A Descriptive Analysis of the South African Presidential State of the Nation Address (SONA) - 1994 to 2023

Ropafadzo Chimuti Tanweer Nujjoo Steven Ellis

The annual State of the Nation Address (SONA) in South Africa serves as a critical national update by the President, delivered to Parliament. This research examines SONA speeches from 1994 to 2023, accessed from the official government website.The study aims to conduct sentiment analysis and topic modeling on these speeches, utilizing analytical skills such as data cleaning, sentiment analysis, Latent Dirichlet Allocation (LDA) for topic extraction, and trend identification.This paper provides insights into how sentiments and key topics have evolved over time, offering a comprehensive view of South Africa's political and societal landscape from 1994 to 2023. {The following findings were discovered }

Table of contents

1 Introduction

The State of the Nation Address (SONA) is a crucial annual event in the political landscape of South Africa. It serves as a comprehensive summary of the nation's current status and a roadmap for its future. The address, delivered by the country's highest-ranking official, outlines achievements, priorities, and strategies for the upcoming year, significantly influencing policy, legislation, and public discourse.

This research paper, undertakes a comprehensive descriptive examination of the content within South African State of the Nation Addresses (SONAs), employing advanced techniques in sentiment analysis and topic modeling. Sentiment analysis allows for the extraction and quantification of emotional and attitudinal tones within the speeches, providing deeper insights into the language used by leaders to engage, motivate, and influence their audience. Understanding prevailing sentiments also offers valuable context for assessing the reception and impact of the SONA.

Furthermore, topic modeling enables the categorization of the diverse subjects addressed within these speeches. This approach clarifies the key areas of government focus and the shifting priorities as they evolve over time. It not only aids in identifying primary trends but also provides a quantitative basis for comparisons between different presidencies.

A comprehensive explanation of sentiment analysis and topic modeling, including a breakdown of their methodologies and an examination of the results as applied to the analysis of South African SONAs, are presented in relevant sections within this report.

2 Materials and Methods

This section describes thoroughly the procedures undertaken to conduct a descriptive analysis of the content of speeches using sentiment analysis and topic modelling. The entire implementation of this task was performed using RStudio, therefore the functions and libraries used throughout the process corresponds to R. However, the equivalence of this analysis can definitely be replicated on other programming platforms. There are various ways to wrangle data and only the "not so obvious" functions were explained in this report. It should be further noted that all the plots produced in this report were generated via the `ggplot()` function from the `ggplot2` library.

2.1 Brief Overview of SONA dataset

The dataset used in this study comprised of speeches obtained from various text files, each of differing sizes. These speeches were presented in a semi-formal structure, with a common characteristic being the inclusion of the date of delivery at the beginning. The data was not inherently structured in a manner conducive to automated tabulation and analysis. Therefore, preprocessing steps were necessary to facilitate subsequent analysis. Detailed information regarding the preprocessing procedures is provided in the following section.

2.2 Data Pre-processing & Cleaning

Raw data retrieval tends to always be messy based on the overview given about the SONA dataset in the previous section. So, tabulating the text files needed some pre-processing where the years at which the speeches occurred were extracted by identifying the first 4 strings from the filenames and attributed them to a new column. Additionally, the names of the presidents were extracted from the filenames. Within the process of extracting the presidents' names, string manipulation was performed to remove unnecessary regular expressions. Although, the dates were not specifically used

in our analysis, they were parsed in a new column for our own perusal. All the unnecessary regular expressions like "`(http.*?(...`" from the speeches were also removed. All those manipulations were done using the `stringr` library. Now that we were only dealing with words, we had to make sure that all the speeches were converted to lower cases and lemmatised to avoid any redundancies. The function used to perform the lemmatisation task, was `lemmatize_strings()` from the `textstem` library. Finally, the data was converted into a tibble.

2.3 Central Preliminary Procedure 1

As an initial step prior to any analyses in the whole methodology section, if the tokenisation (refer to Section 2.4 for an explanation on the concept of tokenisation) involved:

1. words, only the lowercase words would first be detected using the matching pattern [**a-z**] and filtered. Then, the stop words (i.e, prepositions and connecting words) from the SMART lexicon of the **tidytext** library would be removed as they do not convey valid information. In addition, some common and obvious words across all the speeches like "speaker", "madame", "honourable", "chairperson", "development", "national", "ensure", "deputy", "africa", "african", "africans", "south", "southern", "government", "people", "programme", "economic", "economy", "country" and "continue" which would act like noise in our analyses would be filtered out.
2. bigrams, each word would be separated into 2 different columns and then step 1 would be repeated for each column. Then the cleaned separated words would be united back in one column.
3. trigrams, each word would be separated into 3 different columns and then step 1 would be repeated for each column. Then the cleaned separated words would be united back in one column.

2.4 Overall Analysis

For any upcoming analyses, the speeches had to be tokenised accordingly to fit the purpose of our analyses. Tokenisation is the split of a sequence of characters in a text by locating the word boundaries (Palmer 2000). The atomicity of the split could be in terms of per characters, per words, per n-grams, per sentences and more. So, tokenisation is always application dependent. For our purpose, tokenising per words and per n-grams were relevant. This was simply done using the `unnest_tokens()` function from the `tidytext` library. n-gram is a terminology very well known in the world of natural language processing (NLP), and it simply refers to a sequence of n words. If n=1, it is referred to as a unigram, if n=2, it is referred to as a bigram and if n=3, it is referred to as a trigram. The overall analysis sought to reveal the top 20 words, bigrams and trigrams used by all presidents. The steps to achieve these, followed the main procedures highlighted in Section 2.3. Then for each case, the words, bigrams and trigrams were counted, sorted in descending order and sliced to the first 20 elements. Furthermore, they were all plotted as barplots.

A more in-depth overall analysis was performed by aggregating the above per president. However, only the overall trigrams used per presidents were analysed. The words and bigrams ones were omitted as the idea was, if we were to use more than a word anyways to do more in-depth analysis, we might as well use the trigrams as those would convey more meaningful information. So, the exact methods explained in the above paragraph were implemented with the exception that the presidents were first filtered in that same procedural pipeline.

2.5 Central Preliminary Procedure 2

This procedure is sentiment-focused rather than use for an overall analysis purpose. Nevertheless, it follows similar steps as in Section 2.3 with slight updates. Note that the words and bigrams part were omitted here based on the aforementioned argument in Section 2.4. As usual, an initial step prior to any sentiment analyses, if the tokenisation involved:

1. words, only the lowercase words would first be detected using the matching pattern [**a-z**] and filtered. Then, the stop words (i.e, prepositions and connecting words) from the SMART lexicon of the **tidytext** library would be removed as they do not convey valid information. The same common and obvious words across all the speeches as detailed in Section 2.3 would act like noise in our analyses, hence would be filtered out. The relevant dictionary (further description about dictionaries is detailed in Section 2.6) would be left joined to the cleaned dataset on the words so that the words in the speeches would have a sentiment attached to it. Obviously, not all the words in the speeches would be present in the dictionaries. Therefore, words that did not have a label would be mutated to "neutral".

2. trigrams, each word would be separated into 3 different columns and then step 1 would be repeated for each column. Then the cleaned separated words would be united back in one column. Because we would deal with trigrams, the sentiments would have to be polarised (i.e, "positive" = 1, "neutral" = 0 and "negative" = -1) in order to be able to calculate a final sentiment score for all the trio of words. That would be based on the sum of all polarised sentiment belonging to each trio of words. Moreover, if the final sentiment score = 0, the trigrams would remain as "neutral". If the final sentiment score ≥ 1, the trigrams would be "positive" and if the final sentiment score < 1, the trigrams would be "negative".

2.6 Sentiment Analysis

According to Medhat, Hassan, and Korashy (2014), sentiment analysis also known as the opinion mining study people's opinions, attitudes, and emotions towards an entity in a computational manner and can also be considered as a classification process. Our sentiment analysis took a lexicon dictionary-based approach which is considered as a very feasible approach as it does not involve any training data and advanced machine learning techniques (Wankhade, Rao, and Kulkarni 2022). For this reason, some experts like Yan-Yan, Bing, and Ting (2010) also referred to this approach as an unsupervised approach. The 2 main dictionaries used in our analyses were `bing` and `nrc`.

The `bing` dictionary was loaded from the `tidytext` library. Each of the 6786 words in the dictionary is assigned to a binary sentiment of either positive or negative. The main disadvantage of the lexicon approach is that it is highly domain-oriented. Indeed, during some analysis, words like "anti poverty", "anti corruption", and so on were classified as negative sentiment as the word "anti" did not exist in the `bing` dictionary. Therefore, this issue was resolved by adding the word "anti" in the `bing` dictionary and assigned a positive sentiment to it as this label fits the context of this analysis. This part of the sentiment analysis explored the top 20 positive and negative trigrams used by all presidents. The steps to achieve these, followed the main procedures highlighted in Section 2.5. Then for each case, the trigrams were filtered by either "positive" or "negative" sentiment, followed by a count and a sort in descending order and eventually sliced to the first 20 elements. Furthermore, they were all plotted as barplots.

A more in-depth sentiment analysis was performed by aggregating the above per president. The exact methods explained in the above paragraph were implemented with the exception that the presidents were first filtered in that same procedural pipeline. It was also vital to investigate the variation of positive and negative sentiments over the years while excluding deKlerk and Motlanthe as they were only 1-term presidents. Those investigations were done for the trigrams by following the procedure as described in Section 2.5 and then plotted as barplots and line graphs. The only key aspect here

was to group the sentiment and year before counting how many positive and negative sentiments were recorded per year.

This report also employed the **NRC Emotion Lexicon** for sentimental analysis, a resource that associates English words with eight fundamental emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust), along with two sentiments (negative and positive). This lexicon was used to analyse the emotions expressed in presidential speeches, providing insights into the emotional aspects of their leadership and communication.

2.7 Topic Modelling

Topic modelling is a machine learning technique that scans tokens (usually words) in sequences (usually documents), with the aim of detecting word and phrase patterns within the sequences. This is achieved by clustering word groups and similar expressions together into 'topics' that best characterize a set of documents. (Pascual 2019)

Topic models can help offer insights into bodies of text, for better understanding of large collections of unstructured text bodies. (Blei 2012)

In this assignment topic modelling was performed as follows:

A data-set of words that appeared in the corpus of presidential speeches was extracted into a **document-term matrix**, containing rows corresponding to the speeches and columns corresponding to the words.

The **Latent Dirichlet allocation (LDA)** method was then used to extract topics from this sparse and wide document-term matrix. The LDA algorithm is guided by two principles:

- Every document is a mixture of *topics* (for example 90% topic A and 10% topic B) and every topic is a mixture of *words*.
- Every topic is a probability distribution over words.

LDA draws topic distributions from a Dirichlet distribution and uses variational expectation maximisation algorithm, and collapsed Gibbs sampling (both beyond the scope of this assignment) to achieve topic model estimations.

The **LDA()** function from the `topicmodels` package was used to retrieve LDA topic models.

Latent Dirichlet allocation Topic modelling was performed in the following way:

- The top 5 topics were retrieved across all speeches (to retrieve president-specific topics).
- The `tidytext` package was then used to extract **per-topic-per-word probabilities** (the "beta") from the model.
- The **per-president-per-topic probabilities** (the "gamma") were then extracted using the `tidytext` package.
- The log ratio of beta values of topics for different presidents were compared to reveal words with the greatest differences between the two presidents' topics.
- Finally, the speeches of Ramaphosa were retrieved, and the same exercise was repeated for Ramaphosa-specific speeches (to gain insights into the topics that permeated his speeches over the years).

3 Results & Discussions

3.1 Overall Analysis

An overall analysis was executed on the speeches to extract the top 20 words, bigrams and trigrams used by all the presidents to obtain an idea about which of those conveyed meaningful information.

Figure 1 represents the top 20 words, bigrams and trigrams used by all presidents. Plot **A** showcases that all presidents would obviously address the public most of the time. Moreover, the top words were linked to development and were very businessy. Plot **B** were way more meaningful than plot **A** and it seemed like majority of presidents were addressing private sectors more than public sectors in their speeches. They addressed the unity of nations, job creations, justice system and specific locations in South Africa. Mandela seemed to have been a model to the other presidents being the first anti-apartheid activist and president of the country, hence mentioned several times. The 2010 FIFA world cup that was held in South Africa brought such a positive and festive vibe in Cape Town, no wonder why it was in the top 20 bigrams. Plot **C** was even more meaningful and we could observe that the bigram "justice system" and "criminal justice" were the least occurred words in the top 20 bigram plot. That was because those particular bigrams were not that meaningful, as such plot **C**

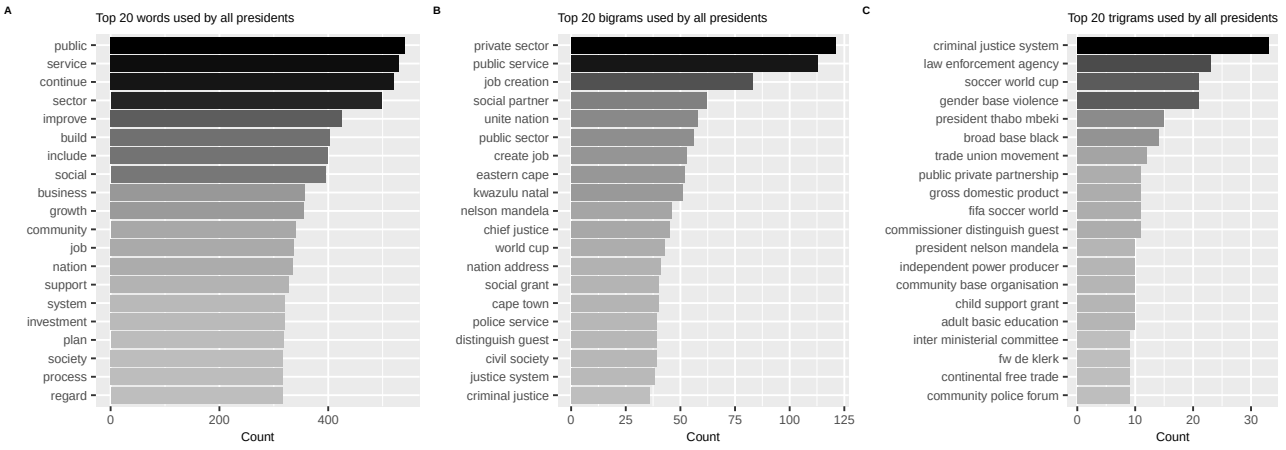


Figure 1: Top 20 words, pair of words (bigrams), and trio of words (trigrams) used by all presidents.

revealed the real meaning of the words being criminal justice system, hence that trigram had the highest occurrence. In addition, a lot of projects like trade union movement, public private partnership, child support grant, adult basic education, continental free trade, community of forum and more were the subject matter amid presidents. The economy of the country was also address in terms of the gross domestic product (GDP), however, problems like gender-based violence outweighed the number of occurrences for the country's GDP, and this priority was valid here.

Figure 2 displays the top 20 trigrams per president. Mandela's trigrams clearly indicated the challenges and priorities, he was dealing at his time of presidency. Without re-emphasising on the trigrams, here's a brief overiew. Mandela's commitment to address social issues, improving public services, maintaining law and order, promoting workers' rights, and ensuring a strong constitutional framework has not gone unnoticed. Once again, his presidency was characterized by a focus on reconciliation and inclusivity, as well as engagement with international matters. Overall, Mandela's presidency was marked by his dedication to nation-building, social progress, and upholding democratic values. Due to the fact that deKlerk was only a 1-term president, all his top words had only a count of 1. Once again, the word constitution appeared multiple times. Indeed, he was dealing with a period of significant political transformation in South Africa. He was therefore working on constitutional negotiations, peace-building, and addressing political and social challenges. His presidency was marked by efforts to move away from apartheid policies and towards a more inclusive and democratic system. Mbeki's

top tier words looked quite diverse. He appeared to address issues related to security, governance, economic development, and education. Moreover, the focus on major sporting events like the soccer world cup suggested a commitment to the country's international standing and promoting sports. Additionally, the reference to traditional leadership and community centres indicated engagement with South Africa's cultural and local governance dynamics. As per the top 20 trigrams for Zuma, he seemed to have focused on economic policies, infrastructure development, and addressing corruption. Apart from the social and economic programs, he had to host the 2010 FIFA world cup during his tenure. Motlanthe's trigrams was not that helpful as he was also only a 1-term president. His top 20 trigrams did not convey meaningful information. The only observation was that it appeared that he may have been dealing with a range of complex challenges, with a focus on social and economic development, justice and anti-corruption. Ramaphosa's trigrams sounded very promising and he seemed to have focused on addressing a wide range of issues, including social challenges like gender-based violence, economic development through trade and job creation, legal and constitutional matters, and engagement with various communities and organizations. His governance reflects a multifaceted approach with policy-making aimed at addressing both immediate and long-term challenges facing South Africa. An observation was made that across the trigrams per president, criminal justice system has appeared various time, which also means that crime is still a persistent problem in South Africa.

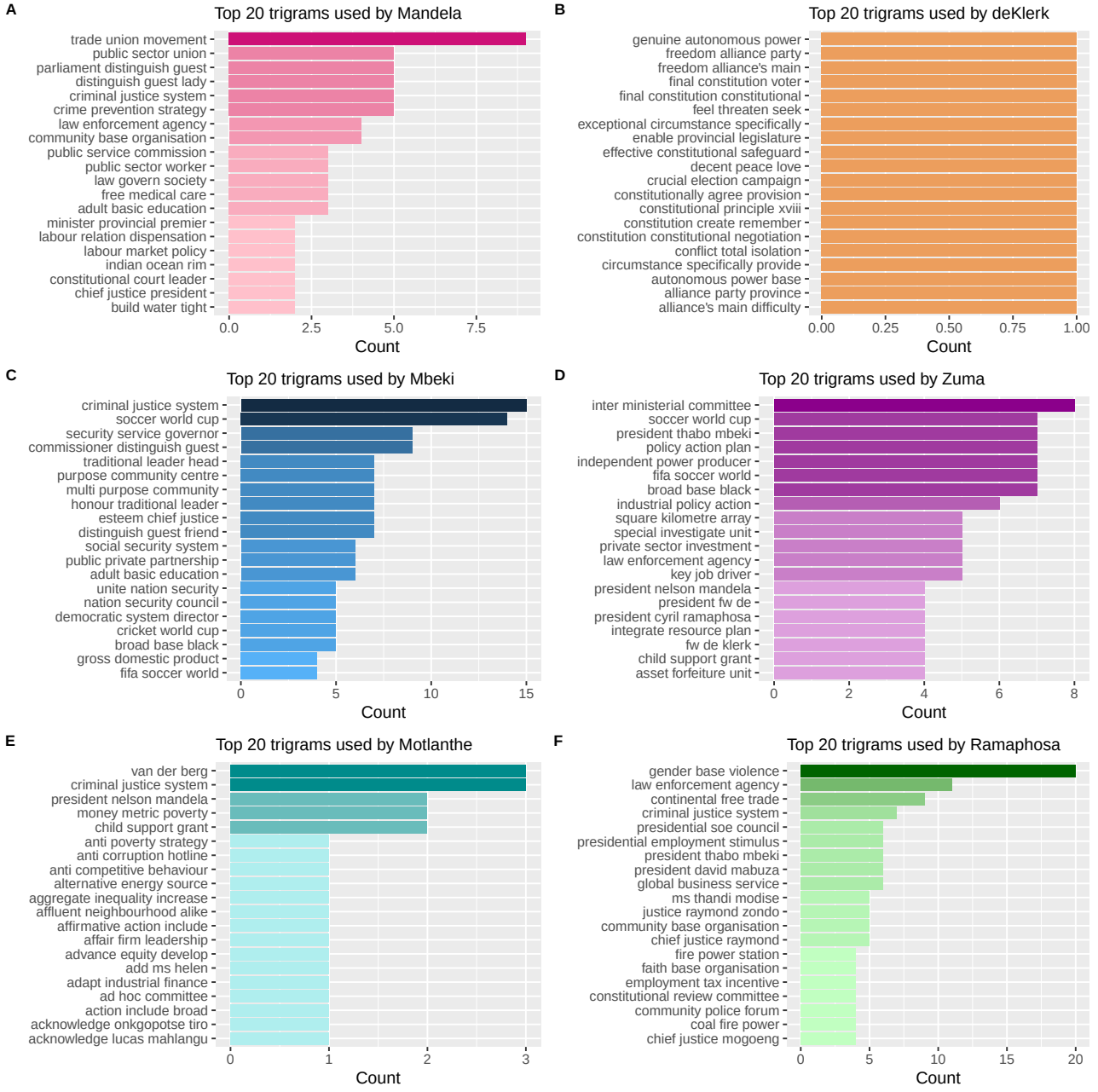


Figure 2: Top 20 trio of words (trigrams) used by each president. **A** relates to Mandela's trigrams, **B** relates to deKlerk's trigrams, **C** relates to Mbeki's trigrams, **D** relates to Zuma's trigrams, **E** relates to Motlanthe's trigrams, and **F** relates to Ramaphosa's trigrams.

3.2 Sentiment Analysis

Plot **A** of Figure 3 is a typical example of the presidents trying to improve the well-being of the local population, especially those who live in informal settlements suffering from basic amenities. So, the positive trigrams highlighted a range of policy areas, projects and achievements, including social welfare, economic development, public services, infrastructure, and international engagement.

For plot **B**, there were several words like gross domestic product, crime prevention strategy, sexual offence court, prosecute authority npa, sexual office unit, passenger rail agency, gross fix capital, fight organise crime, emergency task team, corruption task team, and corruption advisory council which should not have been categorised as negative sentiment. Nevertheless, the problem for crime seemed persistent which was why a criminal justice system needed refinement. Although the criminal justice system is a positive sentiment, we treated it as negative because if there were no crimes, it would be

pointless to have a criminal justice system. So, its existence entailed the unsolved problem of crime in the country. In fact, the same logic could be argued with regards to sexual offence court.

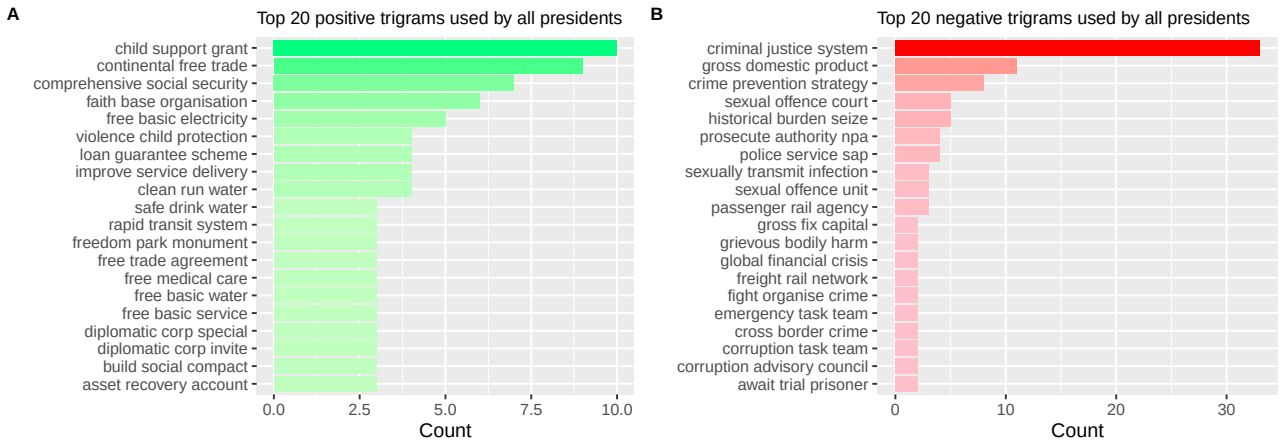


Figure 3: Top 20 positive (**A**) and negative (**B**) trio of words (trigrams) used by all presidents.

Figure 4 depicts the positive sentiment of trigrams per president. Mandela's positive sentiment focused on social welfare, economic development, social justice, reconciliation, and peace. He worked to create a more inclusive and just society while addressing a range of challenges, both domestic and international. He wanted to promote black empowerment as he was fighting for the abolition of apartheid. So, Mandela's commitment to building a post-apartheid South Africa that was democratic, equitable, and prosperous is still engraved in our heart. His main priority seemed to be the provision for free medical care. Since deKlerk was only a 1-term president, only 9 unique positive trigrams were obtained. His focus on building trust, finding practical solutions, and ensuring genuine autonomy and constitutional safeguards indicated a commitment to resolving issues related to apartheid, political reform, and peace-building. It was consistent with his role in initiating negotiations to end apartheid and lead South Africa towards a more inclusive and democratic future. The top 20 positive sentiment trigrams for Mbeki was focused on infrastructure and basic amenities development, social welfare, economic growth, and international cooperation (i.e, afro-asian solidarity). He definitely tried to promote unity and solidarity in South Africa. Zuma's positive sentiment trigrams showcased an all-rounded presidency with a focus on social welfare, diplomacy, economic development, public health, and governance. The positive sentiment trigrams from Motlanthe also being a 1-term president aimed to address a range of issues, including economic growth, social welfare, competitiveness, equitable development, and infrastructure improvement. His administration appeared to have focused on creating a more inclusive and socially responsible society, with an emphasis on social support, economic stability, and equitable access to healthcare. It was no surprise to see that continental free trade was the first positive sentiment trigram for Ramaphosa, as he is always very business-oriented. He was really dedicated to economic growth, social inclusivity, diplomacy, democratic governance, and addressing pressing issues such as land reform and healthcare. He seemed to also focus on collaborative efforts and comprehensive planning to address the nation's challenges and drive progress. Overall, it was observed that most presidents were promoting child support grant and free basic amenities and clean water.

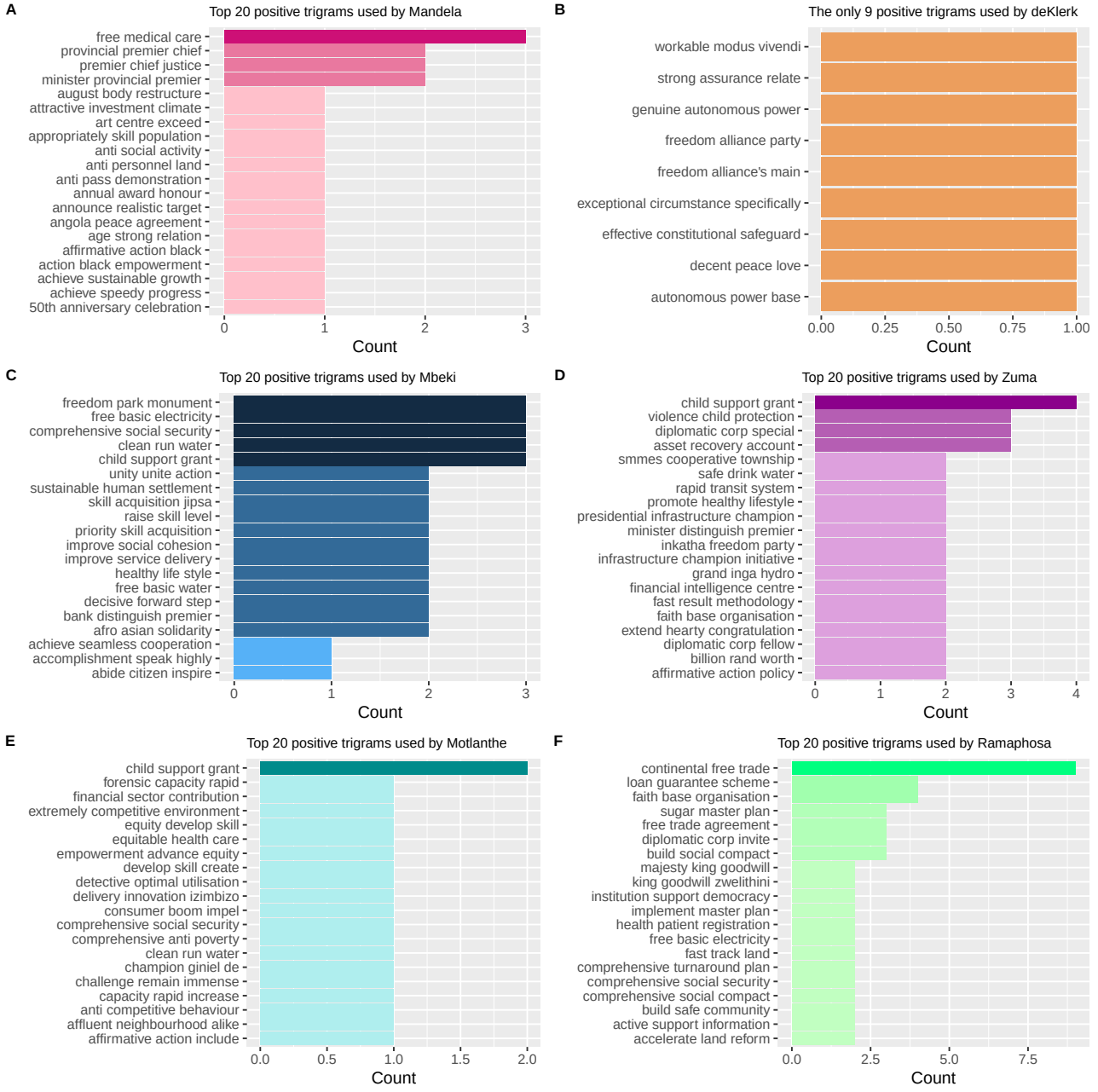


Figure 4: Top 20 positive trio of words (trigrams) used by each president. **A** relates to Mandela's positive trigrams, **B** relates to deKlerk's positive trigrams, **C** relates to Mbeki's positive trigrams, **D** relates to Zuma's positive trigrams, **E** relates to Motlanthe's positive trigrams, and **F** relates to Ramaphosa's positive trigrams.

Figure 5 illustrates the negative sentiment of trigrams per president. Mandela's presidency emphasised on inclusivity, human rights and social justice as mentioned before. During his time, he unfortunately had to deal with significant challenges, such as crime, corruption, land issues, child abuse and rape, poverty, and environmental concerns. The top 2 sentiments were rather positive, but crime related. Due to deKlerk being a 1-term president, he had only 5 unique negative sentiments trigrams which did not really convey much information apart from words like hopeless, hard negotiation, threaten, conflict and difficulty which only highlighted some challenges and hardships during his presidency without any specific context. Ignoring the trigrams which were supposed to have strictly positive sentiments like gross domestic product and post conflict reconstruction; the crime related matter kept on persisting with some other social matters, public health issues and other political challenges. Again, ignoring gross domestic product being a positive sentiment trigram, the crime issue were on top of the list during Zuma's presidency. Furthermore, he had to deal with corruption, economic challenges, infrastructural breakdown, and some social and environmental concerns. Motlanthe's was also a 1-term president and

from the trigrams, “poverty decline substantially”, “metric poverty decline”, “input price decline”, “hungry decline dramatically”, “gross fix capital” and “gross domestic product” should be ignored and treated as positive sentiment. This was a good example of the disadvantages of using lexicon-dictionary based approach for sentiment analysis. Nevertheless, crime, inequality, security measure issue, economic issues, political challenges and illicit activities were still the major concerns during his presidency. We again see some false negative sentiment present in the trigrams for Ramaphosa. Ignoring those, he still had to deal with crime related issues, violence in public, poverty, insanitation, water shortage, drug trafficking, debt perhaps due to the global pandemic which he also mentioned. It was quite clear that throughout all these year the crime issues, poverty, and violence were not able to eradicate or at a bare minimum to mitigate.

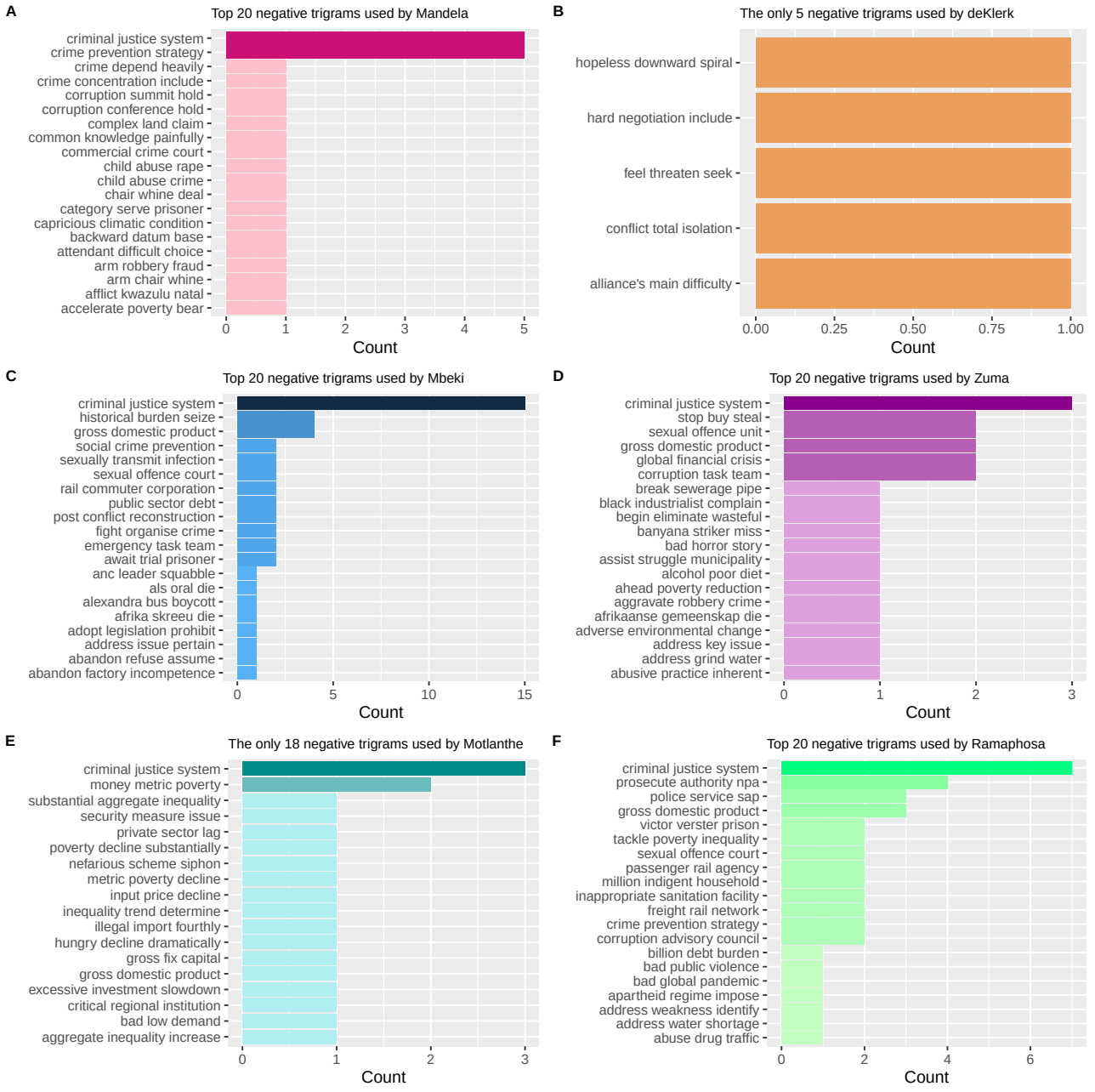


Figure 5: Top 20 negative trio of words (trigrams) used by each president. **A** relates to Mandela's negative trigrams, **B** relates to deKlerk's negative trigrams, **C** relates to Mbeki's negative trigrams, **D** relates to Zuma's negative trigrams, **E** relates to Motlanthe's negative trigrams, and **F** relates to Ramaphosa's negative trigrams.

So, when we investigated the sentiment over the years from Figure 6, we noticed that plot **A** had always more positive than negative sentiment and has increased over the years. Except that during

Mbeki's presidency in 2001, the negative sentiment for the trigrams surpassed the positive ones, perhaps because of the rise in HIV/AIDS in South Africa. Plot **B** shows the fluctuation in count of positive and negative sentiment throughout the years. It is worth noting that the positive sentiments were larger than the negative sentiments. Consider the positive sentiment, the fluctuations occurrences indicates the positive sentiment count in the previous year tend to be higher than the following year relative to the previous year. This means that the presidents were hopeful for country to implement the project as planned, however, in the following year they always have to deal with some societal or economic challenges which caused the drop in the positive sentiment. This observation was supported in Miranda and Bringula (2021) as well, meaning that in general that tends to be the case. Another interesting observation was that whenever there was a peak in negative sentiment, logically we would guess a drop in positive sentiment below the negative sentiment (i.e, more negative sentiment than positive sentiment). However, whenever the negative sentiment was very high, the positive sentiment also increased. That means whenever there's something bad happening, the politicians tend to cover it up with good or positive speeched to decrease tension and mask the actuality. Moreoverm the trends between the positive and negative sentiment had increased over the years and the sudden drop in positive and negative sentiment in 2020 was due to the global pandemic covid-19. So, as a final observation, it seemed as if over the years the South African presidents were improved and did better with an increasing trend in positive sentiment.

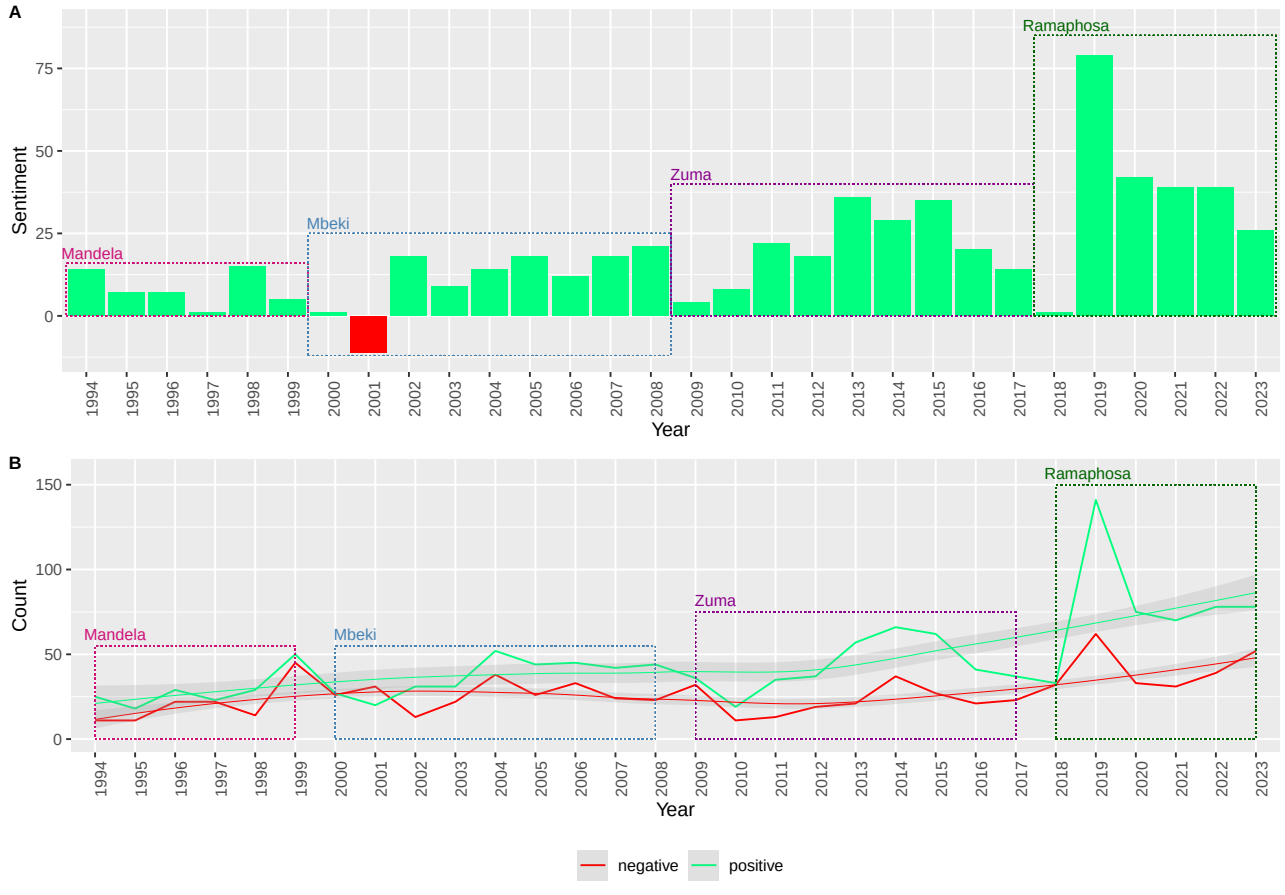


Figure 6: Positive and negative trio of words (trigrams) sentiment over time (excluding 1-term presidents such as deKlerk and Motlanthe). **A** corresponds to how many more positive than negative trigrams (or vice versa) computed over the years [1994-2023]. So, the y-axis for **A** was calculated as *Sentiment = positive trigrams - negative trigrams*. **B** refers to the actual count of positive and negative trigrams throughout the same mentioned period. The bounding boxes indicate which part of this timeline belongs to the presidency period of Mandela, Mbeki, Zuma and Ramaphosa.

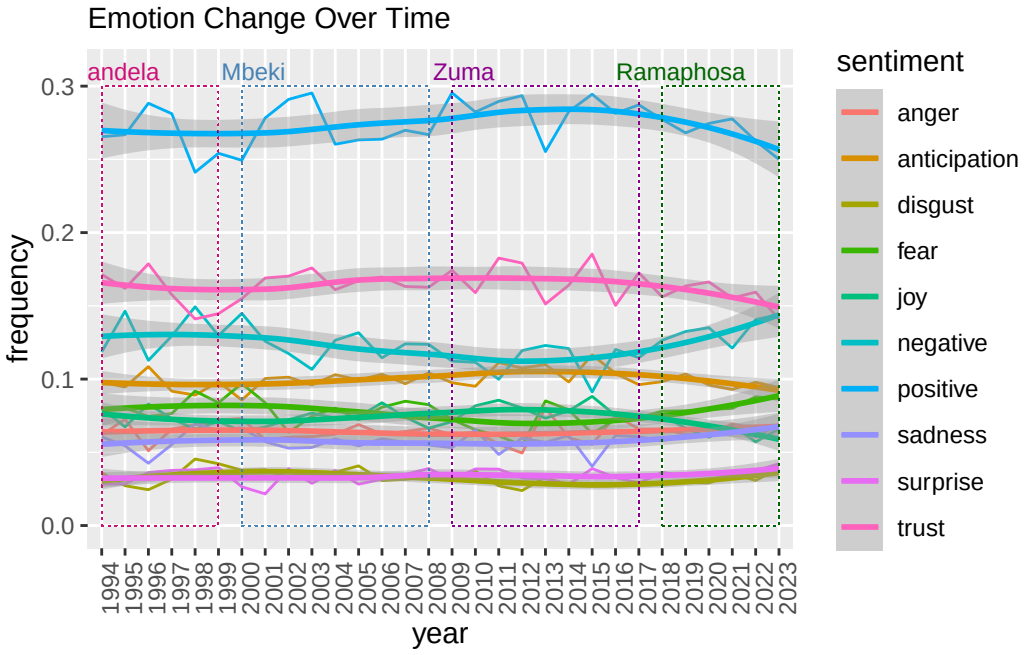


Figure 7: Evolution of emotions expressed over time from 1994-2023 by all presidents

In Figure 7, we observe the evolution of emotional dynamics over time. Notably, there is a distinct upswing in the expression of positive emotions, particularly from 2009 when President Zuma assumed office. This trend potentially signifies President Zuma's leadership style, which may have prioritized positivity, unity, and motivation. Alternatively, it could imply that his speeches served as a means of crisis management, offering reassurance to the public and reinforcing the notion of stability and control during challenging times. We can observe a significant increase in negative emotions during the time frame spanning 2021 to 2023, marked notably by heightened expressions of fear and sadness. This trend can be traced back to the profound global impact of the novel coronavirus, which affected nations worldwide in 2020.

3.3 Topic Modelling

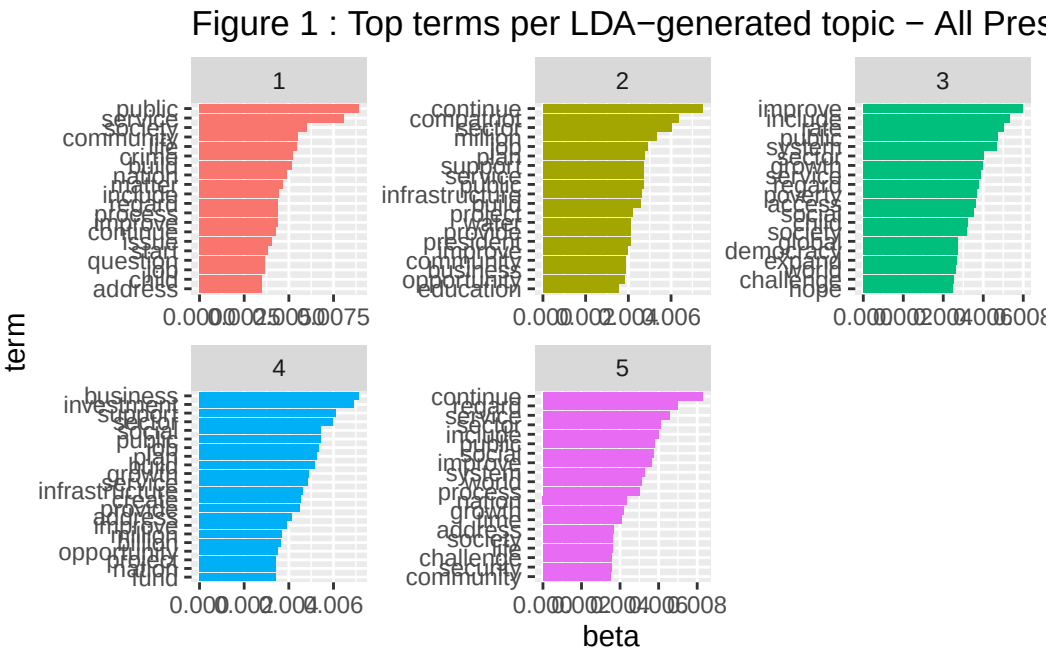


Table 1: President-Per-Topic Probabilities

President	Topic	Gamma
Mandela	1	1
Zuma	2	1
Motlanthe	3	1
Ramaphosa	4	1
deKlerk	5	1

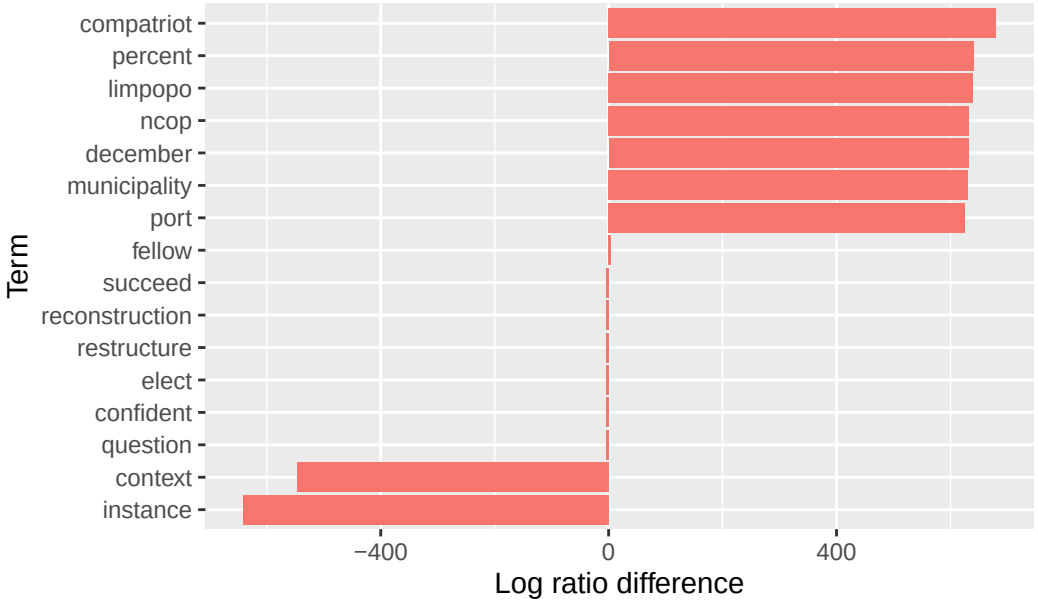
From the LDA topic modelling on all speeches, there is a clear probabilistic relationship between top 5 topics and which president each topic relates to.

Table 2: Topic Summaries

Topic	President	Example Words
Topic 1	deKlerk	constitution, change, improve, parliament, create
Topic 2	Mandela	nation, build, human, crime, job, address
Topic 3	Mbeki	system, world, growth, poverty, challenge
Topic 4	Ramaphosa	business, investment, build, growth, improve
Topic 5	Zuma	compratriot, water, education, community, plan

The comparison of the log ratio of beta values of different president topics revealed interesting insights into the eras characterizing their tenures, their priorities and the challenges each faced.

Figure 2: Log2 ratio of beta in topic 2 (Mandela) / topic

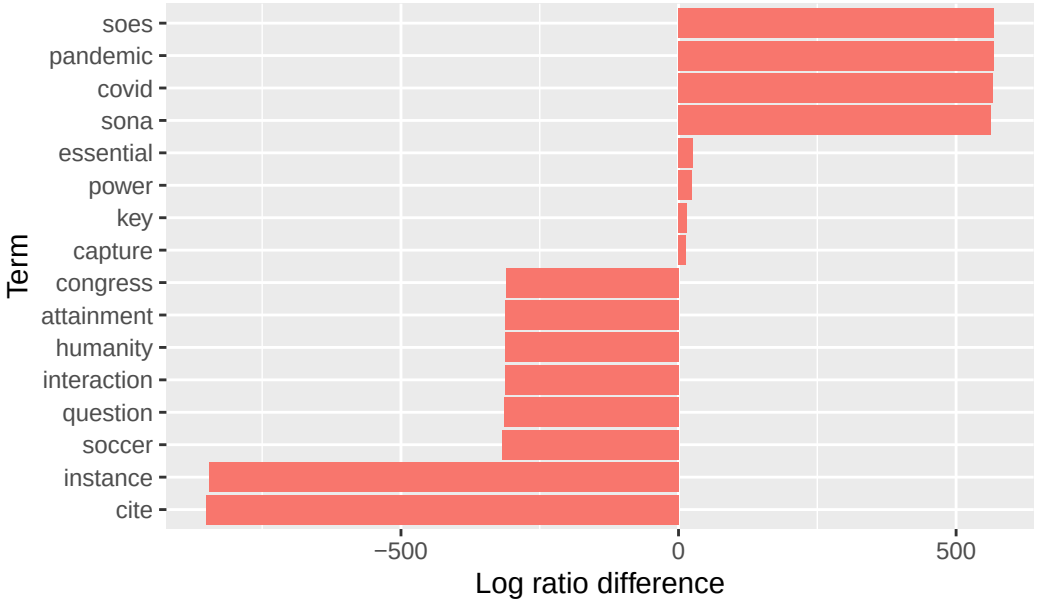


The difference in beta (per-topic-word probabilities) between **Mandela** and **deKlerk** are displayed in *Figure 2*.

deKlerk made a single speech just before the 1994 democratic elections, expressing hope and optimism for a cohesive and cooperative future between all South Africans. Mandela, by contrast, presided over South Africa for the first 5 years of its nascent democracy, and had to grapple and confront many of the social and economic challenges that it presented.

The log ratios between both presidents' topic beta's is very low compared to other subsequent beta comparisons, however, suggesting much commonality in the topics they spoke about and perhaps also their writing styles.

Figure 3: Log2 ratio of beta topic 3 (Mbeki) / topic 4 (Rama

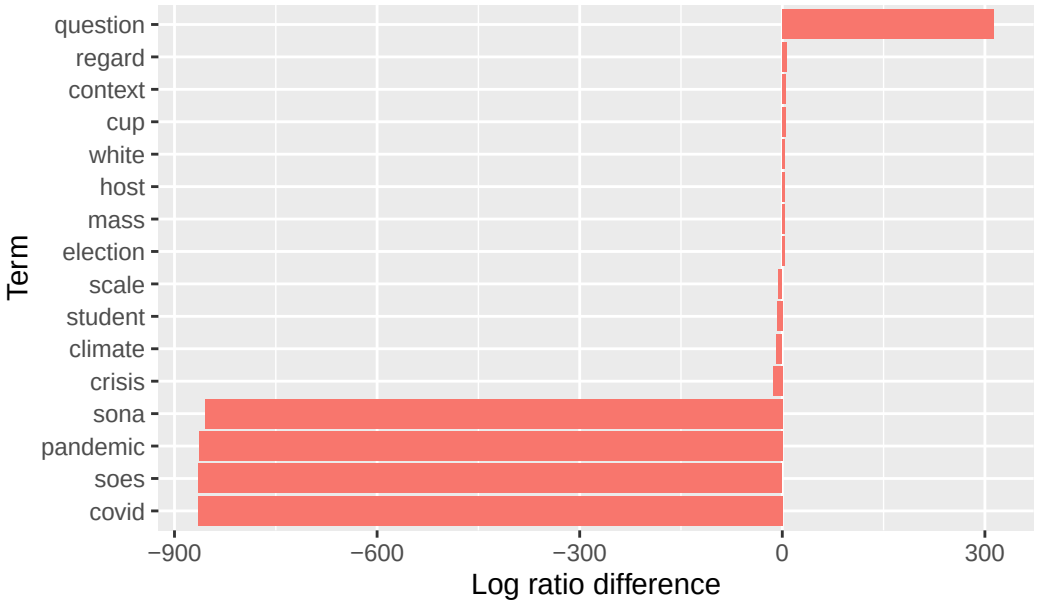


A comparison of the words with greatest differences between **Mbeki** and **Ramaphosa** reveal how they dealt with very different agendas during their respective tenures.

Mbeki presided over the hosting of the 2010 FIFA World Cup (`soccer`, `cup`) whilst Ramaphosa had to deal with a global pandemic in 2020-2021 (`covid`, `pandemic`, `restore`) as well as the Eskom electricity crisis which had started already during Mbeki's presidency (`shed`, `SOEs`, `restore`).

On a side note, Mbeki appeared to be very fond of the phrase `"discharge their responsibilities"` which is why the word `discharge` stands out in the comparison.

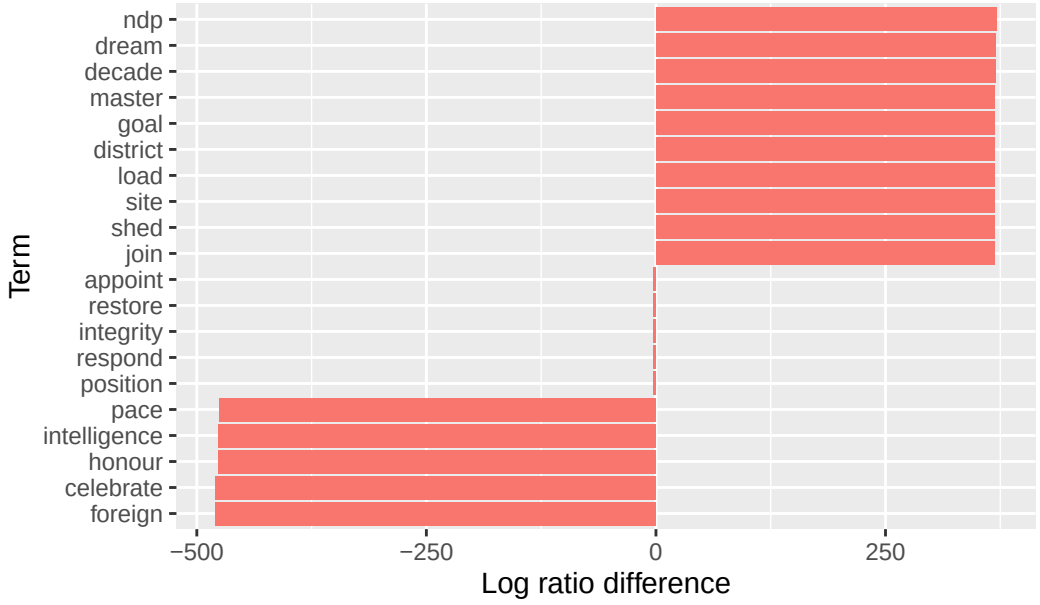
Figure 4: Log2 ratio of beta topic 4 (Ramaphosa) / topic 5



A comparison of the words with greatest differences between **Ramaphosa** and **Zuma** reveal differences in their presidential styles.

Terms like `friend` and `compatriot` by Zuma's indicate his reliance on left-wing populist rhetoric in his speeches, whilst the use of terms like `SOEs` and `capture` by Ramaphosa suggest that Zuma did not to any same degree confront the issues of corruption and inefficiencies in state-owned-enterprises (including Eskom) that beset the legacy of his presidency (and which formed the basis of Ramaphosa's subsequent challenges).

Figure 5: Log2 ratio of beta in 2019 post−election speech



The section of topic modelling is ended with a look at the greatest word differences between Ramaphosa's **2019 post-election speech**, and the one he delivered in **2023**.

Terms in 2019 included `dream`, `comfort`, `bold` and `NDP` (the *National Development Plan*, initiated in 2013, which aims to eliminate poverty and reduce inequality by 2030).

Terms in the 2023 included `pandemic`, `disaster`, `overcome`, `transition`, and tellingly, not a single mention of the *National Development Plan*.

This suggests the use of optimistic and idealistic terms during an election year that support a political agenda, that is not evident in non-election years.

3.4 Review: Using ChatGPT as a Large Language Model

Large Language Models (LLMs) are sophisticated artificial intelligence tools that assist with numerous tasks, such as compiling, processing, and writing reports, such as this one. ChatGPT, in particular, has shown remarkable capabilities in assisting with such tasks, and so it was used to assist with completing this task and the following is a detailed review of the experience.

3.4.0.1 Prompting for Writing:

When it came to the writing process, ChatGPT excelled when it was provided with detailed descriptions of what was required, often in concise bullet points. This structured approach allowed ChatGPT to address each point systematically. While it usually produced well-structured and coherent content, minor human adjustments were occasionally necessary for optimal results.

Its performance deteriorated when clear and detailed instructions were not provided. In such cases, the results were less satisfactory, very generalized and often necessitated additional prompting or fine-tuning.

3.4.0.2 Prompting for Coding:

In the realm of coding, ChatGPT proved to be a valuable resource for understanding and debugging errors. When presented with coding errors, it demonstrated an impressive accuracy rate of identifying the root cause of the issue.

Furthermore, for simpler or common coding exercises, ChatGPT typically provided effective solutions on the first attempt. However, it's with more complex coding problems, even when prompts were tuned, ChatGPT struggled to produce correct answers. This is perhaps indicative of the limitations of current AI models in handling intricate coding challenges.

3.4.0.3 Prompting for Facts and Definitions:

ChatGPT delivered accurate facts and definitions when prompted. However, due diligence remained essential, as the information provided was cross-checking for precision.

One remarkable aspect of ChatGPT was its ability to explain topics comprehensively, providing definitions that allowed for a solid understanding of the subject matter. It excelled in generating examples and rephrasing information in simpler terms, which greatly aided in achieving a clear and thorough comprehension of the topics it addressed.

In summary, ChatGPT has proven to be a valuable tool for a variety of tasks, particularly in the realm of report compilation, coding assistance, and knowledge provision. While it excels with well-defined and detailed prompts, its limitations become apparent when instructions are less clear or when confronted with intricate coding challenges. Nevertheless, ChatGPT's capabilities in simplifying complex concepts and generating good quality written content offer significant value in a variety of applications.

4 Conclusions

References

Blei, David M. 2012. "Probabilistic Topic Models." *Communications of the ACM* 44 (4): 77–84. https://dl.acm.org/doi/10.1145/2133806.2133826.

Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. 2014. "Sentiment Analysis Algorithms and Applications: A Survey." *Ain Shams Engineering Journal* 5 (4): 1093–113.

Miranda, John Paul P, and Rex P Bringula. 2021. "Exploring Philippine Presidents' Speeches: A Sentiment Analysis and Topic Modeling Approach." *Cogent Social Sciences* 7 (1): 1932030.

Palmer, David D. 2000. "Tokenisation and Sentence Segmentation." *Handbook of Natural Language Processing*, 11–35.

Pascual, Federico. 2019. "Topic Modeling: An Introduction." https://monkeylearn.com/blog/introduction-to-topic-modeling/.

Wankhade, Mayur, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. "A Survey on Sentiment Analysis Methods, Applications, and Challenges." *Artificial Intelligence Review* 55 (7): 5731–80.

Yan-Yan, Zhao, Qin Bing, and Liu Ting. 2010. "Integrating Intra-and Inter-Document Evidences for Improving Sentence Sentiment Classification." *Acta Automatica Sinica* 36 (10): 1417–25.