



## **DS6501 Social Data Analytics**

### **Trimester I**

### **Assignment I: Text mining social data**

**Tutor:** Dr. Abubakar Siddique

**Due Date:** 21 April 2024

**Student:** Adya Sinha

**Submission Date:** 21 April 2024

## Table Of Contents

<b>Executive Summary.....</b>	<b>3</b>
<b>Political parties(Greens, Labour, National) Report.....</b>	<b>4</b>
<b>Dataset Analysis and Techniques.....</b>	<b>4</b>
Data Preprocessing.....	4
Sentiment Analysis of Tweets.....	7
Visual representation of sentiment scores.....	9
Finding the most positive & negative tweets.....	11
Creating a word cloud using a TDM(Term Document Matrix).....	11
<b>Conclusion.....</b>	<b>14</b>

## **Executive Summary**

This report analyses the sentiment of tweets from the official Twitter accounts of three New Zealand's parties- Green Party, National Party, and Labour Party. The given analysis classifies tweets as positive, negative, or neutral, and identifies the most positive and negative tweets as it explores the emotional content expressed.

The given dataset had tweets from these parties' accounts, which were first preprocessed to remove any unwanted texts. Then a sentiment analysis using the `get_nrc_sentiment` and `get_sentiment` functions was performed. To help us understand the data more clearly, visualisations were created that include pie charts and word clouds.

The sentiment analysis in the form of a pie chart revealed that the Labour Party had the highest proportion of positive tweets, followed by the Greens Party. The National Party showed a more balanced sentiment, but was leaning slightly negative.

The given word clouds highlights the most frequent words used by each party which provides us with insights into their tweets.

Overall, this report demonstrates applying sentiment analysis techniques to understand the sentiment and emotional resonance in political parties' Twitter communication.

## **Political parties(Greens, Labour, National) Report**

The accumulated tweets of the official Twitter accounts of 3 parties- The Green Party, Labour Party, and the National Party. This provides us with an abundant dataset to understand the public dataset towards the given parties. The Green party is a left-wing environmentalist party whereas the Labour Party is a centre-left political party and the National Party is a centre-right party. They all have considerable power over the political climate of this country along with their strong Twitter presence to communicate with the general public.

## **Dataset Analysis and Techniques**

This dataset consists of tweets collected from the Twitter accounts of the three parties- Greens, National, and Labour parties. To perform the sentiment analysis on the tweets, we can use the 'get\_nrc\_sentiment' and 'get\_sentiment' functions from the 'syuzhet' package in RStudio.

### **Data Preprocessing**

Before conducting the sentiment analysis, the code performs several data preprocessing steps to clean and prepare the tweet data. These steps include:

1. Installing and loading the required packages which are: `tm`, `wordcloud`, `RColorBrewer`, `SnowballC`, and `syuzhet`.

```
# Load required Packages
install.packages("tm")
install.packages("wordcloud")
install.packages("RColorBrewer")
install.packages("SnowballC")
install.packages("syuzhet")

# Load libraries into R Studio
library(tm)
library(wordcloud)
library(RColorBrewer)
library(SnowballC)
library(syuzhet)
```

2. We have to read the CSV files containing the tweets for each party, which is done after loading the required packages.

```
# Reading the csv files
tweets_Greens.df<- read.csv("./Political Parties/NZGreens_tweets.csv")
tweets_labour.df<- read.csv("./Political Parties/nzlabour_tweets.csv")
tweets_National.df<- read.csv("./Political Parties/NZNationalParty_tweets.csv")
```

3. The next step is to extract the text content from the data frames.

```
# Display 'text' field of data frame
# For NZGreens
head(tweets_Greens.df$text)
tweets_Greens.df2<- tweets_Greens.df$text

# For NZlabour
head(tweets_labour.df$text)
tweets_labour.df2<- tweets_labour.df$text

# For NZNational
head(tweets_National.df$text)
tweets_National.df2<- tweets_National.df$text
```

4. After extracting the text content, we can convert the text data to UTF-8 encoding to handle any invalid strings. There were some invalid strings I found in the “Greens Party” data, hence this conversion.

```
#####

invalid_strings <- is.na(iconv(tweets_Greens.df2, "UTF-8", "UTF-8"))
tweets_Greens.df2[invalid_strings]
tweets_Greens.df2 <- iconv(tweets_Greens.df2, "latin1", "UTF-8")

invalid_strings <- is.na(iconv(tweets_labour.df2, "UTF-8", "UTF-8"))
tweets_labour.df2[invalid_strings]
tweets_labour.df2 <- iconv(tweets_labour.df2, "latin1", "UTF-8")

invalid_strings <- is.na(iconv(tweets_National.df2, "UTF-8", "UTF-8"))
tweets_National.df2[invalid_strings]
tweets_National.df2 <- iconv(tweets_National.df2, "latin1", "UTF-8")

#####
```

5. Then I removed the URLs, hashtags, mentions, and other special characters from the tweets using regular expressions to extract junk from the data. This is cleaned using the help of the ‘gsub’ function.

- `www\\. [^\\s]+`: This specific part matches the URLs which start with "[www.](http://www.)" which is followed by one or more non-whitespace characters.
- `@\\S+`: It finds any mention starting with "@" followed by one or more non-whitespace characters.
- `http\\S+`: Now this part will match any URL which starts with "http" followed by one or more non-whitespace characters.
- `#\\S+`: It finds the hashtags starting with "#" after that, by one or more non-whitespace characters.
- `<[^>]+>`: This part of the code will find all HTML tags enclosed within angle brackets (< and >).
- `"[^a-zA-Z0-9\\s]|[[[:cntrl:]]|[[[:punct:]]]"`: This pattern matches all characters which are not uppercase or lowercase characters, a whitespace character, a digit along with any control punctuation characters.

```
# Use a 'find and replace' function to remove garbage from tweets
# For NZGreens
tweets_Greens.df2<- gsub("www\\. [^\\s]+|@\\S+|http\\S+|#\\S+|<[^>]+>", "", tweets_Greens.df2)
tweets_Greens.df2<- gsub("[^a-zA-Z0-9\\s]|[[[:cntrl:]]|[[[:punct:]]]", " ", tweets_Greens.df2)
head(tweets_Greens.df2)

# For NZlabour
tweets_labour.df2<- gsub("www\\. [^\\s]+|@\\S+|http\\S+|#\\S+|<[^>]+>", "", tweets_labour.df2)
tweets_labour.df2<- gsub("[^a-zA-Z0-9\\s]|[[[:cntrl:]]|[[[:punct:]]]", " ", tweets_labour.df2)
head(tweets_labour.df2)

# For NZNational
tweets_National.df2<- gsub("www\\. [^\\s]+|@\\S+|http\\S+|#\\S+|<[^>]+>", "", tweets_National.df2)
tweets_National.df2<- gsub("[^a-zA-Z0-9\\s]|[[[:cntrl:]]|[[[:punct:]]]", " ", tweets_National.df2)
head(tweets_National.df2)
```

This is the input before running the function:-

```
> tweets_Greens.df2
[1] "Great article featuring Green MP @EugenieSage \nhttps://t.co/81XxH6rnx"
[2] "#Mustwatch this emotional speech from Greens co-leader @jamespeshaw in the House today. \n\n\u0093I don\u0092t think I\u0092ve ever\u0085 https://t.co/YcdsvzPTCX"
[3] "#BREAKING The Green Party is today calling on the Government to double the maximum Civil Defence Payment for Auckland\u0085 https://t.co/UckHwBX5NS"
[4] "Greens co-leader @MaramaDavidson with some strong words for the National Party at Waitangi Treaty Grounds earlier t\u0085 https://t.co/8gcfP910q5"
[5] "https://t.co/v1zgX2IkpC"
[6] "https://t.co/r6uHSS0PoY"
[7] "https://t.co/NDnKAGMX3o"
[8] "https://t.co/UdWMBKvHeJ https://t.co/k1417w0PN"
[9] "https://t.co/QkyIj9xUSS\n\nhttps://t.co/udCunQ8Uve"
[10] "\"Poverty is a political choice. We implore the Govt to use their power to freeze rents, like many of the grassroots\u0085 https://t.co/14egj2jJue"
[11] "#BREAKING The Green Party is joining more than 20 community organisations to call for an immediate rent freeze in T\u0085 https://t.co/d3Nj85wuTR"
[12] "For the cost of one year of the fuel tax cut, the Govt could have bought a free e-bike for every uni & high school\u0085 https://t.co/LIceyKVCgH"
[13] "It\u0092s clear that we will be living with the infection for many years to come. \n\nFocus must be on slowing the spread\u0085 https://t.co/G4JHC202SP"
[14] "It\u0092s critically important that the new Health Minister does not lose sight of our COVID response.\n\nThe pandemic is\u0085 https://t.co/4EOP61xL9j"
[15] "This is what climate change looks like. The solutions are the same as they always have been, and they are more urge\u0085 https://t.co/anviuxYyji"
```

This is the output which is free from all unnecessary characters:-

```
> tweets_Greens.df2
[1] "Great article featuring Green MP "
[2] " this emotional speech from Greens co leader in the House today I don t think I ve ever "
[3] " The Green Party is today calling on the Government to double the maximum Civil Defence Payment for Auckland "
[4] "Greens co leader with some strong words for the National Party at Waitangi Treaty Grounds earlier t "
[5] ""
[6] ""
[7] ""
[8] " "
[9] " "
[10] " Poverty is a political choice We implore the Govt to use their power to freeze rents like many of the grassroots "
[11] " The Green Party is joining more than 20 community organisations to call for an immediate rent freeze in T "
[12] "For the cost of one year of the fuel tax cut the Govt could have bought a free e bike for every uni amp high school "
[13] "It s clear that we will be living with the infection for many years to come Focus must be on slowing the spread "
[14] "It s critically important that the new Health Minister does not lose sight of our COVID response The pandemic is "
[15] "This is what climate change looks like The solutions are the same as they always have been and they are more urge "
```

This was done to clean all the texts from all the 3 respective parties.

## Sentiment Analysis of Tweets

The sentiment analysis was performed two ways. The first analysis was based on emotional content. The second analysis is by classifying tweets as either positive, negative, or neutral. The code then indicates the tweets of the highest sentiment scores for each of the parties.

- Getting sentiment score: In the 'syuzhet' package, using the 'get\_nrc\_sentiment' function, the tweets are categorised based on the different emotions which then return a sentiment score based on each emotion of the tweets. The first step is to convert the data frame to a vector. Then we identify the highest rated tweet for each of the emotional categories which are: anger, anticipation, disgust, fear, joy, sadness, surprise, and trust.

```
# Convert data frame into a vector before performing sentiment analysis
word_Greens.df<- as.vector(tweets_Greens.df2)
word_labour.df<- as.vector(tweets_labour.df2)
word_National.df<- as.vector(tweets_National.df2)
```

```
# Perform sentiment analysis to score tweets on emotion
emotion_Greens.df<- get_nrc_sentiment(word_Greens.df)
emotion_labour.df<- get_nrc_sentiment(word_labour.df)
emotion_National.df<- get_nrc_sentiment(word_National.df)
```

The results publish the emotions expressed in each tweet. The frequency of each specific emotion associates with a tweet. Given below are the results for the different parties after combining the tweets to sentiment scores: -

## The Greens Party

```
> # Combine tweets to sentiment scores
> emotion_Greens.df2<- cbind(tweets_Greens.df2, emotion_Greens.df)
> head(emotion_Greens.df2)
```

	tweets_Greens.df2	anger	anticipation
1	Great article featuring Green MP	0	0
2	this emotional speech from Greens co leader in the House today I don't think I've ever	0	0
3	The Green Party is today calling on the Government to double the maximum Civil Defence Payment for Auckland	0	0
4	Greens co leader with some strong words for the National Party at Waitangi Treaty Grounds earlier today	1	0
5		0	0
6		0	0

	disgust	fear	joy	sadness	surprise	trust	negative	positive
1	0	0	1	0	0	1	0	1
2	0	0	0	0	0	2	0	3
3	0	1	1	0	0	1	2	3
4	0	0	0	0	0	1	1	1
5	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0

## The Labour Party

```
> emotion_labour.df2<- cbind(tweets_labour.df2, emotion_labour.df)
> head(emotion_labour.df2)
```

	tweets_labour.df2	anger
1	Auckland flood support	0
2	Make sure to tune in to our Facebook page at 10:30 this morning to catch 2022 Labour Party Conference	0
3	We've just announced a massive infrastructure investment to kick start new housing developments across New Zealand	1
4	Read more above or head here to see the actions we're taking to protect Aotearoa's environment	0
5	As part of our commitment to protecting our environment for future generations we've phased out more problematic pesticides	0
6	For more info on local elections and how to vote head to	1

	anticipation	disgust	fear	joy	sadness	surprise	trust	negative	positive
1	0	0	1	0	0	0	0	0	0
2	0	0	0	0	0	1	0	0	0
3	1	0	0	0	0	0	0	1	0
4	0	0	0	0	0	0	0	0	1
5	0	0	0	0	0	0	1	0	1
6	1	0	0	1	1	1	1	1	1

## The National Party

```
> emotion_National.df2<- cbind(tweets_National.df2, emotion_National.df)
> head(emotion_National.df2)
```

	tweets_National.df2	anger
1		0
2	Congratulations Tama Potaka for becoming the new MP for Hamilton West	0
3	National's latest plan will combat youth offenders and reduce ram raids in New Zealand	3
4	Almost 20,000 Kiwis voted in the first 24 hours on what has been Labour's biggest fail vote here	1
5	More than 20,000 Kiwis have already signed our petition to stop Labour's Jobs Tax This latest tax would take \$34	0
6	In less than 10 hours over 10,000 Kiwis have made it clear they don't want Labour's latest tax grab the Jobs Tax	1

	anticipation	disgust	fear	joy	sadness	surprise	trust	negative	positive
1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
3	3	0	2	1	0	1	0	2	1
4	1	0	0	1	1	1	1	1	1
5	0	0	0	0	1	0	0	1	0
6	0	0	0	0	1	0	1	2	1

- Classification of tweets:** For each of the negative tweets, we use the 'get\_sentiment' function to extract sentiment scores. After that, a consistent approach is used by this function to define a text as positive, negative, or neutral, which is then established on the words that are used within a tweet. As we can see below, a score of 0 indicates neutral, while a negative score indicates negative sentiment. A higher number indicates a more positive tweet.



```

> sent_Greens.value
[1] 0.50 0.50 0.20 0.75 0.00 0.00 0.00 0.00 0.00 -0.65 0.20 0.00 0.35 0.30 1.00 0.00 0.00 0.00 1.25 0.00 0.05 0.80 1.50 0.30
[25] 2.05 0.30 0.75 0.50 0.00 -0.45 1.00 0.90 0.50 -0.25 -0.25 0.80 0.00 0.00 -2.10 0.10 0.00 0.80 -0.25 0.00 0.80 -1.00 -0.40 1.40
[49] 0.25 0.85 1.25 1.50 0.00 0.00 0.75 -0.70 0.85 0.00 1.50 0.65 0.50 0.75 0.00 0.00 0.00
> sent_labour.value
[1] 0.50 0.00 0.55 0.75 0.70 0.00 0.00 0.25 2.05 1.95 0.35 1.40 0.00 1.95 3.75 -0.60 1.55 4.20 1.65 0.80 0.75 1.95 0.80 2.30
[25] 2.25 3.05 0.40 0.75 0.40 0.00 0.50 -0.40 0.50 3.40 -0.30 3.30 1.10 0.10 2.00 1.05 2.40 0.50 1.00 0.00 2.00 2.45 0.80 -0.25
[49] 1.35 -0.50 -0.40 1.10 -0.30 0.90 0.65 0.00 2.20 1.60 0.00 1.25 0.20 0.25 2.50 0.80 0.50 -0.35 0.50 0.85 -0.50 1.00 1.00 1.60
[73] 0.25 1.60 0.75 1.90 0.80 0.90 0.00 0.00 1.85 0.75 2.35 1.60 1.20 1.50 0.80 0.30 1.25 0.25
> sent_National.value
[1] 0.00 1.55 0.45 -0.75 -0.65 0.25 0.50 1.15 -0.40 -1.75 0.80 -0.85 1.50 -1.35 0.50 0.00 -0.50 -0.10 0.75 0.75 -1.00 -0.85 0.70 1.30
[25] -0.50 0.50 0.50 0.75 1.30 -0.75 -1.00 0.70 0.00 1.00 -2.25 0.65 0.00 0.85 0.60 1.30 0.50 2.75 -0.20 -0.85 -1.50 -0.50 -1.00 -0.25
[49] -0.90 0.35 0.25 -2.30 0.65 0.00 -0.35 -1.25 -0.50 0.00 -1.50 -1.25 0.00 -0.95 0.00 -0.85 -0.75 0.00 1.05 0.00 0.75 2.15 -1.75 0.25
[73] 1.55 -0.10

```

After this, we store all the tweets with this: -

```

> # For NZNational
> positive_National.tweets<- word_National.df[sent_National.value>0]
> head(positive_National.tweets)
[1] "Congratulations Tama Potaka for becoming the new MP for Hamilton West "
[2] "National s latest plan will combat youth offenders and reduce ram raids in New Zealand "
[3] "In less than 10 hours over 10 000 Kiwis have made it clear they don t want Labour s latest tax grab the Jobs Tax "
[4] "This morning at Parliament signed the book of condolences for Her Majesty Queen Elizabeth II thanki "
[5] "Sign our petition to reverse Labour s tenant tax Labour campaigned on no new taxes but straight after the 2020 e "
[6] "Since 2017 food prices have skyrocketed 43 now 8 32 42 now 5 34 40 now 12 42 32 "
>
> negative_National.tweets <- word_National.df[sent_National.value < 0]
> head(negative_National.tweets)
[1] "Almost 20 000 kiwis voted in the first 24 hours on what has been Labour s biggest fail Vote here "
[2] "More than 20 000 Kiwis have already signed our petition to stop Labour s Jobs Tax This latest tax would take 834 "
[3] " Help us stop Labour from banning school drop offs and pick ups without consultation in its Reshaping Streets pr "
[4] "Labour s wasteful spending and economic mismanagement are driving up prices "
[5] "A National Government will do more to steer young people away from a life of isolation and dependence on welfare a "
[6] "A National Government would scrap the Auckland Regional Fuel Tax Labour s plans for an Auckland Light Rail Tax "
>
> neutral_National.tweets <- word_National.df[sent_National.value == 0]
> head(neutral_National.tweets)
[1] ""
[2] ""
[3] "Chris Luxon and Sam Uffindell on the campaign trail in Tauranga "
[4] "" 12 30pm ""
[5] ""
[6] "Average household costs could be up 150 per week by the end of this year when inflation is running laps around wa "

```

## Visual representation of sentiment scores

By processing the data from all three variables, we are able to quantify the number of tweets corresponding to each sentiment category. This allows us to create a pie chart that visually represents the distribution of tweets categorised as positive, negative, or neutral.

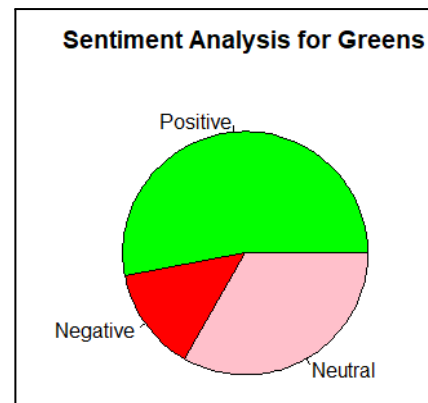
- Firstly, we count all the tweets which are positive, negative, or neutral. As a result, using these counts stored in **pos**, **neut**, and **neg**, we can generate a pie chart.

pos_Greens	35L	neg_Greens	9L	neut_Greens	22L
pos_labour	71L	neg_labour	9L	neut_labour	10L
pos_National	32L	neg_National	32L	neut_National	10L

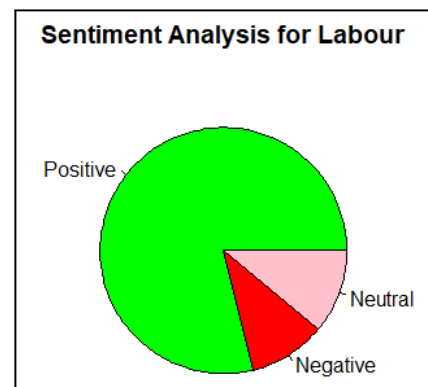
- Next, I chose my colours and combined the values to then assign it to a vector. And then state the labels used in the pie chart.

- After that, we can plot the pie chart for all the 3 parties:-

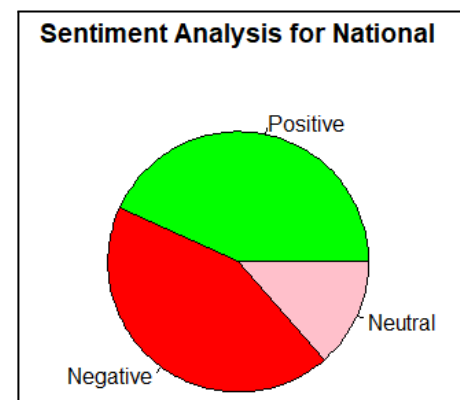
The Greens: As we can see, there is a larger proportion of positive tweets (green slice) compared to negative and neutral tweets. This suggests that the Green Party's twitter tends to be more positive in tone and sentiment.



Labour: Similar to the Greens, the pie chart for Labour shows a larger green slice, indicating a higher percentage of positive tweets. However, the proportion of negative tweets (red slice) appears slightly larger compared to the Greens' chart. Although, the neutral tweets are still slightly higher than the negative tweets.



National: In the case of the National Party, the pie chart displays a more balanced distribution between positive (green) and negative (red) tweets. However, there is relatively a higher presence of the negative sentiments as shown in the red part compared to the green slice.



## Finding the most positive & negative tweets

To help us find the most positive and negative tweets in our dataset, we use the `sent.value` variable.

```
> most_Greens.positive <- word_Greens.df[sent_Greens.value == max(sent_Greens.value)]
> most_Greens.positive
[1] "we need to make it clear and understandable again the simple steps people can take to keep others safe   Everywher  "
>
> most_labour.positive <- word_labour.df[sent_labour.value == max(sent_labour.value)]
> most_labour.positive
[1] "we re in a good position to continue to support Kiwis   and take advantage of the opportunities ahead   Read more ab  "
>
> most_National.positive <- word_National.df[sent_National.value == max(sent_National.value)]
> most_National.positive
[1] "we are forever grateful for those who have served New Zealand fighting for the ideals of freedom  democracy  and pe  "
>
> # Selecting and displaying the most negative sentiment (lowest sent.value)
> most_Greens.negative <- word_Greens.df[sent_Greens.value == min(sent_Greens.value)]
> most_Greens.negative
[1] "The pandemic is not over   Forty people have died from COVID in the last week alone   nearly 35 000 cases   and 418 ho  "
>
> most_labour.negative <- word_labour.df[sent_labour.value == min(sent_labour.value)]
> most_labour.negative
[1] " 2 3   The country s debt is low   GDP is up   our exports are in demand   and we re seeing more people in work   with higher wages  "
>
> most_National.negative <- word_National.df[sent_National.value == min(sent_National.value)]
> most_National.negative
[1] "Gang membership is violent crime is  "
```

## Creating a word cloud using a TDM(Term Document Matrix)

A Term Document Matrix is a mathematical matrix that represents the frequency of terms (words) across the documents (tweets) in the corpus. We analyse tweets to build a list of the most common words used, starting with the most popular. After that, we can finally plot a word cloud with the words that occur at least 3 times.

- First we create a corpus, i.e., a collection of text documents from a data frame called `word.df`. The `Corpus` function from the `tm` package is used to do the same.
- Then we use the `control` argument to apply transformations to the text, like converting all texts to lowercase, removing punctuations, filtering out words based on length, clearing stop words, and removing numbers.
- After that, we have to convert the TDM object into a matrix format to help the calculation of word frequencies easier. The `sort` function is used to order the word frequencies in decreasing order.
- Next, from the `word_freqs` vector a data frame `dm` is created that contains the words (`word` column) and their corresponding frequencies (`freq` column).
- The `dm$word` and `dm$freq` vectors provide the words and their corresponding frequencies, respectively. The `min.freq` argument specifies that only words appearing at least 10 times will be included in the word cloud. The `random.order=FALSE` argument ensures that the words are placed in the word cloud in a deterministic order.
- From the `RColorBrewer` package, the `colors` argument specifies a colour palette which is used for the word cloud.

Word cloud for the NZ Green Party, where the 3 most frequently occurring words are- Public, Party, and Green.



Word cloud for the NZ Labour Party, where the 3 most frequently occurring words are- Today, Zealand, People.



Word cloud for the NZ National Party, where the 3 most frequently occurring words are-  
Living, Labour, Kiwis.



## **Conclusion**

The sentiment analysis of the tweets from the official Twitter accounts of The Labour Party, the National Party, and the Green Party revealed some interesting insights:

1. **Sentiment Distribution:** The sentiment analysis for the tweets showed us that the Green's Party has been the most neutral on Twitter. However, the Labour Party had a higher proportion of positive tweets as compared to the Greens Party and less neutral tweets. Out of all the 3 parties, the National Party had the most negative tweets, with a similar amount of positive tweets as The Greens Party.
2. **Emotional Content:** According to the analysis based on emotional content, we can see that there were a varied range of emotions expressed in the tweets of the three respective parties. This means that the parties use a lot of emotional pleas in their Twitter communication.
3. **Most Positive and Negative Tweets:** The analysis identified the most positive and negative tweets for each party which helps us provide valuable insights into the types of tweets perceived as particularly positive or negative by the general public/twitter users.
4. **Word Frequency:** The word clouds revealed the most frequently occurring words in the tweets from each party. For the Green Party, the words "public", "party", and "green" were most frequent. The most frequent words for the Labour Party, the words "today", "zealand", and "people". For the National Party, the words "living", "labour" and "kiwis" were most frequent.

In conclusion, the analysis provides an insight of the sentiment towards the three political parties based on their Twitter activities.