Can Twitter Data Support or Contradict Theories about why Labor Lost Their Unloseable Election?

Angharad Dare-Edwards - 45633801
Swaantje Grunefeld - 43430163
Lachlan Gibson - 42670908
Jiayu Zhou - 45478776

The University of Queensland

8 June 2019

Can Twitter Data Support or Contradict Theories about why Labor Lost Their Unloseable Election?

Abstract

The Australian Labor Party (Labor) lost the 2019 Australian federal election to the Liberal National Party (LNP) contrary to the expectations of many. As a result, many journalists, politicians and polling organisations have been attempting to explain the election outcome. Our project explores the possibility of using Twitter data to support or reject some of their popular theories. We found that although the data could provide useful commentary on each theory, we could not obtain any 'silver bullet' that could emphatically explain the outcome alone. Rather, the Twitter data seemed to give some support to multiple theories, which probably cooperatively culminated in the LNP victory. The data suggested that the biggest election topics were about climate change, finances and jobs. The LNP avoided promoting their own policies and leader, and aimed to smear Labor's policies as financially risky. This strategy appeared to be particularly useful in more rural and mining relient regions like Queensland where people would be more directly affected by Labor's climate policies. Furthermore, the LNP avoided drawing negative attention from more apathetic voters on the issue by avoiding mentioning climate change directly.

Can Twitter Support or Contradict Theories about why Labor Lost Their Unloseable Election?

Introduction

The Australian federal election was held on 18th May 2019. In the lead up to the election, the polls were clear: the Liberal National Party (LNP) were fighting a losing battle. Two party preferred polls consistently showed the Australian Labor Party (ALP or Labor) to be ahead of LNP, as shown in Figure 1. Newspaper articles had headlines such as "majority of voters think Bill Shorten will be the winner on Saturday"[1]. In fact, the expected outcome of this election was so certain, that the betting agency Sportsbet made the bold decision to pay out $5.2m to Labor punters 2 days ahead of the election[2]. In a shock result, the LNP won, Labor lost their unloseable election, and political analysts scrambled to decipher what happened. Various theories emerged about both why Labor lost, and also why politicians, punters, polls and the public mis-predicted. These theories are covered below.

The social media platform Twitter is strategically used by political candidates around the world, and elections are popular topics for Twitter users[3]. Twitter data has been used for many election-related studies, including for topic classification, election prediction[4], studies of political polarisation and for its role in political engagement[5]. Given the huge amount of data available on Twitter regarding the 2019 Australian Federal Election, we decided to mine this data to see if it could support or contradict the various theories, and thus demonstrate an additional use for Twitter data in the sphere of public elections.

---

[1] "Essential poll: majority of voters think Bill Shorten will be the winner on ...." 16 May. 2019, https://www.theguardian.com/australia-news/2019/may/16/essential-poll-majority-of-voters-think-bill-shorten-will-be-the-winner-on-saturday. Accessed 1 Jun. 2019.

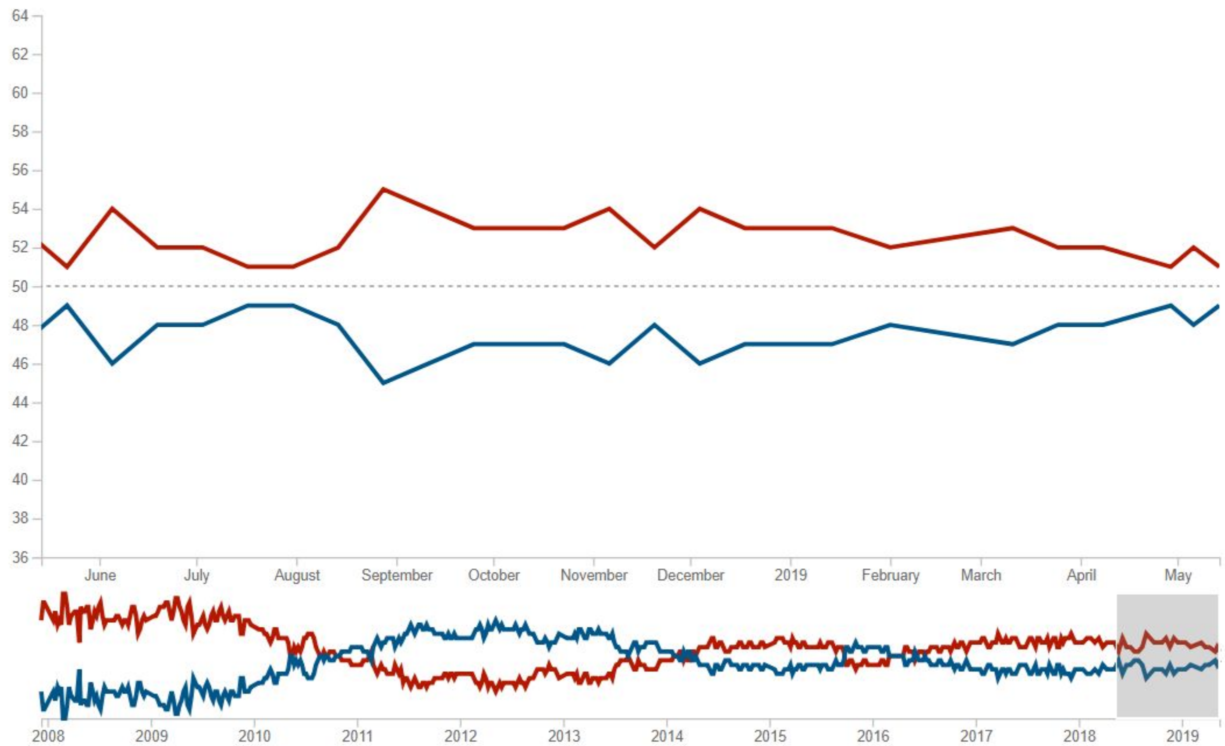
[2] "Betting agency's costly election stuff up - News.com.au." 19 May. 2019, https://www.news.com.au/sport/sports-life/betting-agency-makes-monumental-and-costly-error-ahead-of-election/news-story/a9a49b6aef6d2b03e9a8d0841a376d81. Accessed 2 Jun. 2019.

[3] Dhiraj Murthy (2015) Twitter and elections: are tweets, predictive, reactive, or a form of buzz?, Information, Communication & Society, 18:7, 816-831, DOI: 10.1080/1369118X.2015.1006659

[4] Dwi Prasetyo, N. & Hauff, C. (2015). Twitter-based Election Prediction in the Developing World. *Proceedings of the 26th ACM Conference on Hypertext & Social Media* (pp. 149-158). New York, ACM. DOI: 10.1145/2700171.2791033

[5] Dhiraj Murthy (2015) Twitter and elections: are tweets, predictive, reactive, or a form of buzz?, Information, Communication & Society, 18:7, 816-831, DOI: 10.1080/1369118X.2015.1006659

Figure 1: Poll conducted by the Guardian showing the two party preferred vote[6]. Labor (red) has maintained greater percentage than LNP (blue) for more than a year.



Theories

Some of the theories that have been presented to explain Labor's loss are as follows:

1. ***"Cult of Personality" or the "Morrison Tactic".*** Shorten was less popular than Morrison, and LNP focused on promoting their leader rather than the party as a whole. Though the Labor party consistently scored higher in polls than the LNP, polls for the preferred prime minister show Morrison to be in the lead in the months prior to the election[7].
2. **"*Attack Ads Triumph Over Policy Detail".*** Labor came to this election with very bold, progressive policies, whereas the Liberals provided very little policy. The theory is that the Liberals distilled the contest to a referendum on Labor's ambitious plans and focussed on discrediting their policies.
3. ***"Climate Election".*** Climate change was supposedly one of the most important issues of the election, with more than 80% of Australians wanting more action on climate change, according to

[6] "The Guardian Essential Report, 16 May results | Australia news | The ...." https://www.theguardian.com/australia-news/ng-interactive/2019/may/16/the-guardian-essential-report-16-may-results. Accessed 1 Jun. 2019.
[7] "The Guardian Essential Report, 16 May results | Australia news | The ...." https://www.theguardian.com/australia-news/ng-interactive/2019/may/16/the-guardian-essential-report-16-may-results. Accessed 3 Jun. 2019.

ABC's Vote Compass[8]. The election was often referred to as the "climate election". Labor's campaign heavily featured their stance on climate change, with the range of actions they would take on climate change. Now it is theorised that Labor's focus on climate actually led to their downfall, with voters more interested in financial stability[9].

4. **"Quexit"**. The election results showed a huge swing against Labor in Queensland, contributing significantly to their overall loss. Explanations for this swing include the "rural vs city divide" theory—that rural voters are more aligned to LNP than to Labor—as well as lower sentiment towards Bill Shorten and a greater than average concern about jobs. "They saw him as a southern trade unionist with a green agenda" Dr Paul Williams, a political scientist from Queensland's Griffith University, told SBS News. None of those things appeal to regional Queenslanders who are keen to get the extractive industries moving to generate blue collar jobs."[10])

Key Findings

In this section we describe our key findings against each theory, and finish with our own theory and conclusion.

1. ***"Cult of Personality" or the "Morrison Tactic"***
   Analysis of the Twitter data collected did not reveal any evidence of this tactic and thus the success of it. In fact, our analysis showed tweets with strong positive sentiment tended to discuss Bill Shorten more than Scott Morrison.

2. ***"Attack Ads Triumph Over Policy Detail"***
   Our analysis revealed evidence of the LNP tweets focussing on attacking/discrediting Labor's policies. This was demonstrated in the use of wordclouds and in analysing statistics on the number of mentions of certain key words along with the word "Labor".

3. ***"Climate Election"***
   Analysis on word frequency counts revealed "climate" to be the top topic tweeted by the public. Labor's tweets on 'climate' showed a strong correlation with the public's (in terms of percentage of tweets each day), whereas LNP's did not. We did not however find any evidence that Labor's focus on climate led to their downfall, but it is clear that despite the strong focus on climate and Labor's alignment with the public on this topic, it was not enough to win them the election.

---

[8] (2019, May 14). Federal election 2019: Vote Compass finds broad desire for ... - ABC. Retrieved June 3, 2019, from https://www.abc.net.au/news/2019-05-15/federal-election-vote-compass-climate-change/11110912

[9] (2019, May 19). Election 2019: What happened to the climate change vote we ... - ABC. Retrieved June 3, 2019, from https://www.abc.net.au/news/2019-05-20/what-happened-to-the-climate-change-vote/11128128

[10] "It was the election they couldn't lose, so what went so wrong for Labor?." 19 May. 2019, https://www.sbs.com.au/news/it-was-the-election-they-couldn-t-lose-so-what-went-so-wrong-for-labor. Accessed 2 Jun. 2019.

4. ***"Quexit"***
   Contrary to what the theory would lead one to expect, our analysis on the Queensland subset of public tweets showed "Bill Shorten" featuring more heavily in positive sentiment tweets than Scott Morrison. This is the same finding as with the full public dataset.
   The data did reveal slightly lower interest in the 'climate' topic in Queensland than for the full public dataset, and higher discussion on "Adani". Adani was a controversial topic in the election, as the opening of a new Adani mine would be a step backwards for climate change, but would bring many jobs to Queensland. This finding could be used to help support the theory Queenslanders voted for jobs over climate.

Overall, it is unlikely any one theory is the reason LNP won and Labor lost. Some of our analysis supports and some contradicts elements within each theory. Our analysis revealed that LNP tweet far less than the other two main parties, especially on controversial topics such as climate change, or even their own candidates. We believe this supports a theory that LNP won by avoiding controversial topics.

Getting the Data We Need

What Data do we Need?
To answer the problem statement, we needed historical tweets from both the general public and politicians. See Figure 2 for an example tweet.

Figure 2: A "tweet" on Twitter has a variety of data types, including the tweet text, often an image, links, and other metadata.

In addition to knowing what was tweeted (the "text"), we required key metadata including: who tweeted ("screenname") and the timestamp ("created_at"). Each tweet has a long list of other metadata, most of which was not relevant to the problem statement. A few additional attributes were selected, which were hoped to be useful.

To get the tweets from the politicians, we needed to identify which Twitter accounts were "politician" accounts. To do this, we required the screennames of all election candidates, plus the screennames of the main party accounts. We decided to limit the politician accounts to just those from the 3 main parties (LNP, Labor and The Greens), as they were most relevant for the problem.

Sources of Data

The politician screennames were sourced from each of the party websites[11] [12] [13]. The tweets were sourced from the official Twitter API, using private developer accounts. This will be discussed further in the following section.

Data Ingestion

Ingestion Parameters

We required only historical tweets (rather than real-time), which could be achieved through a "pull" on the Twitter API, using the Twitter Standard API Search and Tweet Timelines.

The standard API search only returns tweets for the last 7-9 days, so we started collecting data early, then ingested additional data every few days.

Politician Screenname Scraping

Before downloading the tweets, we first needed to collect the Twitter accounts of all candidates for the LNP, Labor and Greens party. This was done by scraping the official party websites[14][15][16] using the python package scrapy. The LNP Twitter accounts were more difficult to collect; they do not have an official website listing the candidates from both the liberal and national party. Due to the dynamic page scrolling on the Liberals webpage, it was not possible to implement the simple python code used for scraping Labor and Greens websites. Instead, the web address for each liberal candidate was built from the official list of candidates provided by the AEC[17] and the social media accounts scraped. Finally, the Nationals party website[18] did not provide the Twitter accounts of their candidates, so these had to be found manually. In addition, each party's official Twitter account was manually collected.

An example of the code used to get the screennames is shown in Figure 3.

---

[11] "Our Federal Candidates | The Australian Greens." https://greens.org.au/candidates. Accessed 4 Jun. 2019.

[12] "Our People - Australian Labor Party." https://www.alp.org.au/our-people/senators/. Accessed 4 Jun. 2019.

[13] "Liberal Party of Australia." https://www.liberal.org.au/. Accessed 4 Jun. 2019.

[14] Our People - Australian Labor Party. Retrieved May 6, 2019, from https://www.alp.org.au/our-people/senators/

[15] Our Federal Candidates | The Australian Greens. Retrieved May 6, 2019, from https://greens.org.au/candidates

[16] Our Team | Liberal Party of Australia. Retrieved May 6, 2019, from https://www.liberal.org.au/our-team

[17] (2019, April 24). House of Representatives and Senate candidates - Australian .... Retrieved May 6, 2019, from https://www.aec.gov.au/election/candidates.htm

[18] The Nationals - Meet Our Team. Retrieved May 6, 2019, from http://nationals.org.au/our-team/

Figure 3: A sample of the code used to scrape candidate Twitter screennames.

```
# -*- coding: utf-8 -*-
# install scrapy with pip install scrapy
#run using scrapy runspider labor1.py
import scrapy
import logging
logging.getLogger('scrapy').setLevel(logging.WARNING)

class spider1(scrapy.Spider):
    name = 'labourcandidates'
    start_urls = ['https://www.alp.org.au/our-people/our-people/']

    def parse(self, response):
        NAME_SELECTOR = 'div.ml-card__title::text'
        LINK_SELECTOR = 'a.ml-card__link'
        for i in response.css(LINK_SELECTOR):
            str1 = 'https://www.alp.org.au'
            str2 = i.css('::attr(href)').extract()[0]
            print('"'+str1+str2+'",')
            #print(response.css(LINK_SELECTOR).css('::attr(href)').extract())
```

Figure 4: An example of the output, for the Labor candidates.

```
http://twitter.com/alifrance5
http://twitter.com/chrisgambian
http://twitter.com/emma4dobell
http://www.twitter.com/@lukejgosling
http://www.twitter.com/cowperlabor
http://www.twitter.com/julianhillmp
http://www.twitter.com/kimbakit
http://www.twitter.com/senraffciccone
https://twitter.com/AlboMP
```

Accessing the Twitter API

To access the tweet data, we signed up for developer accounts with Twitter, and accessed the official Twitter API using the Python package "tweepy". It is necessary to keep developer account credentials private; to ensure this, each team member had their own developer account and the credentials were imported to the main python file.

Politician Tweets

Once the screennames were collected, we used Python and tweepy to download the tweets from each politician's timeline back to 1st March 2019. Although the official Twitter API rate-limits prevent us from downloading *all* tweets from an account, we *can* download 3200 tweets from each account. This is sufficient for our purposes, as no politician had more than 3200 tweets since 1st March.

The table below gives an overview of the number of candidates, the number of candidates with Twitter accounts and the total number of tweets made by each party since 1st March. Evidently there are similar numbers of candidates in each party, but the actual number of accounts and tweets differ wildly. ALP have by far the highest number of accounts and the greatest number of tweets. Though there are approximately half as many Greens candidates on Twitter than LNP, they have almost double the number of tweets.

Figure 5: Number of candidates, candidate Twitter accounts, and tweets per party.

Party	# Candidates	# candidate Twitter accounts	# tweets (since 1st March)
ALP	188	102	31582
Greens	186	29	12586
LNP	196 (169 Liberals, 27 Nationals)	69	8689

Public Tweets

We wished to access tweets from the "general public" who were tweeting on election topics. Without having specific screennames, we needed to download these tweets from a search of the API. This necessitated having specific queries/search terms on which to run the search. We decided to use a short list of specific election-related hashtags and keywords. It is important to note that a Twitter user is not required to include any specific keywords or hashtags when tweeting on a particular subject, so our dataset only includes those who chose to use these terms. We do not believe this introduces any significant bias.

The hashtags chosen were party-neutral, and featured in the trending hashtags for Australia during the election campaign.

Figure 6: Search terms used for the Twitter API Search to access public tweets on general election topics

Election	auspol	AustraliaVotes	AusVotes19
AusVotes2019	AusVotes	AustraliaDecides	

With the searchterms often quite general, for example "election", and with the fact any Twitter user (Australian voter or not) is free to use any hashtag they wish (for example, a journalist in Great Britain could tweet and use the hashtag #auspol); it was necessary to limit our search to Australian tweets. This was achieved by specifying a geocode, which specifies a longitude and latitude, plus a radius sufficiently large to include the Australian mainland and Tasmania.The location is preferentially taken from the Geotagging API, but will fall back to the user's Twitter profile. The search was also filtered for english language tweets (english detection is "best effort").

The attributes selected for each tweet are listed and described in Figure 6.

Figure 6: Table of attributes selected for each tweet

Attribute	Description
created_at	The date and time of the tweet, in GMT.
id_str	This is the unique id of the tweet. We refer to this as the "tweet_id".
screen_name	The user's account name
retweet	yes/no whether the tweet is a retweet or not. This attribute was used to ensure the full text of any retweet was downloaded.
tweettext	The full text of the tweet/retweet.
retweeted	Whether this this tweet was itself retweeted.
user.location	The user's location as per their profile.

The Twitter standard API search is rate-limited, meaning there is a limit on the number of requests that can be sent per 15 minutes. The python code for the data ingestion thus took this into account, pausing for 15 minutes after the limit was reached.

On average the search returned about 30,000 tweets per day, taking about 3.6 hours to download each day's tweets.
An example of the code used to pull this dataset is provided in Appendix A.

Exploratory analysis on the above dataset, as well as analysis of Google trends, revealed that climate change was a key topic in the election. An additional dataset of public tweets was downloaded using climate-change-related hashtags and keywords.

Figure 7: Search terms used for the Twitter API Search to access public tweets on climate change topics

ActOnClimate	ClimateElection	GlobalWarming	ClimateDisaster
ClimateEmergency	Carbon Tax	Adani	Environment
Climate	ClimateCrisis	Environment Policy	

Is our Data Fit for Use?

Data Exploration

Exploratory data analysis (EDA) was performed on the various datasets to gain an understanding of the quality of the data. This analysis led to:

i) improved data collection (i.e. re-ingesting the data after rewriting search queries)
ii) further data collection (for example after discovering gaps in our data due to ingestion errors)
iii) data transformation (discussed below)
iv) abandoned analysis (e.g. some data not deemed fit for use)

EDA included the following techniques:

Word Clouds
In addition to eyeballing the raw data, we used the Python package "WordCloud" to create wordclouds on the tweets, hashtags and user location fields of our data. This helped gain an understanding of the quality and content of these fields. For example, we explored the user_location field to confirm that the filtering of the location to around Australia worked. From eyeballing the user_location field in our data, it was clear that this was a free-text field in a Twitter user's profile and thus included many variations for the same/similar location (e.g. New South Wales vs NSW) as well as many nulls plus nonsense locations (e.g. "Left of Centre"). Performing a wordcloud on this field helped to verify that the most frequent locations were indeed in Australia, and useable.

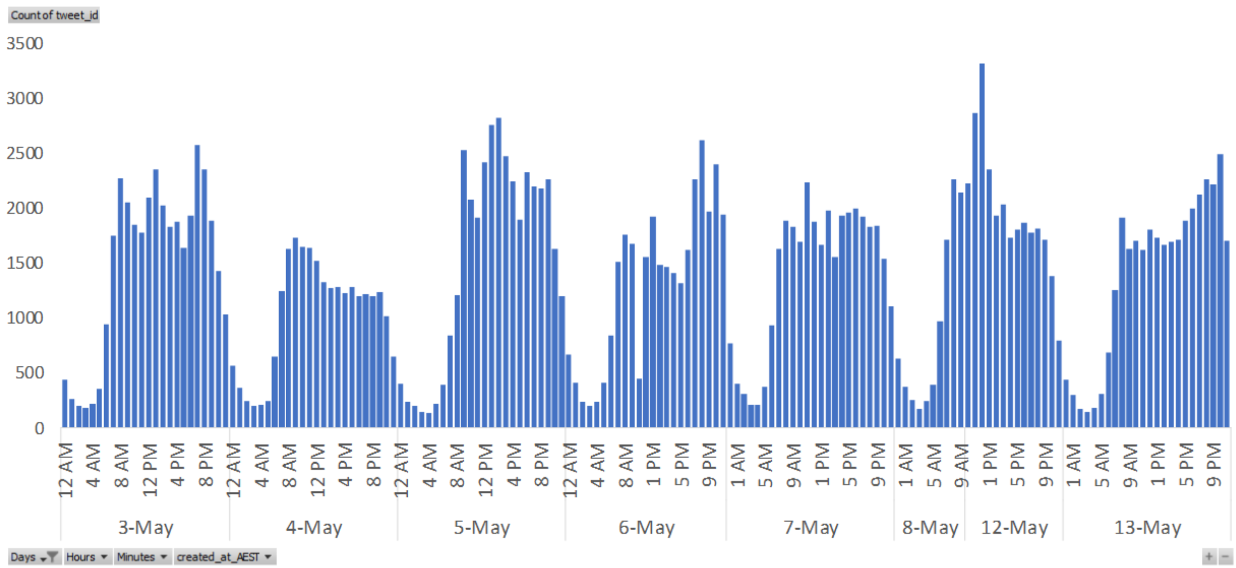
Figure 8: Example of EDA WordCloud. This wordcloud was formed from a sample of the public tweets

Timeseries Plots of Tweet Counts
Twitter's API documentation states that their search service, and thus search API is not an exhaustive source of all tweets[19]. By graphing the number of tweets by hour by day, we were able to verify that despite our dataset being a sample of tweets matching our search query, the daily pattern of tweet counts made sense for Australian Eastern Standard Time and looked like a representative sample. These graphs

[19] "Standard search API — Twitter Developers - Developer — Twitter." https://developer.Twitter.com/en/docs/tweets/search/api-reference/get-search-tweets.html. Accessed 3 Jun. 2019.

also helped us to identify visually where we had gaps in our dataset due to data ingestion errors, allowing us to fix/fill these gaps before conducting our detailed analyses.

Figure 9: Plot of tweet count by hour by day for a sample of the public 'election' dataset.



Word Frequency Counts

Examining the frequency and sentiments of tweets related to key election topics could be a useful way to quantify how connected or disconnected each party was from the public. In order to extract these topics from the public and politicians, R was used to count word frequencies in the "text" field of the tweets. In doing this, it was necessary to also "clean" the unstructured text to remove common English stopwords, emoticons, user tags, hash tags, links and URLs. Looking at the top 20 most frequently tweeted words by the public reveals that the top two topical words are "climate" and "tax", appearing 32,453 and 31,266 times respectively. The remainder of these top words relate to the parties, their leaders and general election related words.

Figure 10: Top 20 most frequently tweeted words by the public and major parties

Public		LNP		ALP		Greens	
election	climate	will	million	labor	change	climate	need
will	tax	new	people	will	australia	will	plan
labor	party	labor	support	government	vote	people	australia
morrison	australia	great	plan	today	election	greens	like
vote	campaign	government	community	liberals	labors	election	time
just	liberal	today	labors	morrison	plan	can	get
shorten	scott	tax	funding	one	great	vote	now
people	government	australians	taxes	people	time	just	today
bill	now	australia	bill	cuts	scott	change	coal
one	can	economy	can	tax	australians	one	labor

Top 20 most frequently tweeted words by the public and major parties in descending order. Words about tax and finances are shaded orange while words about climate are shaded green.

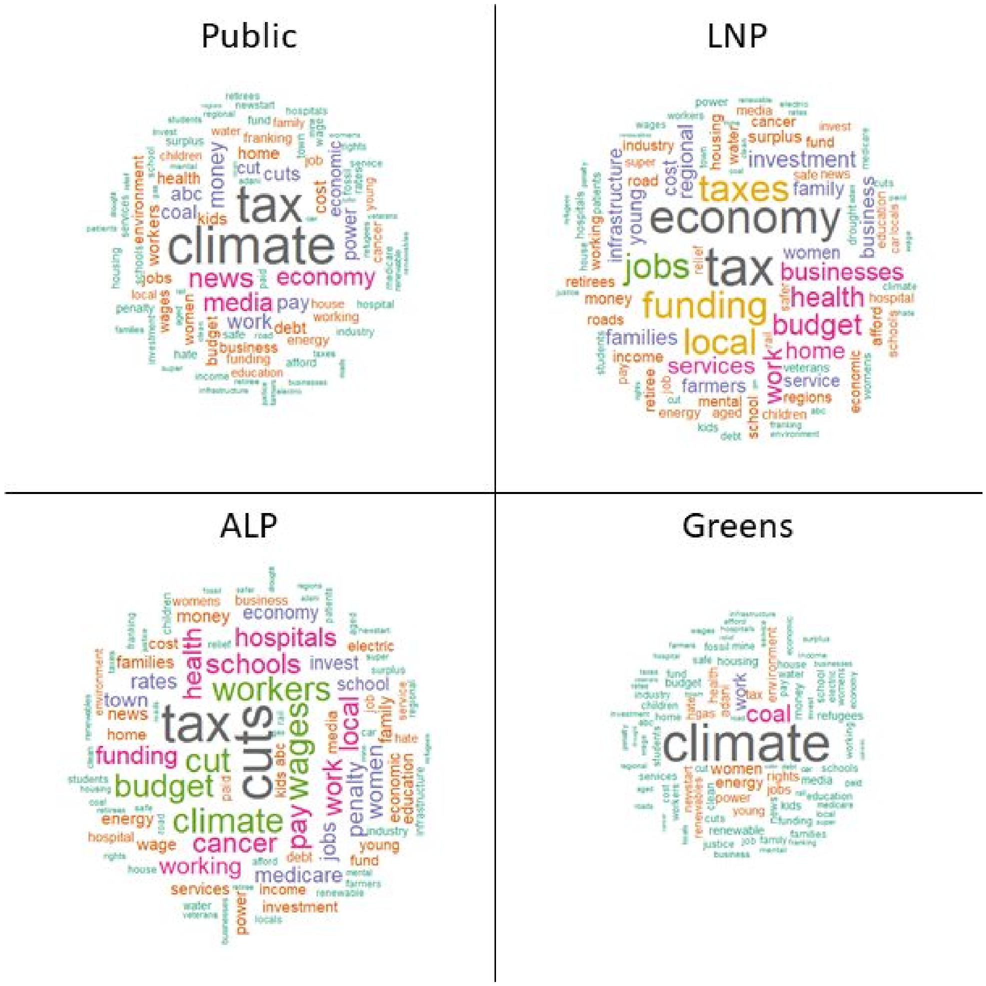
Further key topical words were identified by manually checking the list of most frequent words that appeared at least 0.01 times per tweet on average. Most of the frequently tweeted topical words are either environmental or financial and the word clouds below illustrate their relative frequencies. These topical keywords were identified as the key topics in the election, and this was confirmed by analysis on Google Trends (see Appendix B). We used these keywords to identify topics covered by each tweet, discussed in Data Enrichment below.

Figure 11: Top 10 most frequently tweeted topical words by the public and major parties

Public		LNP		ALP		Greens	
climate	0.063	tax	0.084	cuts	0.064	climate	0.124
tax	0.060	economy	0.074	tax	0.064	coal	0.051
news	0.030	funding	0.057	climate	0.039	work	0.031
media	0.027	taxes	0.057	workers	0.038	women	0.027
economy	0.025	local	0.054	cut	0.038	energy	0.024
money	0.021	jobs	0.052	wages	0.038	rights	0.021
work	0.021	budget	0.041	budget	0.036	environment	0.020
pay	0.019	health	0.041	schools	0.031	tax	0.019
cuts	0.018	work	0.036	cancer	0.031	adani	0.019
abc	0.018	businesses	0.035	pay	0.030	jobs	0.019

Top 10 most frequently tweeted topical words in descending order by the public and major parties, along with average frequency per tweet.

Figure 12: Word Clouds to illustrate the relative frequency of topical words in each dataset



Data Transformation and Enrichment

Various transformations and enrichments were performed on the data to make it fit for use. These transformations are discussed below.

Converting dates to AEST

The created_at field in our data gives a time-date stamp for each tweet. This was downloaded in the default Greenwich Meantime (GMT). An additional column was added to the dataset when necessary to convert this time/date to Australian Eastern Standard Time (AEST).

Adding factors to identify whether a tweet covered certain identified top election topics

After distilling the key topics (discussed in the section on EDA above), additional factor columns were added to the datasets to identify whether or not a tweet covered a certain topic.

This was a “blunt tool” approach, for example not allowing for topic word synonyms, however was felt to be sufficient to provide insight on the relative frequencies and sentiment (discussed later) of each topic.

Creating a subset of data with valid location information

Location information is an important part in learning the Twitter user, although there are many missing data (MNAR) in the user location and many records come from defaulted user profile, which means it is hard to acquire accurate imputation. Bad location example is room 101 or northern Australia. However, location information still be contained in our datasets because it is meaningful to figure out the distribution of sentiment value in eight regions(Australian Capital Territory, New South Wales, Victoria, Queensland, Western Australia, South Australia, Northern Territory, Tasmania). According to the rank of population, 100 cities has been selected to classify the location in our datasets. After filtering the valid location information, a large amount of data was removed, for example, in the “politics sentiment” dataset, we have 519132 data before, only 113263 data was left for location analysis. This step was implemented in R and Excel.

Adding Sentiment Scores

In order to determine the opinions of the public and politicians tweeting about election topics and candidates, we decided to perform sentiment analysis on the tweets we collected. Sentiment analysis is a quantitative way of determining how people are reacting to certain topics. Accurate sentiment analysis can be done with machine learning techniques or by comparing against a sentiment lexicon.

We considered using Google’s Natural Language API[20], but this was foiled by the limited number of documents we could analyse without payment. Instead we chose to use the TextBlob Python library[21] which performs basic sentiment analysis using a sentiment lexicon. This lexicon was trained on movie reviews, and so is not expected to be highly accurate for our tweets.

Some examples of tweets with negative, neutral and positive sentiment are shown below. The sentiment analysis cannot pick up on more complex sentiment such as sarcasm, which is perhaps the reason the third neutral example tweet does not have a more negative sentiment score.

[20] (2018, November 5). Natural Language API Basics - Google Cloud. Retrieved May 12, 2019, from https://cloud.google.com/natural-language/docs/basics

[21] Tutorial: Quickstart — TextBlob 0.15.2 documentation. Retrieved May 12, 2019, from https://textblob.readthedocs.io/en/dev/quickstart.html

Very Negative (Sentiment = -1)	“Day 2069 of the perpetual election campaign we've had since this terrible Government took office. What will Morrison blame @billshortenmp for today? Does anyone even care? #AusVotes2019”
	“She’s THE worst Environment Minister EVER! (I know I’m staying the bleeding obvious...) #LibFail #ClimateEmergency #auspol https://t.co/4l0vLY5DNU”
	“It is outrageous for Josh Frydenberg to argue has a high company tax rate when WE ALL KNOW that our richest companies are not paying tax at all. What a con. What a rort. #ausvotes #insiders”

Neutral (Sentiment = 0)	“Scott Morrison has named health and education as two of his priorities if he holds power at the federal election #auspol https://t.co/moFS1v6bQR”
	“Pls, next election can it just be told through the medium of dogs? thx https://t.co/4vkYc2Senj”
	“Well the LNP want to modernise Australia’s economy by selling coal to the poorest countries that can’t afford to modernise their energy grid. Oh that will be Australia soon if we have another 3 years of this government. #auspol #AUSVote2019 #AusLibChat https://t.co/jJ3EyxoA1k”

Very Positive (Sentiment = +1)	“The people of McEwen know it’s time to end the cuts and chaos, restore fairness and opportunity and elect a Shorten Labor Government. With @RobMitchellMP - he’s here for us! #AusVotes2019 https://t.co/17eWI6ZsuL”
	“This is the Climate Election, and The Greens have the best policies for the climate. Check out our scorecard. #vote1greens #afutureforallofus #climateelection https://t.co/ZcO0PQ19J4”
	“It would be exquisite if @billshortenmp turns out to be the Prime Minister we all thought Malcolm Turnbull could have been. #qanda #auspol #ausvotes”

The advantage of applying sentiment analysis is that we could then investigate which topics were mentioned most often with a positive and negative sentiment. This could be done through frequency counts and word clouds, and could be tracked over time.

Data Quality

Social media[22], including Twitter[23], has been shown to be a good source of data to identify public opinion and make election predictions. The Twitter data was of a fairly high standard, being scraped using the official Twitter API and carefully designed queries. However, there are some limitations and data quality issues about how accurately the Twitter data represents the real opinions of Australians. It is worth discussing these issues and how they might affect the data's usefulness in commentating on the election theories.

Using Tweets as a Proxy for Public and Party Opinion

The official Twitter API gave us fresh and reliable access to their indexed tweets, including the tweet text, geo coordinates, username and user location. However, not all Australians use Twitter, and not all Twitter users Tweet. The subgroup of people who tweet might not form a representative sample of all voting Australians. Furthermore, information that could verify the validity of the sample was often missing. For example, most tweets did not include geographical coordinates to pinpoint their precise location, and the user location was a free form text field that, when filled, was not consistently formatted and often inaccurate. Despite these limitations, we believe that there are sufficiently diverse users and tweets from the public to still provide valuable information about the popularity of election topics and political figures. Furthermore, the tweets from the politicians and official party accounts were fairly exhaustive and are very likely targeted to more than just Twitter users since other media sources often report these tweets to the wider public. This makes Twitter a good platform for the parties to campaign and, therefore, a good source of data to analyse their strategies. This means that our dataset should be especially informative in validating election theories based on party strategies.

Sentiment Calculation

Inaccuracy and imprecision of tweet sentiment scores are potential data quality issues. Inaccurate sentiment scores could misrepresent the opinion of the public or political parties, while imprecise sentiment scores would make it difficult to draw any conclusions from aggregated results.

It is possible to use bounds on standard errors to estimate the number of tweets required to find statistically significant results when calculating means. The sentiment scores are bounded by -1 and 1 so the Popoviciu's inequality limits the variance of a group of sentiment scores to be less than or equal to 1. Therefore the standard error of the mean sentiment of a group of tweets is bounded by $1/\sqrt{n}$ where n is the number of tweets in the group. Without assuming independence, the standard error when comparing the difference between the mean sentiments of two groups is also bounded by $2/\sqrt{n}$ where n is the number of tweets in the smaller group. These bounds on the standard errors also give values for n that guarantee sufficient precision for statistically significant results. For example, if the difference in mean sentiment between two groups of tweets is 0.1 then having $n \geq 400$ guarantees that the standard error is

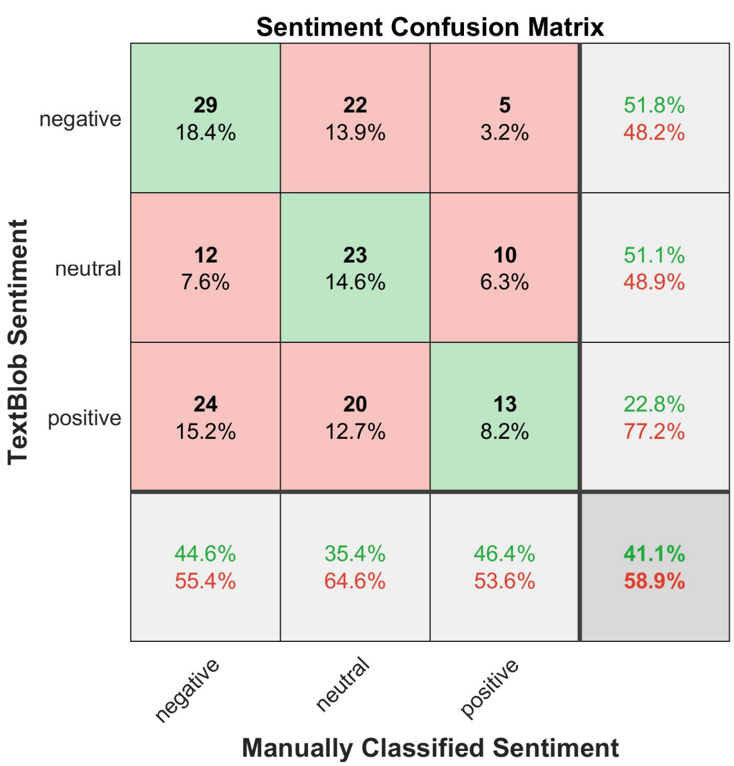
[22] (2019, May 19). Election 2019: How the polls got it so wrong in predicting a ... - ABC. Retrieved June 4, 2019, from https://www.abc.net.au/news/2019-05-19/federal-election-results-how-the-polls-got-it-so-wrong/11128176
[23] Dwi Prasetyo, N. & Hauff, C. (2015). Twitter-based Election Prediction in the Developing World. *Proceedings of the 26th ACM Conference on Hypertext & Social Media* (pp. 149-158). New York, ACM. DOI: 10.1145/2700171.2791033

less than the difference between the two groups. Similarly, having $n \geq 1600$ would guarantee a standard error less than 0.05. Note that these values are just bounds and, in practice, even smaller values of n would likely be sufficient to constrain the sentiment variance. In our dataset, the total number of tweets from the public was more than 500,000 while the number of tweets from political parties was more than 8,000 each. Therefore, we expect that our aggregated sentiment scores are sufficiently precise to compare the sentiments of different groups of tweets.

The accuracy of sentiment scores given by the TextBlob package on tweets is not immediately clear. Therefore, to quantify its performance we manually classified the sentiments of 158 tweets as either negative, neutral or positive. These tweets were based on a random ordering of 300 tweets (different for all four of us) that were randomly sampled from the public political tweets dataset using three strata of negative, neutral and positive TextBlob scores. Of the 158 tweets that were human classified, 33 were classified by more than one person. Of those 33, the sentiment classification was unanimous 22 times suggesting a consistency score of $67\% \pm 16\%$ (95% confidence). This is much higher than the expected consistency of random guessing. In the case of the 11 contentious tweets where the human scores were not unanimous, the final human sentiment score was estimated as the sign of the average of each human classification.

The confusion matrix in figure 13 compares the manual sentiment classification with the sentiments estimated by the TextBlob package.

Figure 13: Confusion matrix of manual sentiment classification vs the sentiments estimated by TextBlob



The overall accuracy of the TextBlob sentiment classification is $41\% \pm 8\%$ (95% confidence), which, while seems a bit low, is above the 33% accuracy of random guessing. A chi-squared test confirms that the human and TextBlob sentiment classifications are correlated ($p = 0.05$). It is worth noting that the TextBlob sentiment scores are not discrete, and the human classifications also contain some error. Therefore, the true accuracy of the TextBlob scores is likely to be higher than this estimate. Furthermore, a neutral TextBlob classification indicates that the sentiment was exactly zero, while small non-zero scores, such as ± 0.1, would still classify as positive or negative. With this system, errors involving neutral classifications are more likely and not as significant as misclassification of positive tweets as negative and vice versa. The total error rate of positive tweets to negative and vice versa is only $18\% \pm 6\%$ (95% confidence), much less than 33%.

Making the Data Confess

Politician vs Public Tweet Count Comparison

From hashtag word clouds and word frequency counts, we were able to isolate topics that were covered most often by the politicians and general public. Interestingly, the electorate names that were mentioned most often in the politician tweets, such as Higgins or Dickson, often correspond to key seats in the election. We can hypothesise that politicians tweeted more about electorates that were marginal seats and vital to the election outcome in an attempt to sway voters.

Closer analysis of the tweets related to electorates showed some very interesting behaviour across the Labor, Greens and LNP parties. Here we present only some of the plots to illustrate our results. Across the three parties, the electorates that were mentioned most often include Warringah, Kooyong, Dickson and Higgins. Warringah was the electorate that was held by LNP's Tony Abbott (a former prime minister), for more than twenty years. Dickson and Kooyong were also held by well-known LNP figures - Peter Dutton (Minister for Home Affairs) and Josh Frydenberg (treasurer) respectively.

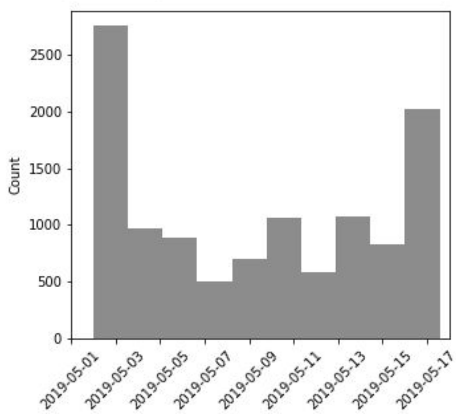


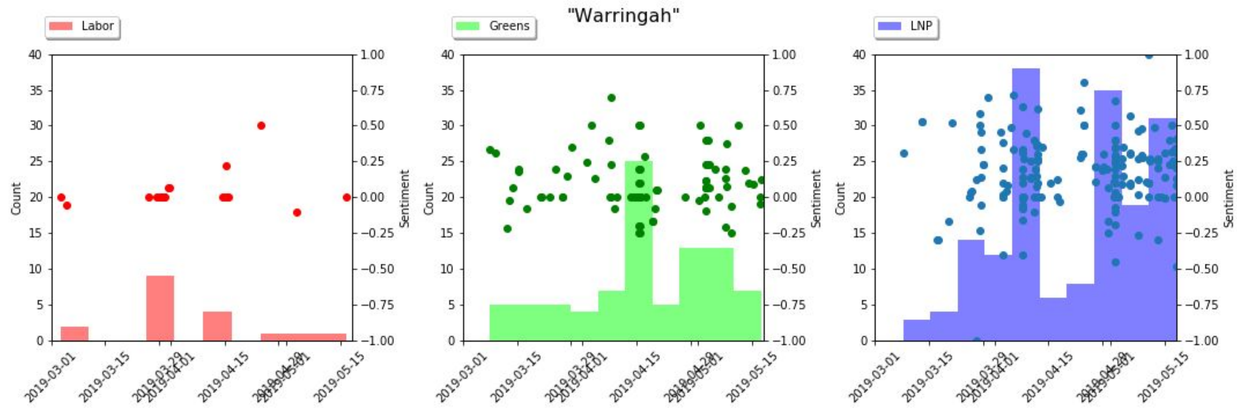
Figure 14: The number of tweets containing "Warringah" from 1. - 17. May in the public tweet dataset.

The number of tweets in the public dataset can give us an indication of the general interest in a particular electorate. We have found that increased activity occurs often at the time of some political event or news/poll result. The figure above demonstrates the high interest of the general public in the Warringah electorate, with thousands of tweets in a week. The particularly high number of tweets at the beginning of this collection period are likely due to the Warringah debate on 2. May with Tony Abbott and Zali Steggall[24].

The sentiment analysis of these tweets was not useful as they show wild fluctuations, with no discernable pattern. This is likely because people have no particular sentiment related to the electorate itself, but rather to the candidates of the electorate.

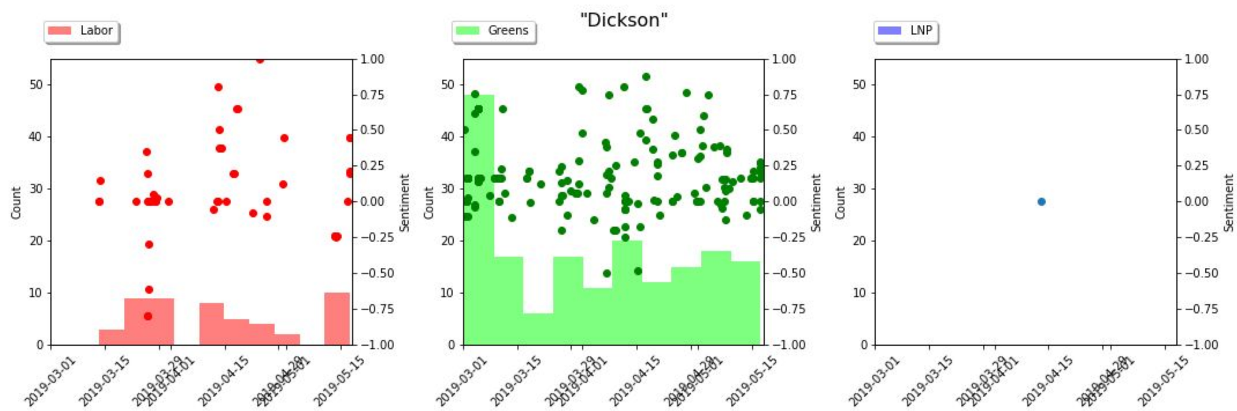
[24] (2019, May 2). Tony Abbott, Zali Steggall clash as battle for Warringah heats up - ABC. Retrieved June 3, 2019, from https://www.abc.net.au/news/2019-05-02/tony-abbott-zali-steggall-in-debate-battle-for-warringah/11074498

Figure 15: Number and sentiment of tweets containing “Warringah” for the three parties



In the politician tweets, we see that Warringah was most often mentioned in LNP tweets and least often by Labor. Despite this, Abbott lost his seat to an Independent. On the other hand, the Dickson, Kooyong and Higgins electorates were won by the LNP party. For each of these electorates, LNP had far fewer tweets - even for the marginal seats! This is demonstrated in the Figure below showing the number and sentiment of tweets containing “Dickson” for each party.

Figure 16: Number and sentiment of tweets containing “Dickson” for the three parties



These results indicate that fewer tweets may actually be more beneficial than many. Though we cannot state causality, it is possible that LNP were able to avoid controversial topics and controversial candidates by tweeting less, and thereby did not risk alienating their followers.

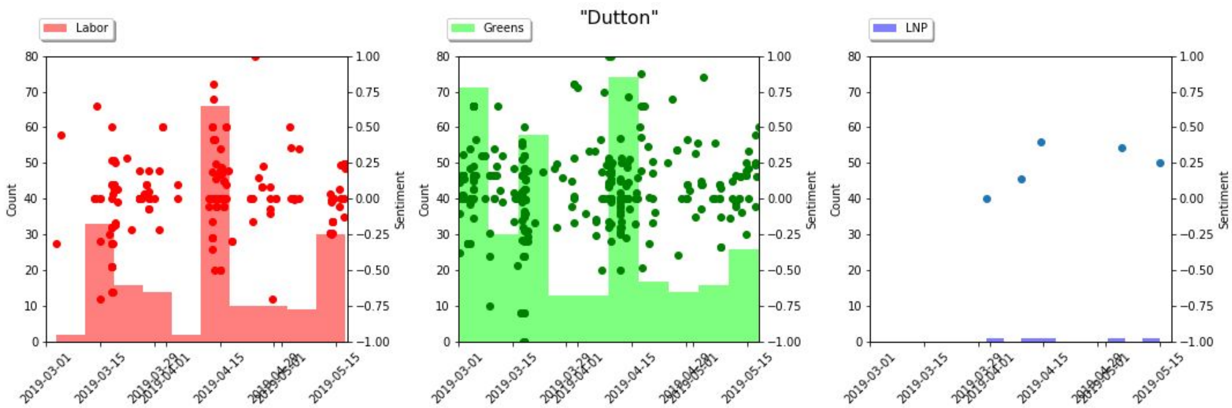
The Dickson candidate, Peter Dutton, is quite a controversial LNP member. Former prime minister, Paul Keating, prompted voters to “drive a political stake through his dark political heart”[25]. In previous figures we could see that politician tweets tend to have a positive sentiment when their tweets contain are related

---

[25] (2019, May 13). Paul Keating urges voters to 'drive a political stake through' Peter .... Retrieved June 3, 2019, from
https://www.abc.net.au/news/2019-05-14/paul-keating-peter-dutton-political-stake-through-dark-heart/11110814

to an electorate. This may be because they are addressing the voters in a positive manner, or promoting their own candidate. In contrast, the Greens and Labor tweets about Peter Dutton are more negative, as seen in the Figure below. This is reflective of how controversial Dutton was.

Figure 17: Number and sentiment of tweets containing "Dutton" for the three parties.



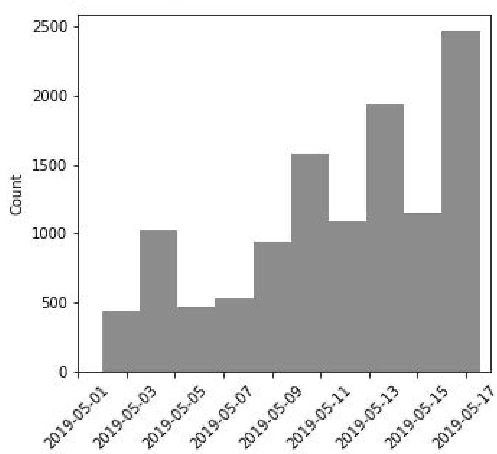


Figure 18: Number of tweets from the public data set that contain "Dutton" from 1. - 17. May

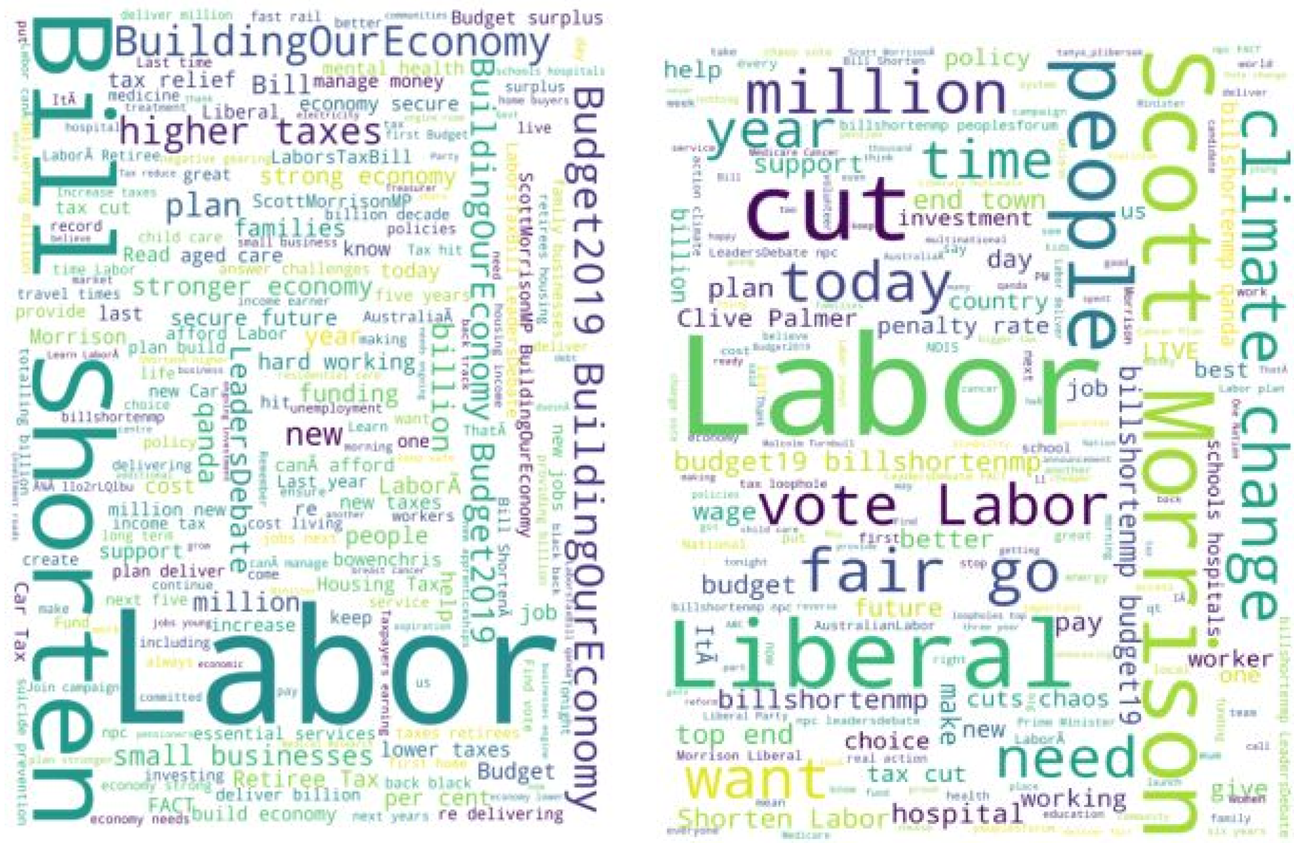
An increasing interest in Dutton was shown in the general public, with the number of public tweets containing "Dutton" increasing from the 1st May to 17th May. The Labor and Greens candidates also tweeted about Dutton often, whereas the LNP candidates tweeted about Dutton only five times since the 1st March. Though there was clear public interest in Dutton and his electorate, LNP did not discuss this on Twitter. Despite LNP's lack of engagement regarding this electorate on Twitter, they won the marginal seat. Similar results were shown in the electorates of Kooyong and Higgins. This supports our theory that tweeting less (and avoiding controversial topics) may actually be beneficial.

Analysing the Theories

1. "Cult of Personality" or the "Morrison Tactic"

We created word clouds to explore visually whether the LNP mentioned their leader significantly more than policies, or than Labor did their leader. These wordclouds are shown in Fig 19 below.

Fig 19: Word Cloud for Liberals official account (left) and Labor official account (right)



These wordclouds do not obviously support this theory. The liberal's official account barely mentioned Scott Morrison at all and instead appear much more focused on the opposing party. Similarly, Labor had many tweets about Scott Morrison and Liberals, but also about Labor themselves.

Despite the well documented unpopularity of Bill Shorten[26], wordclouds formed from tweets rated with a strong negative sentiment (Figure 20) show Scott Morrison as one of the most frequent mentions, with Bill Shorten appearing far less frequently. Conversely, in the wordcloud formed from positive sentiment tweets (Figure 21), Bill Shorten clearly stands out and Scott Morrison can't be detected.

---

[26] "Bill Shorten's path to victory is to avoid a two-man ... - The Guardian." 29 Apr. 2019, https://www.theguardian.com/australia-news/commentisfree/2019/apr/30/bill-shortens-path-to-victory-is-to-avoid-a-two-man-slugfest. Accessed 7 Jun. 2019.

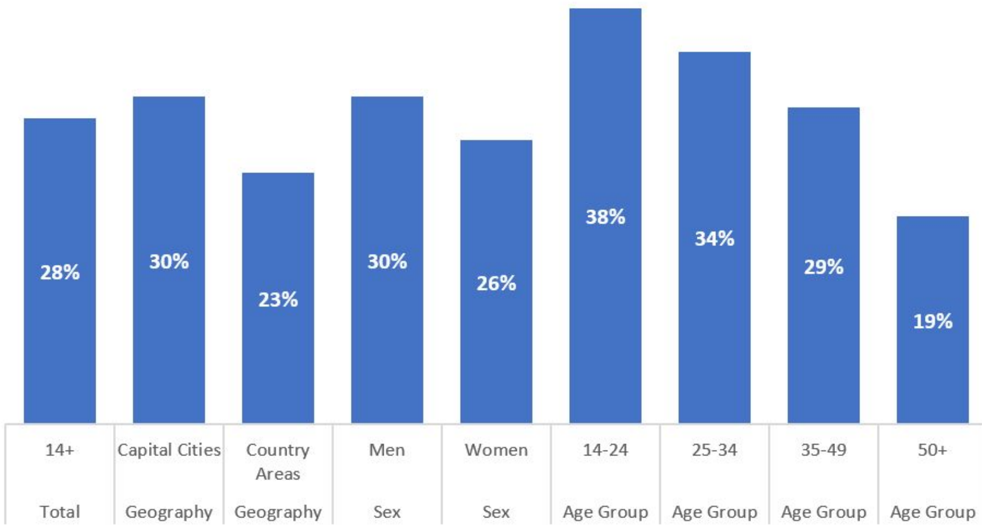
Figure 20: Word Cloud formed from strongly negative tweets from the public dataset

Figure 21. Word Cloud formed from strongly positive tweets from the public dataset

A significant explanation for these results could be the demographic of the Twitter users. Figure 22 demonstrates the demographic of Australian Twitter users. Australian Twitter users are demographically skewed to younger, urban and/or well-educated people[27]. Also, note that Twitter users can be as young as 14 years old, but the voting age in Australia is 18. We hypothesise that these younger users do not contribute significantly to the election discussions on Twitter. Unfortunately, however, we do not have the age profile of the users in our dataset to confirm this.

[27] "Twitter: a bird in the hand for 5.4m Australians - Roy Morgan Research." http://www.roymorgan.com/findings/6859-Twitter-active-users-australia-march-2016-201606210913. Accessed 4 Jun. 2019.

Figure 22. Roy Morgan - Twitter Usage Across Major Demographics in 2016[28]



2. Attack Ads Triumph Over Policy Detail

Labor came to this election with very bold, progressive policies, whereas the Liberals provided very little policy. The theory is that the Liberals distilled the contest to a referendum on Labor's ambitious plans and focussed on discrediting their policies.
We explored one example of this, given in ABC's coverage[29].

> *"Labor failed to sufficiently counter the Coalition's depiction of its franking credits policy as a "retiree tax", its curbing of negative gearing and capital gains tax as a "housing tax" and its removal of a superannuation concession as a "superannuation tax""*

The percentage of tweets falling within each key election topic for each party was calculated. For LNP, 12% of tweets fell under "tax", compared with 6% for Labor and 2% for Greens. Of LNP's "tax" tweets, 55% also mentioned "Labor", whereas only 2% of Labor's "tax" tweets mentioned LNP[30]. This supports the ABC's example that LNP followed a smear campaign against Labor's policies.

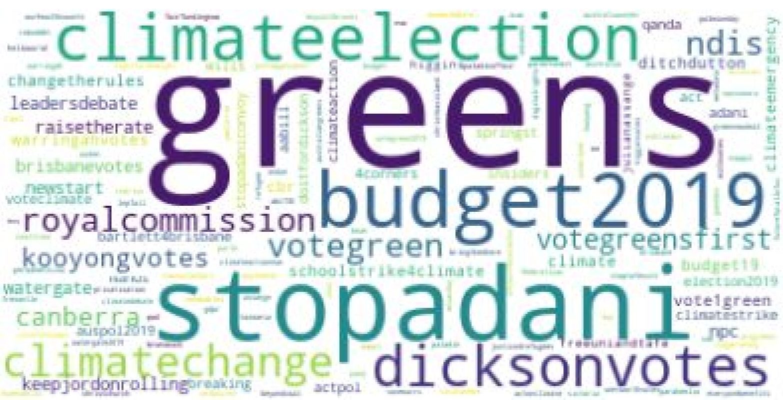
Party wordclouds formed from the candidate tweets are shown in Fig 23 below. These visually highlight the different policies or strategies of the parties. The Labor wordcloud shows key labor policies including

[28] "Twitter: a bird in the hand for 5.4m Australians - Roy Morgan Research." http://www.roymorgan.com/findings/6859-Twitter-active-users-australia-march-2016-201606210913. Accessed 4 Jun. 2019.
[29] "Scott Morrison goes from accidental Prime Minister to ... - ABC." 19 May. 2019, https://www.abc.net.au/news/2019-05-20/election-morrison-gains-authority-to-change-coalition-direction/11128460 . Accessed 2 Jun. 2019.
[30] Or "Liberals".

the more controversial ones: Franking credit, Change the rules, Climate election. The Greens wordcloud highlights their policies to stop the Adani mine and tackle climate change. By contrast, the Liberals don't cover controversial topics, and instead stick to "Building our economy", "putting locals first" and discreding Labor's policies (e.g. "labor's tax bill").

Figure 23: Wordclouds of hashtags used most often in tweets from Labor (top), Greens (centre) and LNP (bottom)

Labor



Greens

LNP

3. Climate Election

With climate change a key topic in the election, we explored and compared how the politicians and public tweeted on the subject. The graph below shows the number of tweets which use the word "climate", for each party and the public. It clearly shows an increase in tweets on climate topics on days of substantial "climate" news for the public, the ALP and The Greens, however LNP consistently say very little on the subject.

Figure 24: Number of climate tweets per day for each dataset

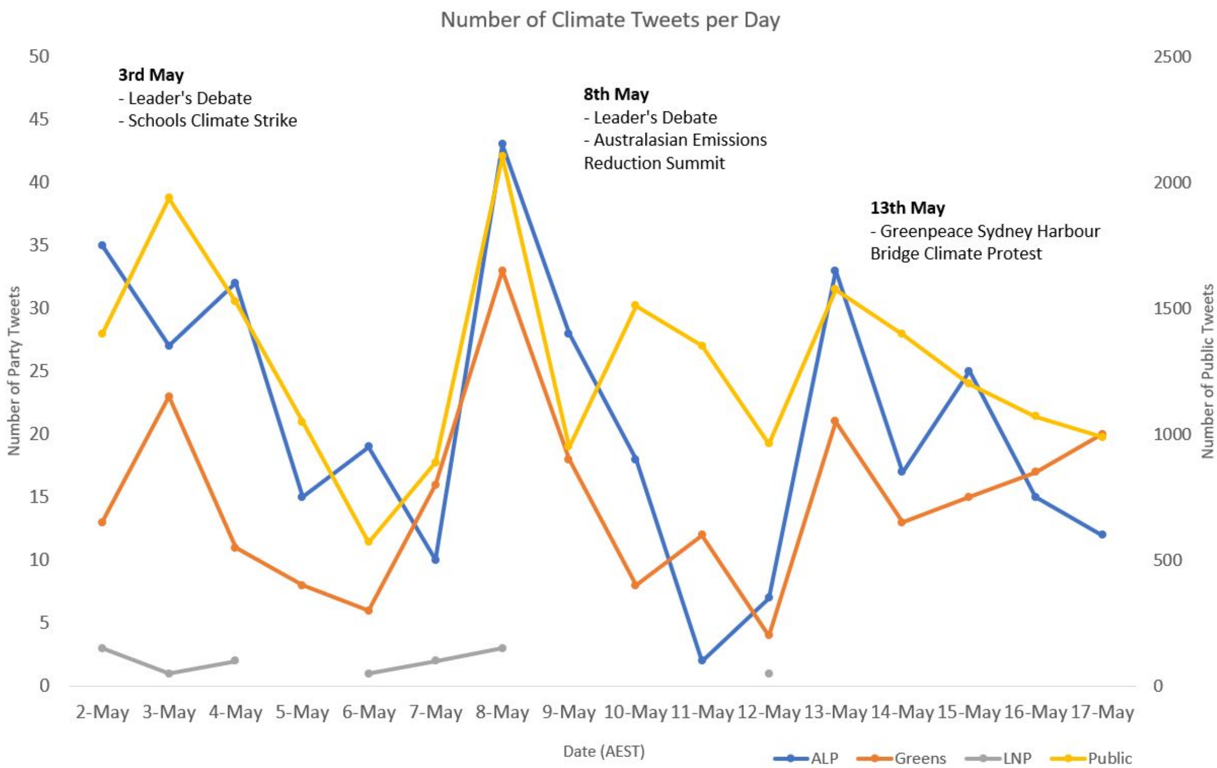
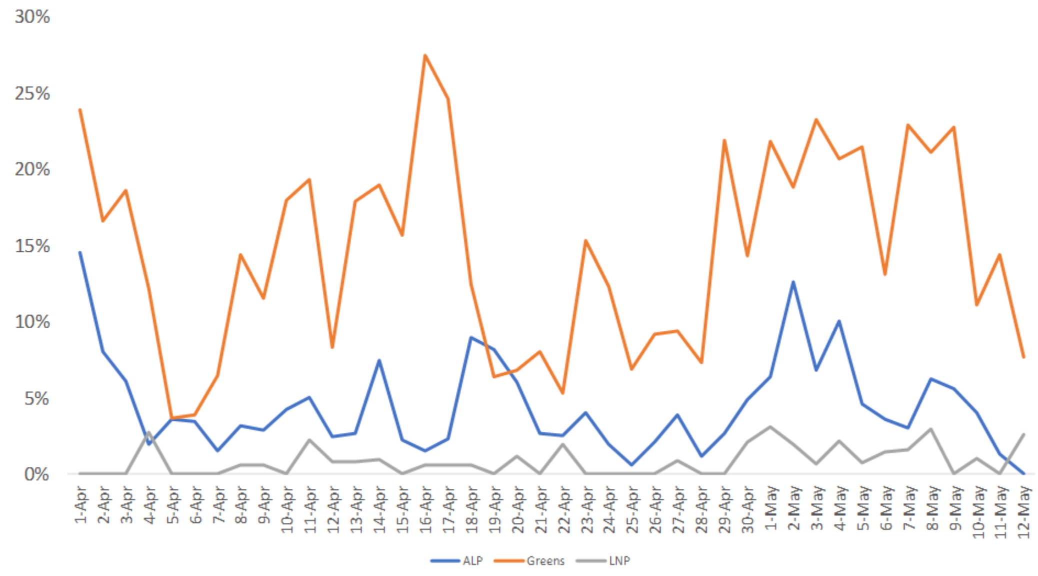


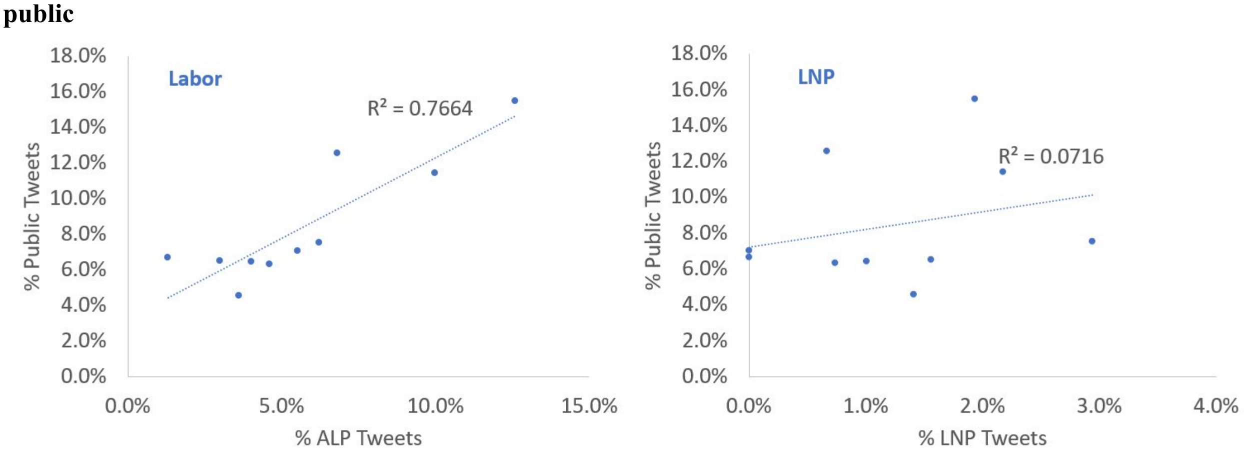
Fig 25, showing the percentage of tweets each day mentioning "Climate" for each party, also demonstrates how much more The Greens and Labor were tweeting about "Climate" than LNP.

Fig 25: Percentage of tweets each day mentioning "Climate" for the three main parties



To quantify the alignment between the parties and the public on this topic, regression analysis on the percentage of tweets was performed. Scatterplots show the correlation (or otherwise) of the percentage each day between Labour and the public, and LNP and the public.

Figure 26: Scatterplots of percentage of "Climate" tweets each day showing relationship between party and public



Regression analysis on these relationships showed a significant relationship between the percentages of Labor and the public (significance F of 0.000904) but no significant relationship between LNP and the public (significance F of 0.45495). With Labor tweeting more on Climate, and also following the spikes when significant climate events happened, it is perhaps not surprising there is more of a correlation with the public, who also react to these "climate" events in their tweeting.

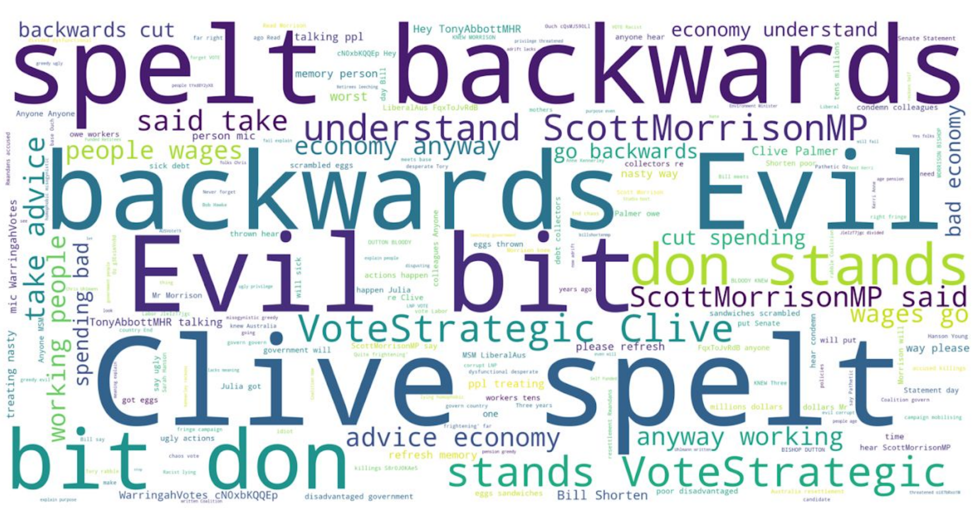
It is interesting though that despite this evident interest in "Climate" for this election, and with Labor clearly showing more alignment with the public on this topic, it is was not important enough to override all other election topics and win Labor the election.

4. Quexit

We have seen that Climate Change was a big election topic, however one of the explanations for the large swing against Labor in Queensland is that Queensland voters found Labor's stance on climate change less appealing than the Southern states, with the money in their pocket and jobs being more important issues for the electorate. Bill Shorten was also claimed to be particularly unpopular in the state, with one political scientist quoted as saying "They saw him as a southern trade unionist with a green agenda...None of those things appeal to regional Queenslanders who are keen to get the extractive industries moving to generate blue collar jobs."[31]

We created wordclouds of positive and negative sentiment on the subset of public data from Queensland to explore these explanations. Whilst not much can be discerned from the negative sentiment wordcloud, the positive sentiment wordcloud is certainly interesting. This wordcloud shows that Bill Shorten and "Vote Labor" featured heavily in these strongly positive-sentiment tweets, which one would not expect. As discussed above, this result could partly be because of the demographics of Twitter users. Appendix C explores the average and minimum sentiment of tweets by region, though unfortunately this does not give any conclusive evidence for or against the Quexit theory.

negative



---

[31] "It was the election they couldn't lose, so what went so wrong for Labor?." 19 May. 2019, https://www.sbs.com.au/news/it-was-the-election-they-couldn-t-lose-so-what-went-so-wrong-for-labor. Accessed 4 Jun. 2019.

positive

The proportion of tweets covering each of our key topics is the same across the whole public tweet dataset as in the Queensland-only dataset. The only two topics with a difference of greater than 0.1% are Climate and Adani. There are 27 such topics altogether, but a table of some key topics is shown below to demonstrate the finding.

Figure 27: Proportion of tweets covering each topic in full public dataset and Queensland subset.

Topic	All Public Tweets	QLD Subset
Climate	7.8%	6.9%
Tax	5.8%	5.7%
Adani	0.9%	1.5%
Indigenous	0.3%	0.3%
Housing	0.7%	0.6%
Refugee	0.6%	0.6%

This finding with “Climate” and “Adani” may help support the theory that Queensland voters are less interested in climate change issues and, in particular in rural Queensland, are more interested in the new jobs that would be created by Adani’s new mine.
It is important to note that the method used for filtering the tweets to Queensland biases towards city locations. Inner-city electorates are generally richer and have a better-educated population with

less-pressing economic and physical needs. This encourages greater focus on topics such as climate change in these electorates as compared to in more rural electorates[32].

Responses to Feedback

One of the main pieces of feedback we received, both from the lecturer and peers, is that our project could be extended by considering the location of Twitter users. In particular, it was suggested to focus on how Queensland tweets differed from other areas. In response to this, we added a section about the “Quexit” theory and our analysis of tweets per location, and discussed the demographic of these Twitter users. More information can be found in the section “Quexit” under “Making the Data Confess”.

It was also suggested that the candidates’ total tweets should be compared to determine how active each party is. This was an interesting piece of analysis (in section “Getting the data we need”) which demonstrated how different the tweeting behaviour of each party is. In addition, we added more explanation about sentiment analysis, as per peer feedback.

Finally, we would have liked to apply some of the algorithms that were suggested, such as “summarizer” or bag of words model. Unfortunately, due to time constraints, we were unable to do this. We would also have liked to investigate more closely the location and demographics of the Twitter users but, again, our analysis was limited by time.

Related Works

Though our work has given hints as to why the election unfolded as it did, our data could not have been used to predict the result. This is of course partly due to the limited time frame we considered, but previous studies have also shown that Twitter data has limited predictive power[33]. In addition, the study by Murphy concluded that many tweets by candidates do not necessarily lead to a successful election, which is also supported by our finding that LNP candidates tweeted the least out of all three parties.

A literature review by Jungherr in 2016 showed no particular link between the amount candidates used Twitter and electoral success. Jungher also stated that political tweets are “skewed toward the opinions of a small, nonrepresentative, politically interested, and partisan subgroup of a population”[34]. The limitations of Twitter data for predictive purposes include the demographic bias and that it is self-selected[35] - people use Twitter voluntarily.

---

[32] "Election 2019: What happened to the climate change vote we ... - ABC." 19 May. 2019, https://www.abc.net.au/news/2019-05-20/what-happened-to-the-climate-change-vote/11128128. Accessed 4 Jun. 2019.
[33] Dhiraj Murthy (2015) Twitter and elections: are tweets, predictive, reactive, or a form of buzz?, Information, Communication & Society, 18:7, 816-831, DOI: 10.1080/1369118X.2015.1006659
[34] Jungherr, A. (2016). Twitter use in election campaigns: A systematic literature review. *Journal of information technology & politics*, *13*(1), 72-91.
[35] Gayo-Avello, D. (2013). A meta-analysis of state-of-the-art electoral prediction from Twitter data. *Social Science Computer Review*, *31*(6), 649-679.

However, there is also evidence that Twitter data *can* be used to make election predictions. Recently Griffith academic Prof. Bela Stantic has been featured in the news, as he predicted the 2019 Australian election result correctly using social media data[36]. He had also predicted the Brexit vote and America's 2016 election correctly. Unfortunately, there are currently no papers published that explains the methodology for his findings.

Conclusion

Our analysis of the Twitter data for the 2019 Australian election has given hints as to why the election was won by LNP rather than the Labor party. We have analysed some of the popular theories about the election result, and found that the "Cult of personality" theory was not supported by the Twitter data. No conclusive evidence for or against the Quexit theory could be determined. However, we did find an alternative theory within our data - that LNP avoided most controversial topics. This behaviour appeared in many areas of our data - LNP avoided controversial topics such as climate change, they did not tweet about their more controversial candidate, Peter Dutton, and they managed to win the key seats even with very few tweets about these electorates.

It is important to note that we only investigated *Twitter* data, which has inherent biases due to the self-selection and demographics of the users. Though we have started an investigation of the demographic of the Twitter users, there is much more that can be done to explore the socioeconomic status of Twitter users. If we were to extend our project, we could also consider other social media, such as Facebook, which are likely to be used by different demographics. Though our data was unable to predict the election outcome, it gave us valuable insights as to the way the election unfolded. In fact, these insights could be used to guide a party's Twitter approach in the next election.

[36] (2019, May 19). Election 2019: How the polls got it so wrong in predicting a ... - ABC. Retrieved June 4, 2019, from https://www.abc.net.au/news/2019-05-19/federal-election-results-how-the-polls-got-it-so-wrong/11128176

APPENDIX A

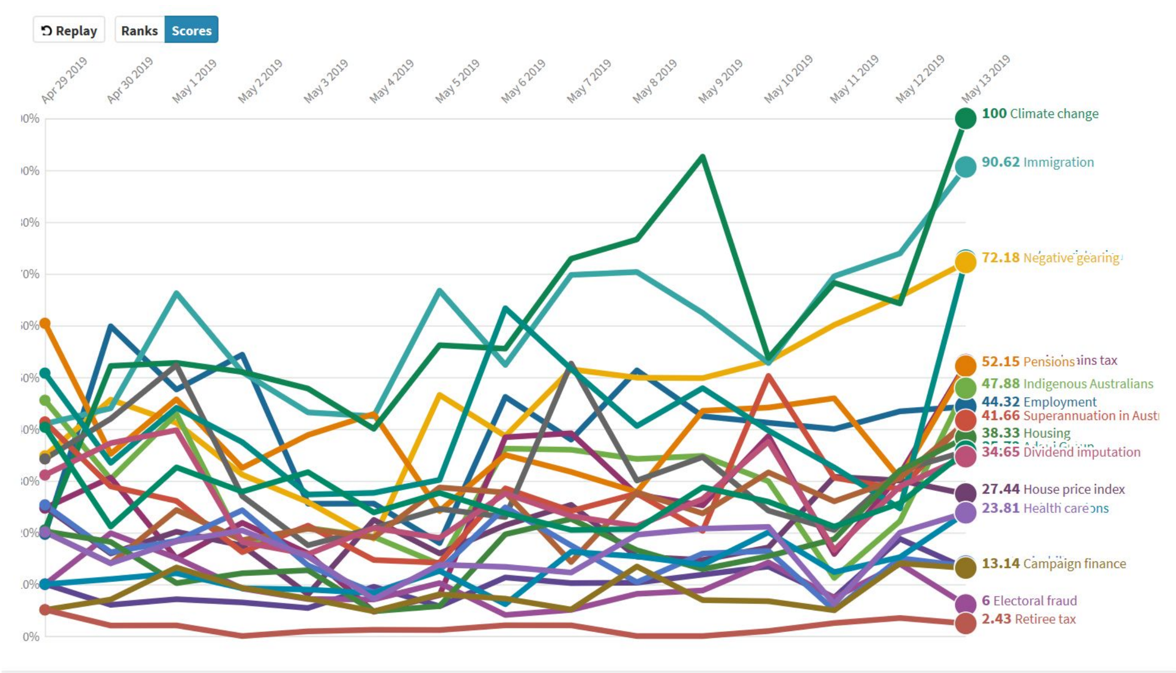
An example of the python code used to collect the Twitter data related to the Australian election.

```
import tweepy
from tweepy import OAuthHandler
from tweepy import Stream
from tweepy.streaming import StreamListener
import csv
import twitter_credentials
import sys
import time
import datetime
non_bmp_map = dict.fromkeys(range(0x10000, sys.maxunicode + 1), 0xfffd)
auth = tweepy.OAuthHandler(twitter_credentials.consumer_key, twitter_credentials.consumer_secret)
auth.set_access_token(twitter_credentials.access_token, twitter_credentials.access_token_secret)
api = tweepy.API(auth)

#API.search(q[, lang][, locale][, count][, page][, since_id][, geocode][, show_user])
# Open/create a file to append data to
csvFile = open('resulttest.csv', 'w', newline='',encoding="utf-8")
#Use csv writer
csvWriter = csv.writer(csvFile)
csvWriter.writerow(["created_at","tweet_id","screen_name","retweet","text","hashtags","retweet_count","r
etweeted","user_location"])

#query with selected search terms
query = ('Election OR auspol OR AustraliaVotes OR AusVotes2019 OR AusVotes19 OR AusVotes OR
AustraliaDecides')
max_tweets = 9999999
from_date="2019-05-05"
to_date="2019-05-06"
search_filter=""
#Limit tweets to around Australia
```

APPENDIX B
Google Trends

Analysis on Google trends supported the finding that climate change was a top election topic, as were the other topics selected from our word frequency count analysis.

Figure B.1: Google Trends graphs showing top election-related search terms in Australia

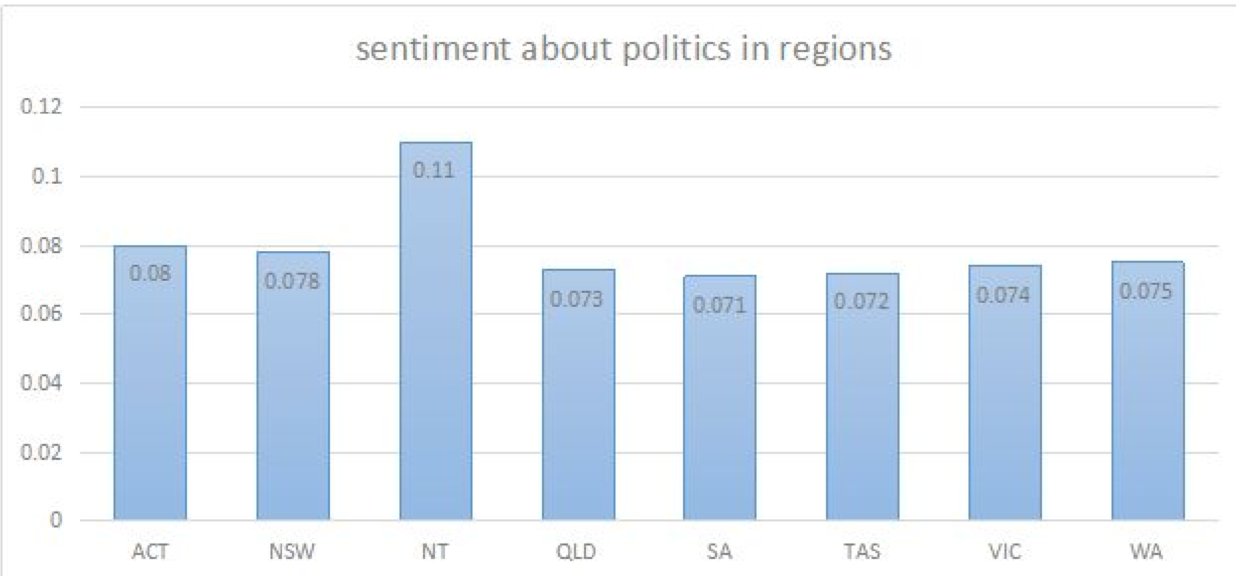


APPENDIX C

Average Sentiment by Region

The average sentiment of tweets was investigated across location in Australia. This was done in an attempt to discern whether there was a specific attitude towards politics across all states and territories. Figure C.1 shows that the sentiment of tweets in the public dataset does not vary much between regions, as they are all approximately zero. This indicates that there must be, on average, similar amounts of positive and negative sentiment to cancel out.

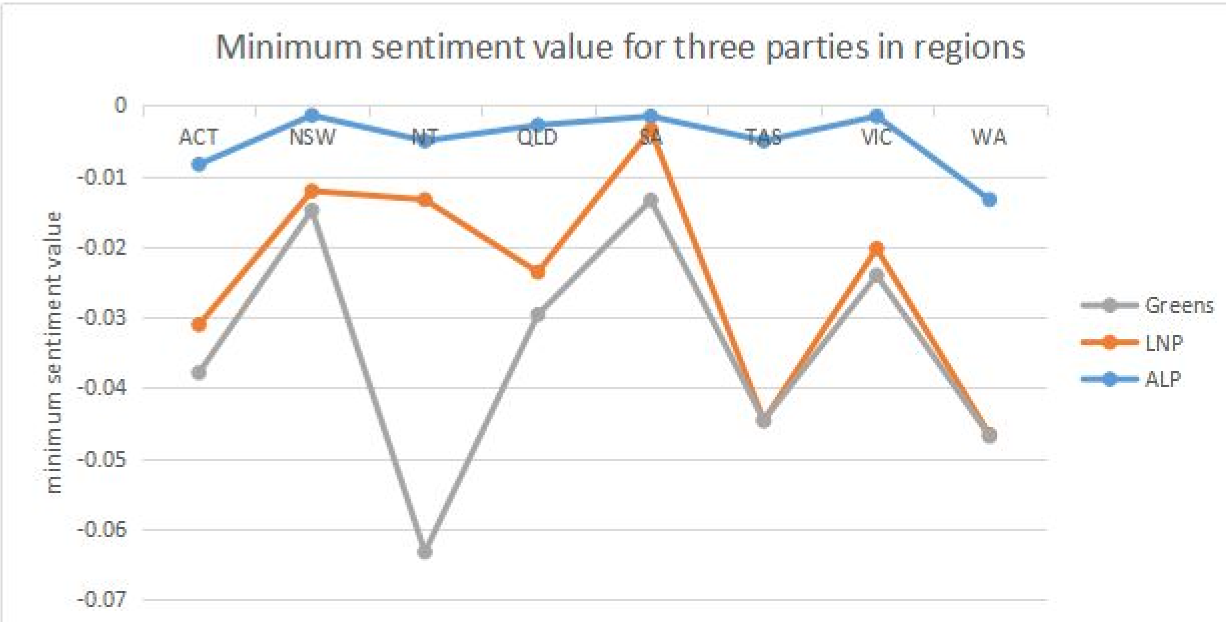
Figure C.1: Average sentiment of tweets related to the Australian election from each state/region



The average sentiment of tweets by each party was also investigated for all regions, but it was found that there were only very slight differences between regions. Given that the results are inconclusive, the plots are not shown here.

The maximum sentiment for every party's tweet in every region is 1.The minimum sentiment of tweets from all parties across all regions are shown in Figure C.2. ALP has a stable trend with a minimum score near zero in all regions. It is interesting to see that the Greens Party has the worst score in every region, compared with other two parties. This may indicate that their rhetoric can be more negative than the other two parties. Note, however, that the tweets analysed here are only a subset of the total party tweets, since not all tweets have useable location data. As such, this gives us only a possible indication of the behaviour of the three parties.

Figure C.2: Minimum sentiment of party tweets across all regions



An interesting avenue to pursue would be to look at the sentiment of different regions to specific topics in the election. This analysis would also give more promising results if the regions were more fine-grained, so that we could discuss the difference between rural areas and cities in Australia.