in the set of organic tweets

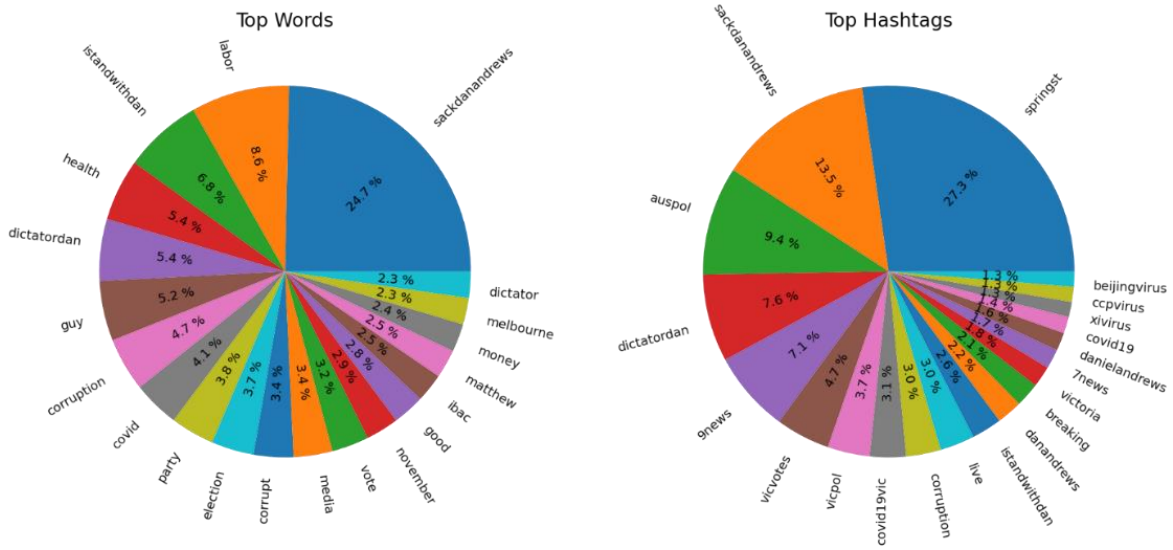


Figure 2

Tweet length descriptive statistics

Words *Characters*

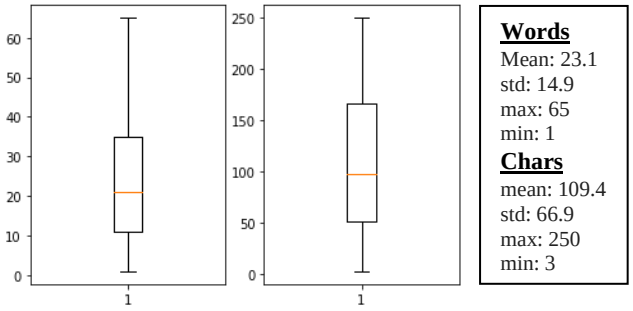


Figure 3

N retweets descriptive statistics

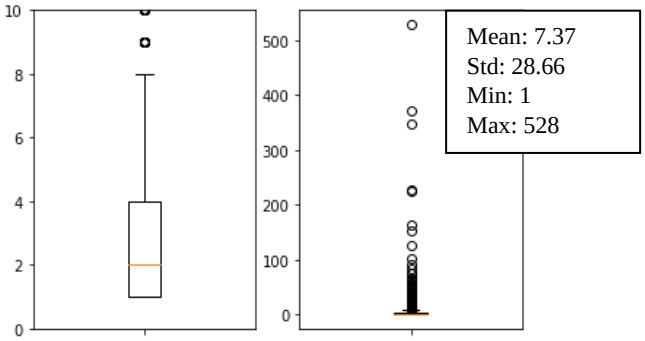


Table 1

Top 5 most retweeted tweets.

Tweet	Retweets
It's not going to be safe for people who are not vaccinated to be roaming around the place spreading the virus, that's what they'll be doing	528
Guy s pollblivion. The Andrews government has increased its lead over the opposition on 60.5 ALP vs 39.5. The horror result for Matthew Guy includes a scathing assessment of his woeful performance far behind as preferred premier on 34 vs Dan Andrews 66.	371
This is not an apology it is just more excuses from a guilty man who knows his decisions destroyed so many Victorians not forgiven #sluggate #sackdanandrews #sackbrettsutton	348
As the vic election draws near will be asking you to re-elect him he will talk about how proud he is of you for the sacrifices that you made over the past 2 yrs n never forget he blamed you never himself he lectured you amp belittled you daily	227
Dan Andrews hotel quarantine staff had an illegal party as Victorians were going through the longest lockdown in the world his famous words then were don t have a drink with your mates more lies discovered today in corruption hearing with emails discovered	225

Table 2.
7 positive and negative concordances from the corpus

Positive
Hit jobs by pro liberal media amp disgruntled corrupt haters on Dan are disgraceful
The liberal political pool is so shallow that there is no legitimate alternative to Dan Andrews
Should I say still Dan's going to win best get used to it
Andrews was not the subject of direct adverse comments opinions or findings by IBAC or the ombudsman Dan has been vindicated.
Because of this our families and our economy bounced back stronger than ever thanks Dan
I absolutely love Dan Andrews where is the evidence of these supposed rorts
How dictator Dan defied a dangerous Murdoch media and led Australia to covid victory
Negative
I have lost faith in IBAC after they cleared Daniel Andrews like he always gets called teflon Dan and gets away with everything
If they were wrong about dictator dan then why didn't he sue them for defamation?
The former speaker and deputy speaker for using taxpayer money for their personal benefit dodgy dan must resign immediately
The rot is so deep we need to flush out and sack all Dans mates he's placed in power a delayed report into a controversial supervised injecting room
Can I identify as Dan Andrews and commit multiple crimes making myself rich while never getting held to account
Daniel Andrews Labor must go same police that investigated Dan Andrews wifes car crash where the police report told of a male driver meaning it was Daniel
The 4 000 beds never existed dan promised and lied about them he is criminally insane

Positive, negative and compound sentiments were calculated for every tweet in the organic dataframe (Figure 4). To gain a more accurate representation of an individual users sentiments, the mean compound sentiment of all tweets from each unique user was calculated. If the mean was positive, the maximum sentiment tweet was placed in its own dataframe. If the mean was negative, the minimum sentiment tweet was taken. The distributions for unique, polarised tweets can be seen in Figure 5. Figure 6 shows the distribution for duplicate tweets, as calculated via hamming distance. Figure 7 shows a linear combination between the compound sentiment of a retweet, and the number of times it was retweeted.

Figure 8 shows a scatterplot between author description sentiment and mean tweet sentiment (Pearsons $R = 0.01$, $p = 0.28$). Table 3 shows descriptive statistics for reply, retweet and like counts, grouped by compound sentiment (positive/negative). Given the relative sparsity of datapoints in the historical period from January 1st to July 1st, the average daily compound sentiment over this period (Figure 9) was plotted separately to the July 1st to August 20 period (Figure 10). The running averages had a window of 100 and 500, respectively.

Figure 4.
Sentment distributions for all tweets.

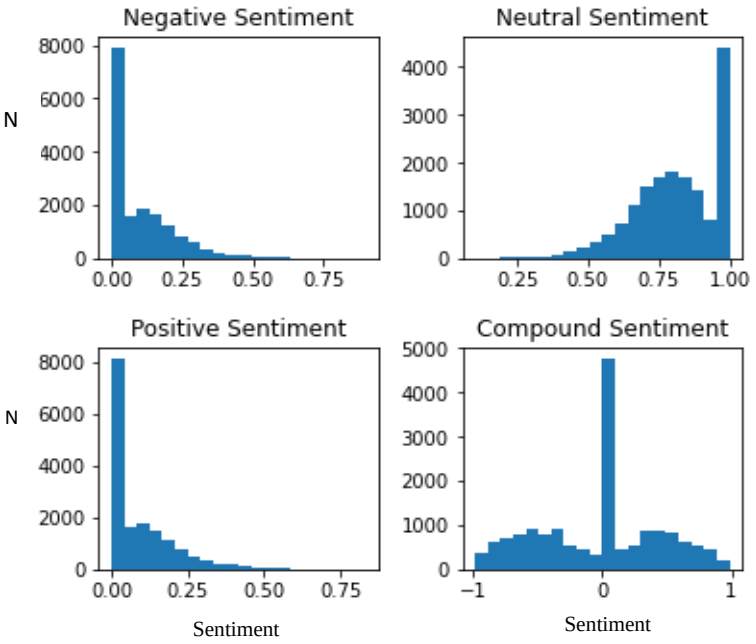


Figure 5.
Maximum/Minimum sentiment distribution for each user

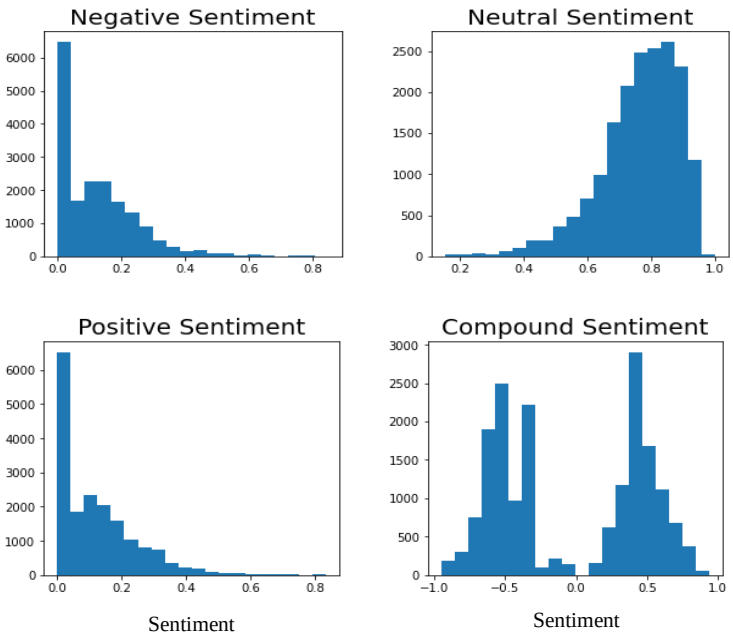


Table 4 shows main topics gleaned through topic mining, along with their number of mentions and average sentiment. Figure 11 shows word clouds for two topics found via LDA analyses.

Figure 6.
Sentiment distribution for duplicate tweets

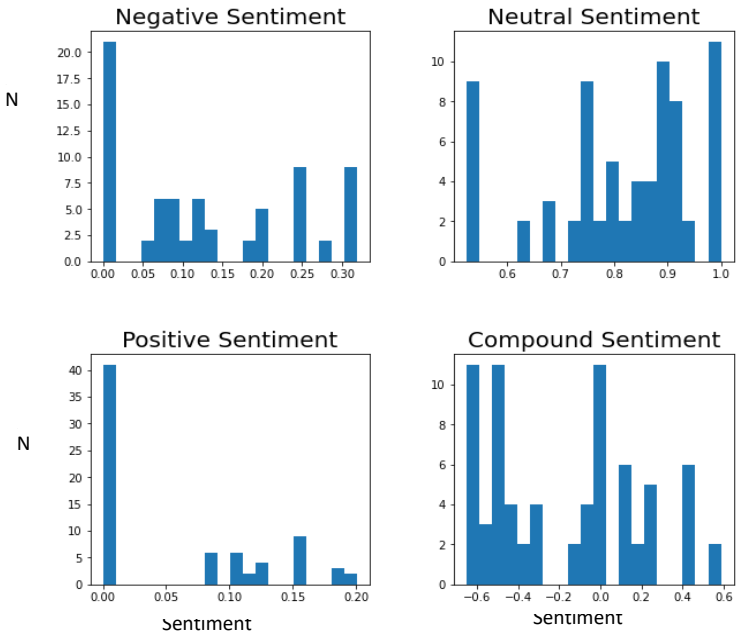


Figure 7
*Retweet sentiment popularity score (Compound sentiment * N retweets)*

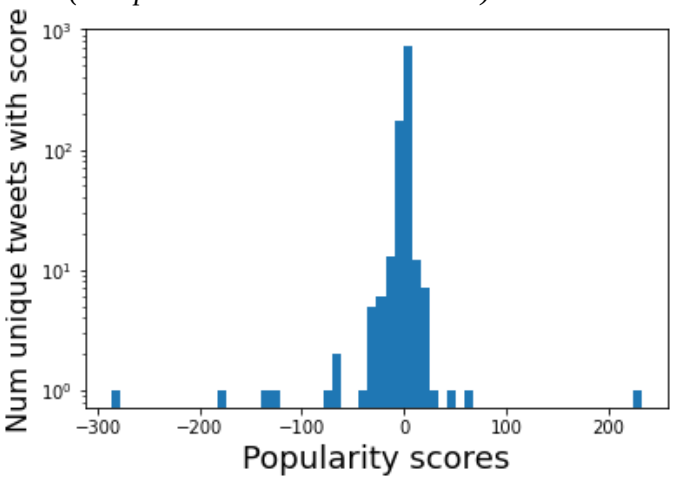


Figure 8

Scatterplot between compound sentiments for author description and mean tweet sentiment

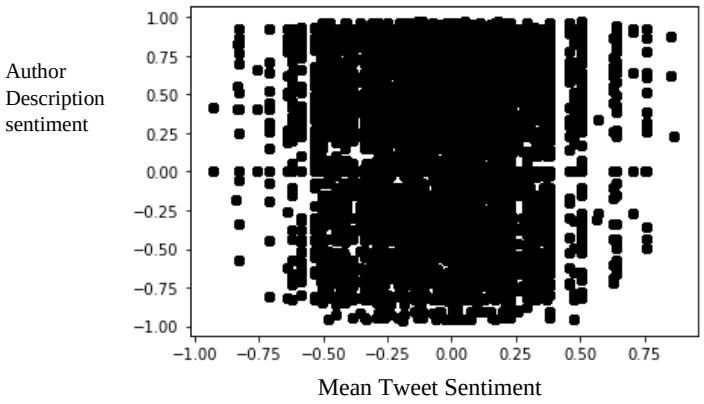


Table 3

Descriptive statistics for replies, retweets and likes count, grouped by positive or negative sentiment

	Positive			Negative		
	Replies	Retweets	Likes	Replies	Retweets	Likes
Mean	6.5	6.0	34.3	9.1	11.3	56.7
Std	34.4	43.0	224.6	44.1	89.6	450.6
Median	0.0	0.0	1.0	0.0	0.0	1.0
Max	690	985	3683	712	1974	9193

Figure 10

Compound sentiment (01/01/22 – 01/07/22)

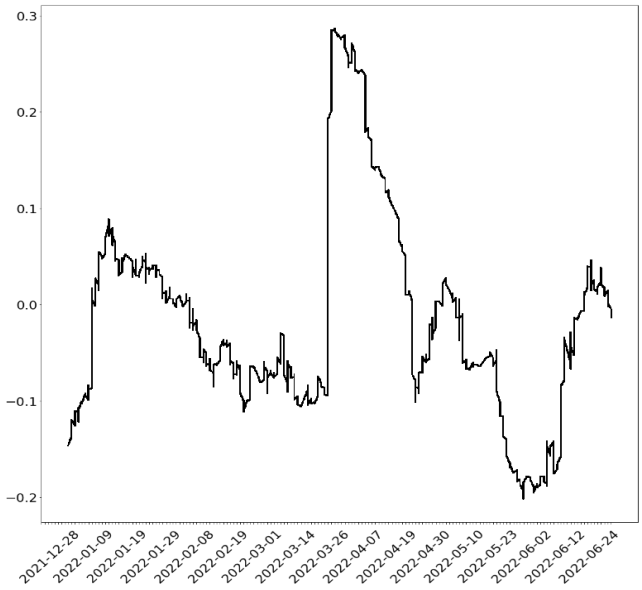


Figure 11

Compound sentiment (01/07/22 – 20/08/22)

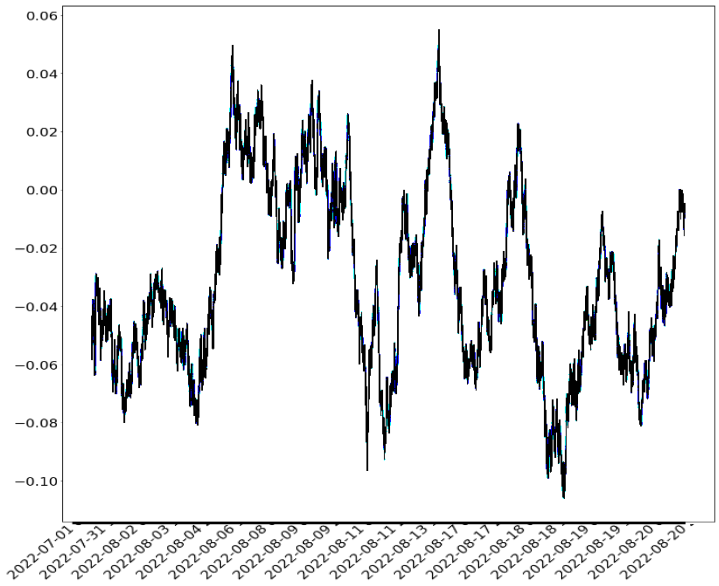


Table 4

Main topics, number of mentions, and average compound sentiment for topic

Topic	**Sluggate**	**Car crash**	**4,000 beds**	**Pandemic**	**IBAC**	**Protests**	**Election**	**Red shirts**
N	37	2	9	222	395	24	615	139
Sent.	-0.15	-0.70	0.17	-0.11	-0.06	-0.56	0.10	-0.14
Topic		**Hotel quarantine**	**East west link**	**North east link**	**Westgate tunnel**	**Lockdown**	**Ambulance**	
N		78	22	1	3	340	112	
Sent.		-0.23	-0.27	0.48	-0.14	-0.14	-0.26	

Table 5

News events by month

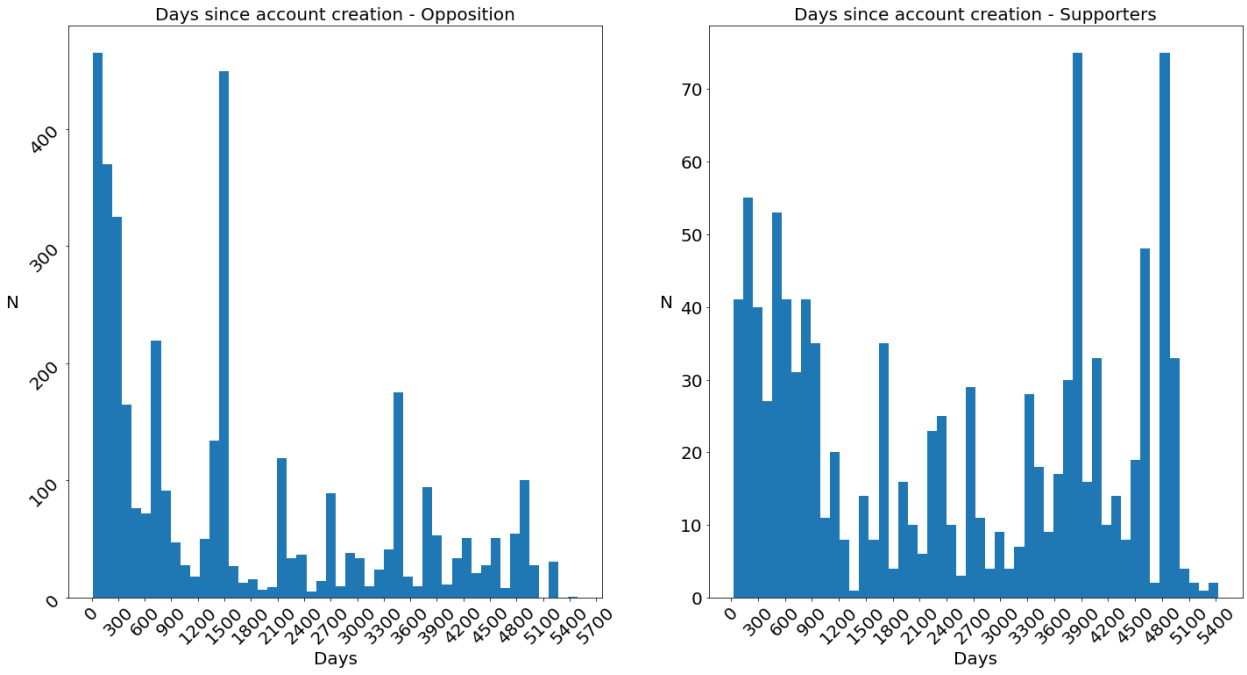
January	Novak Djokovic Saga	Pandemic Power extension	166 million rat test order
February	PRGuy lawsuit begins	Mask mandates removed	Premier bullying allegations
March	Shane Warne funeral	Premier gets covid	Tippers union strike
April	Commonwealth games confirmed	Auditor general political ad scandal	Police force staffing crisis
May	IBAC corruption hearing	Power saving bonus announced	Premiers cat dies
June	4 cabinet ministers resign	Free prep/kindergarten announced	Jacinta allen made deputy premier
July	Red shirt investigation dropped	Vicroads privatized	Ambulance crisis
August	Anti-protest bill announced	Free nursing degrees announced	Lockdown class action begins

Figure 12

Word clouds extracted via LDA analyses

Figure 13

Distribution of Days since account creation for a) opponents of the premier; b) supporters of the premier



Discussion

Whereas previous years were marked with co-ordinated campaigns from both sides [1], it is evident that the 2022 twitter conversation around the premier has comprised more of opposition groups. For example, while the two most retweeted tweets were supportive, the remaining 8 out of the top 10 were in opposition (Table 1). Out of the main catchphrases used by both sides, #SackDanAndrews and #IStandWithDan, the former dominated the latter 24.7% to 6.8% (Figure 1). The mean (102.9 characters) tweet length was closer to the old twitter limit (140 characters) then the current 280 characters (Figure 2). There is preliminary evidence linking the length of a tweet and its civility [2][3], and given the prevalence of incivility in this dataset, short tweet lengths are unsurprising.

Mentions of conspiracy theories were abundant in the dataset. Hashtags related to the lab leak theory of coronavirus (#ccpvirus, #xivirus) were in the top 20 hashtags (Figure 1). Tweets requesting for the arrest of world leaders associated with the WEF numbered 422. The majority of these, if not all, were from users opposed to the premier. The duplicate dataframe as determined via hamming distance was mostly composed of news articles, commonly republished by multiple organizations under a parent company. Mining N-grams from the data did not result in any meaningful insights. Most were from news stories. The most mentioned quadgram was ‘premier Daniel Andrews has’ with 837 mentions. ‘Dan Andrews is a’ was mentioned 113 times (the remainder of the sentence is left to the imagination).

The distribution of sentiment for all tweets was symmetric (Figure 4). When taken for unique users and their most polarized tweet (Figure 5), slight variations between the groups became apparent. Whereas positive sentiment tweets followed a normal distribution, negative tweets were positively skewed and multimodal. Whereas the positive skew could indicate higher polarisation, the multimodal distribution is harder to explain. The distribution for the linear combination of compound sentiment and N retweets was normal, with a slight positive skew (Figure 7). No correlation was found between the author description sentiment, and their mean tweet sentiment (Figure 8). One insight found from this was that author descriptions were more polarised than tweets regarding Dan Andrews,

The average compound sentiment was negative for much of the year (Figure 10, Figure 11), apart from limited periods in January, April and early August. The highest sentiment period was March/April. News events occurring at that time (See Table 5) that may be responsible for this high sentiment include the state funeral for Shane Warne, the premiers covid diagnosis, and the announcement of the commonwealth games for Victoria. The lowest sentiment occurred in July, when the privatization of Vicroads was announced, and the ambulance crisis was peaking. Negative sentiment posts attracted more

likes, replies, and retweets than positive ones. This is a known phenomenon, where negative posts can attract up to two times as much engagement as positive ones [4].

Much of the discussion centred around 5 topics: the pandemic, IBAC investigations, the upcoming election, the red shirts scandal and lockdowns (Table 4). Of these, only the election had a positive average sentiment. This may be due to a combination of tweets from news media, and those in opposition posting about their anticipation of the election. Only two topics were shown as word clouds in Figure 12. The first topic is comprised of anti-dan words, and the second of pro-dan. Further topics did not add information, and it was difficult to glean the topic they were supposed to represent.

One outcome of the previous study [1] was the insight that most accounts discussing this topic were 'sockpuppet' accounts; accounts created specifically for the purpose of posting inorganically. One of the main hallmarks of these accounts is that they are created shortly before tweeting, and contain little to no alternative tweets. In the present study, accounts in opposition to the premier were more characteristic of sockpuppet accounts. For example, the number of accounts created less than 60 days from posting the tweet in question was 365 for opponents, and only 32 for supporters. Although it can not be stated with certainty that these accounts are sockpuppets, a visual inspection of the tweets showed that 213 had characteristics of them, such as repeated tweets with slightly different wording, or just repeating hashtags with no content whatsoever (Table 6). In contrast, only 7 tweets from supporters had characteristics of being sockpuppet accounts.

Table 6

Days between account creation and tweet, the tweet itself, and the number of times variations of the same tweet was posted

Days	Tweet	N Repeats
41	take a stand vote for change in november #libdems #sackdanandrews #rememberinnovenber	4
45	all we have had from dan andrews is scandal after scandal his labor govt is rotten to the core #sackdanandrews #jaildanandrews #auspol #springst #labortrash	2
41	#sackdanandrews #sackdanandrews #sackdanandrews #rememberthisnovember	106

s

Limitations

In the preceding analyses, it became evident that many tweets had been mischaracterised, due to the high level of sarcastic tweets, and other entities being discussed (the opposition leader, chief health officer, IBAC etc). For example, one of the phrases found to be most supportive ('glorious supreme

leader') is obviously sarcastic. Additionally, one of the most negative ('Suck **** filthy shrine *******') was likely in response anti-dan protests, and thus from a supporter.

The hamming distance algorithm was effective at isolating tweets from news organizations. However, it was evident that it could not detect duplicates due to inorganic activity. Consider the following two tweets from the dataset:

- stop drinking the kool aid no one is buying what hes selling #sackdanandrews #rememberinnovember
- stop drinking the kill aid no one's buying this bs anymore #sackdanandrews #neverforget

Firstly, the strings are not the same length, and thus these tweets were skipped in duplicate checking. Secondly, the authors have changed more than 10 characters, yet retained the same message. A more effective way of implementing this filtering mechanism would be to cut the length of the longer string down to the length of the shorter string and take the hamming distance up to shorter strings length.

Although most of the analysis had been completed by this point, the word movers distance algorithm was used with a word2vec embedding to measure sentence similarity in accounts less than 60 days old at the time of the tweet, to check whether this would be a more effective method. This algorithm faced its own issues. The following two tweets were classified as the least distant:

- you need to worry more about what dictatordan and gutlessalbo is doing to labor
- important to remember the politicians who played us lets sackdanandrews in november for starters

Whereas the most distant were:

- enough of the deflecting puff posts we arent buying what your selling why won't you release the elective surgery waiting list figures #sackdanandrews (hashtag repeated 7 times)
- sackdanandrews vote vic labor last

It was not effective at filtering these inorganic tweets. It is suggested in further studies that the hamming distance with trimmed tweets be used.

Conclusion

It is evident that the discussion surrounding the premier has changed since the 2020 study [1]. Whereas the results of the aforementioned study showed more support for the premier than opposition, the current study has found the opposite. Support for the premier is sparse, and opposition is prevalent. There is evidence of an increase in inauthentic activity on the opposition side however.

Much has changed since the previous study. It is difficult to determine how much these interim events affected twitter sentiment regarding the premier. However, it is known to be much more difficult for

leaders to gain and maintain popularity, than to lose it. Some decrease in popularity over time is to be expected.

Whereas in 2020 the premiers supporters approved of pandemic measures, now that they have been removed, many supporters have become disillusioned, using hashtags such as #IStoodWithDan. Groups in opposition have continued their opposition. The implications of this for the upcoming 2022 Victorian election are unknown, as political polling suggests that the premier will be re-elected. However, it is evident that the highly polarised debate regarding the premier will likely continue.

References

[1] Graham, T., Bruns, A., Angus, D., Hurcombe, E. and Hames, S., 2020. *#IStandWithDan versus #DictatorDan: the polarised dynamics of Twitter discussions about Victoria's COVID-19 restrictions*. [online] NCBI. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7754160/>.

[2] Jaidka, K., Zhou, A. and Lelkes, Y., 2019. *Brevity is the Soul of Twitter: The Constraint Affordance and Political Discussion*. [online] Available at: <https://academic.oup.com/joc/article/69/4/345/5547032?guestAccessKey=54a38170-de3f-4437-bbdc-cb3aa1ba1b5e&login=false>.

[3] Boot, A., Sang, E., Dijkstra, K. and Zwaan, R., 2019. *How character limit affects language usage in tweets*. [online] Available at: <https://www.nature.com/articles/s41599-019-0280-3>.

[4] Rathje, S., Van Bavel, J. and van der Linden, S., 2021. *Out-group animosity drives engagement on social media*. [online] PNAS. Available at: <https://www.pnas.org/doi/10.1073/pnas.2024292118>.