Sentiment Analysis

Introduction

Sentiment analysis is a text mining technique that aims to extract the thoughts and feelings of script to determine their polarity, i.e. positive, negative or neutral. Three approaches are available to conduct sentiment analysis: supervised, lexicon-based and hybrid. The supervised method utilizes machine learning algorithms to train the classifier. It is superior in performance to the lexicon-based method, however it requires a substantial amount of labelled data. Lexicon-based sentiment analysis uses lexicons (dictionaries) to describe polarity. This method is also computationally efficient, but the results may vary depending on the lexicon and domain. A word may be subjective or objective depending on the context, e.g. in the clause "crude oil", this is an objective use of the word *crude*; when it is used as "crude language", it is now subjective and has a negative sentiment. Dealing with negation and sarcasm is also a challenge with this approach. The hybrid method is an amalgamation of the supervised and lexicon-based methods (Sadia et al, 2018).

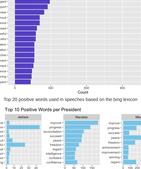
This section will discuss the results obtained from employing the lexicon-based approach to determine the outlook of the South African Presidents. The *bing* and *nrc* lexicons were explored. In the instance that a word was not in the lexicon, the default label was set to "neutral". The *bing* lexicon was developed by Minqing Hu and Bing Liu as the *Opinion Lexicon*. It comprises of 6786 words, where 2005 are "positive" and 4781 are "negative".
The *nrc* lexicon has 13872 words and incorporates more sentiments in addition to positive and negative: *anger, anticipation, disgust, fear, joy, sadness, surprise* and *trust*.

Words in the *nrc* lexicon

Category	No. of Words
anger	1245
anticipation	837
disgust	1056
fear	1474
joy	687
negative	3316
positive	2308
sadness	1187
surprise	532
trust	1230

Word-Level Analysis

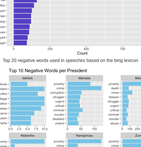
In this subsection, a word-level analysis will be conducted. The top 20 positive and negative words in the speeches as a whole, as well as a breakdown per president using the *bing* lexicon are presented below:



Top 20 positive words used in speeches based on the bing lexicon

Top 20 positive words by president based on the bing lexicon

Overall, the word **improve** was the most used positive word in all of the speeches. This is also true when the words are grouped by president, except for deKlerk, whose most used word was **freedom**. Other recurring words include **progress** and **peace/peaceful**.



Top 20 negative words used in speeches based on the bing lexicon

Top 20 negative words by president based on the bing lexicon

Poverty, **crime** and **corruption** were the most commonly used negative words. They also appeared in the breakdown by president, except for deKlerk, who did not have any of these three words in his top ten negative words. deKlerk commonly used the words **violent**, **illegal** and **discrimination**.

Sentiment-Level Analysis

The general feeling of the speeches, broken down by sentiments, for each president were examined using both the *nrc* and *bing* lexicons. The neutral words were excluded, as they dominate over the other sentiments.

Proportion of positive to negative words by president based on the bing lexicon

The proportions of positive to negative words do not seem to differ by a large amount. However, the plots indicate that **Zuma**, **Mbeki** and **Mandela** expressed relatively more positive sentiments in comparison to the other presidents. To further delve into the presidents' attitudes, the more comprehensive *nrc* lexicon was applied.

Proportion of positive to negative words by president based on the bing lexicon

From the plots, it can be observed that, besides *negative* and *positive* sentiments, prevalent themes include *trust* and anticipation. To gain some insight of the words specifically associated with these themes, the top 2 words with *trust* or *anticipation* for each president are presented in a table below.

Top 2 words by president associated with either *anticipation* or *trust* according to the *nrc* lexicon

President	Word	nrc	bing
deKlerk	constitutional	trust	neutral
deKlerk	freedom	trust	positive
Mandela	public	anticipation	neutral
Mandela	nation	trust	neutral
Mbeki	public	anticipation	neutral
Mbeki	continue	anticipation	neutral
Mbeki	continue	trust	neutral
Motlanthe	public	anticipation	neutral
Motlanthe	system	trust	neutral
Ramaphosa	economy	trust	neutral
Ramaphosa	public	anticipation	neutral
Zuma	continue	anticipation	neutral
Zuma	continue	trust	neutral

Most of the words associated with *trust* or *anticipation* in the *nrc* lexicon are categorised as "neutral" in the *bing* lexicon. This highlights the variability that different dictionaries can yield in an analysis. Additionally, it is important to note that due to the multi-level nature of words in the *nrc* lexicon, some words are associated with more than one sentiment, such as **continue**.

Changes of the sentiments over time were also investigated. Since there is a large number of neutral words, these were excluded from the analysis. Line plots with corresponding smoothed lines were generated to assess any shifts in positive or negative sentiments.

To assess whether the shifts are statistically significant, a logistic regression model was employed. Logistic regression is a modelling technique, based on regression, in which the dependent variable is binary. This is an appropriate model to use because the dependent variable has a binary outcome: positive or negative.

Proportion of positive to negative words by president based on the bing lexicon

Results from applying logistic regression

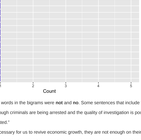
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0	1.0694006	0	1
date	0	0.0000745	0	1

From the plot, it appears that there has generally been a decline in negative sentiment over the years.
The p-values obtained from the logistic regression, however, are above the traditional threshold of 0.05, indicating that there was no significant decrease in the proportion of negative sentiments over time.

Bigrams Analysis

An n-gram is a sequence of n words in a text. Bigrams, in particular, are pairs of words that are adjacent to each other. Certain words, such as "not" change the meaning of the subsequent word by negating it. It is therefore important to consider this when conducting sentiment analysis.

This subsection section will explore the most common positive and negative bigrams in the speeches using the *bing* dictionary. To determine the net sentiment of a bigram, positive words are assigned a value of 1 and negative words are assigned a value of -1. Words that are preceded by a negation word (*not*, *no*, *never*, *without* or *lack*) had their sentiment reversed. During the analysis, a few of the labels given to words in the *bing* dictionary did not align with this context and were thus adjusted. Consequently, 'honour' and 'hope' were changed from *neutral* to *positive*. The word 'oppression' was reclassified from *neutral* to *negative*.



The findings illustrated in the plots partially align with what was observed in the top 20 positive and negative words. In the case of positive words and bigrams, several words appeared in both sets, such as **improve**, **support**, and **progress**. For the negative bigrams, the main themes were similar as before: **poverty**, **crime** and **corruption**. Interestingly, none of the negative bigrams contained negation words; these were subsequently extracted separately.

The most frequently occuring negation words in the bigrams were **not** and **no**. Some sentences that include the "not enough" bigram are:

- "But the response is that not enough criminals are being arrested and the quality of investigation is poor."
- "Not enough jobs are being created."
- "While structural reforms are necessary for us to revive economic growth, they are not enough on their own."

These examples highlight the importance of taking negation words into account as they reverse the sentiments of the sentences.

Discussion

In this section, a sentiment analysis of the State of Nation Addresses was conducted to discern the overall feelings expressed by the South African presidents over time. It was discovered that most of the presidents, with the exception of deKlerk, had common interests: peace, freedom and progress. Recurring concerns were centred around poverty, crime and corruption. The analysis indicated that concerns over these themes did not change over time, as the logistic regression model indicated that the sentiment trend was consistent. It is worth noting that this analysis had some limitations, such as the under-representation of some presidents and the subjective nature of lexicons.

Use of Large Language Models

An example of applying ChatGPT to generate code was requesting for a function that extract sentences containing specific bigrams. Before providing much context, the response given was steps to manually do it in a programming language. When prompted for code, ChatGPT assumed that it was being done in Python. To provide context, the programming language as well as the structure of the tidy data was provided as a response. R code was subsequently returned and it worked exactly as expected. Comments were also provided to guide the user.

```
# Define the specific bigram
specific_bigram <- "specific bigram" # Replace with your desired bigram

# Initialize a list to store sentences containing the bigram
sentences_with_bigram <- character(0)

# Search for the bigram in each sentence
for (sentence in sona_sentences$sentence) {
  if (grepl(specific_bigram, sentence, ignore.case = TRUE)) {
    sentences_with_bigram <- c(sentences_with_bigram, sentence)
  }
}

# Print or manipulate the extracted sentences
print(sentences_with_bigram)
```

References

Sadia et al (2018). https://ieec.neduet.edu.pk/2018/Papers_2018/15.pdf