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

Ropafadzo Chimuti Tanweer Nujjoo Steven Ellis

Summary of motivation and outcome. Start with context, task and object, finish with findings and conclusion. This is written last.

Table of contents

1 Introduction

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 either via the `ggplot()` function from the `ggplot2` library.

2.1 Brief Overview of SONA dataset

Ropa's 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 were first 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 were 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, hence were filtered out.
2. bigrams, each word was separated into 2 different columns and then step 1 was repeated for each column. Then the cleaned separated words were united back in one column.
3. trigrams, each word was separated into 3 different columns and then step 1 was repeated for each column. Then the cleaned separated words were 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 words and trigrams used per presidents were analysed. The bigrams ones were omitted as the idea was, if we were to use more than a word 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 bigram part is 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 were first 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 were 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 were filtered out. The relevant dictionary (further description about dictionaries is detailed in Section 2.6) was 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 were present in the dictionaries. Therefore, words that did not have a label were mutated to "neutral".

2. trigrams, each word was separated into 3 different columns and then step 1 was repeated for each column. Then the cleaned separated words were united back in one column. Because we were dealing with trigrams, the sentiments had 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 was 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 words and trigrams used by all presidents. The steps to achieve these, followed the main procedures highlighted in Section 2.5. Then for each case, the words and 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 both words and 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.

Ropa: nrc dictionary descriptions

2.7 Topic Modelling

Steven's section Summary of topic modelling, LDA. Tools used for LDA (R packages), how data was pre-processed for LDA, and techniques used.

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. Nelson 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**

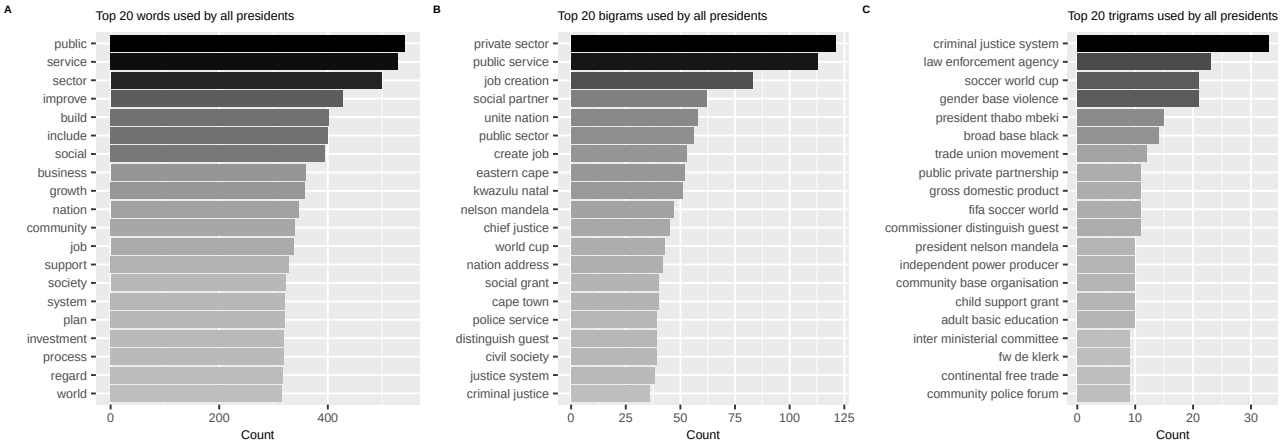

Figure 1: Top 20 words, pair of words (bigrams), and trio of words (trigrams) used by all presidents.

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.

The top 20 words used by each president are represented Figure 2. Based on the words of Mandela, he really demonstrated a presidency that was committed to public service, inclusivity, community building, and addressing a range of social, economic, and security-related challenges. Mandela's leadership was marked by a dedication to nation-building, reconciliation, and addressing the needs of the people. In general, deKlerk's top 20 words were very political, which completely makes sense because he had to promote a peaceful post-apartheid transition. His presidency was characterised by significant political, constitutional, and social changes, aimed at ending apartheid and establishing a more inclusive and democratic South Africa. The top 20 words reflected the complex and multifaceted nature of the challenges he faced and the efforts he made to transform the country's political landscape. Based on the top 20 words from Mbeki, it seemed like his presidency was marked by a commitment to improving public services, addressing social issues, promoting economic growth and development, and addressing significant challenges. His top 20 words were not that meaningful because that is basically what all presidents would want to achieve in a society. Among the top 20 words of Zuma, the ones that really stood out were energy and water as during his presidency, he was dealing with some energy and water crisis. The top 20 words for Motlanthe became quite repetitive at this point. We saw that he still dealt with some crisis, poverty and other challenges. Interestingly, he was the first president from which the word poverty was first appeared in the top tiers. Ramaphosa's top 20 words revealed a lot about his personality. He is a businessman, so his approach as president was very business-oriented. He seemed to have priorited economic development, job creation, and social welfare during his term as president.

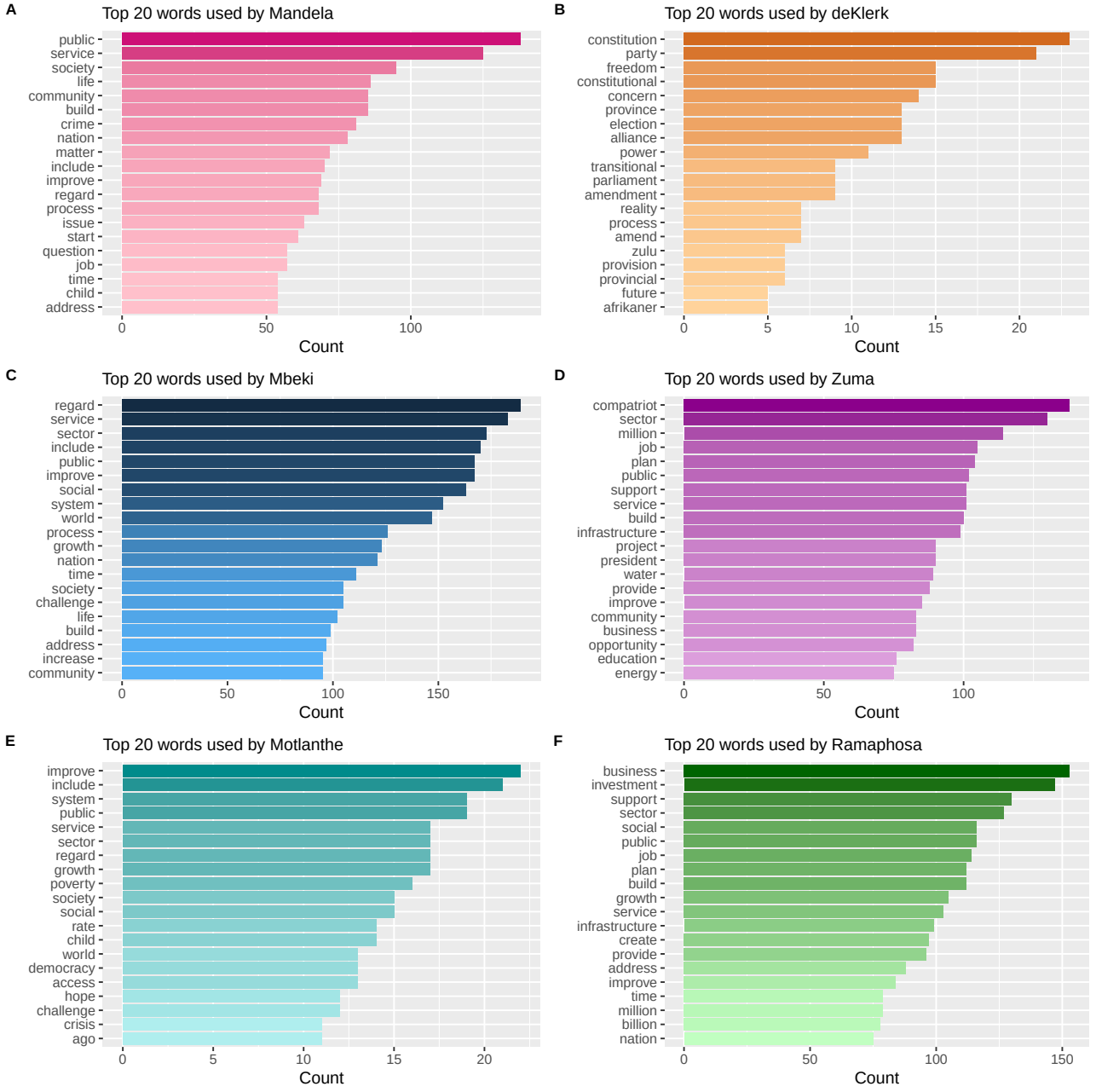

Figure 2: Top 20 words used by each president. **A** relates to Mandela's words, **B** relates to deKlerk's words, **C** relates to Mbeki's words, **D** relates to Zuma's words, **E** relates to Motlanthe's words, and **F** relates to Ramaphosa's words.

Figure 3 conveys more meaningful information as we are now dealing with 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.

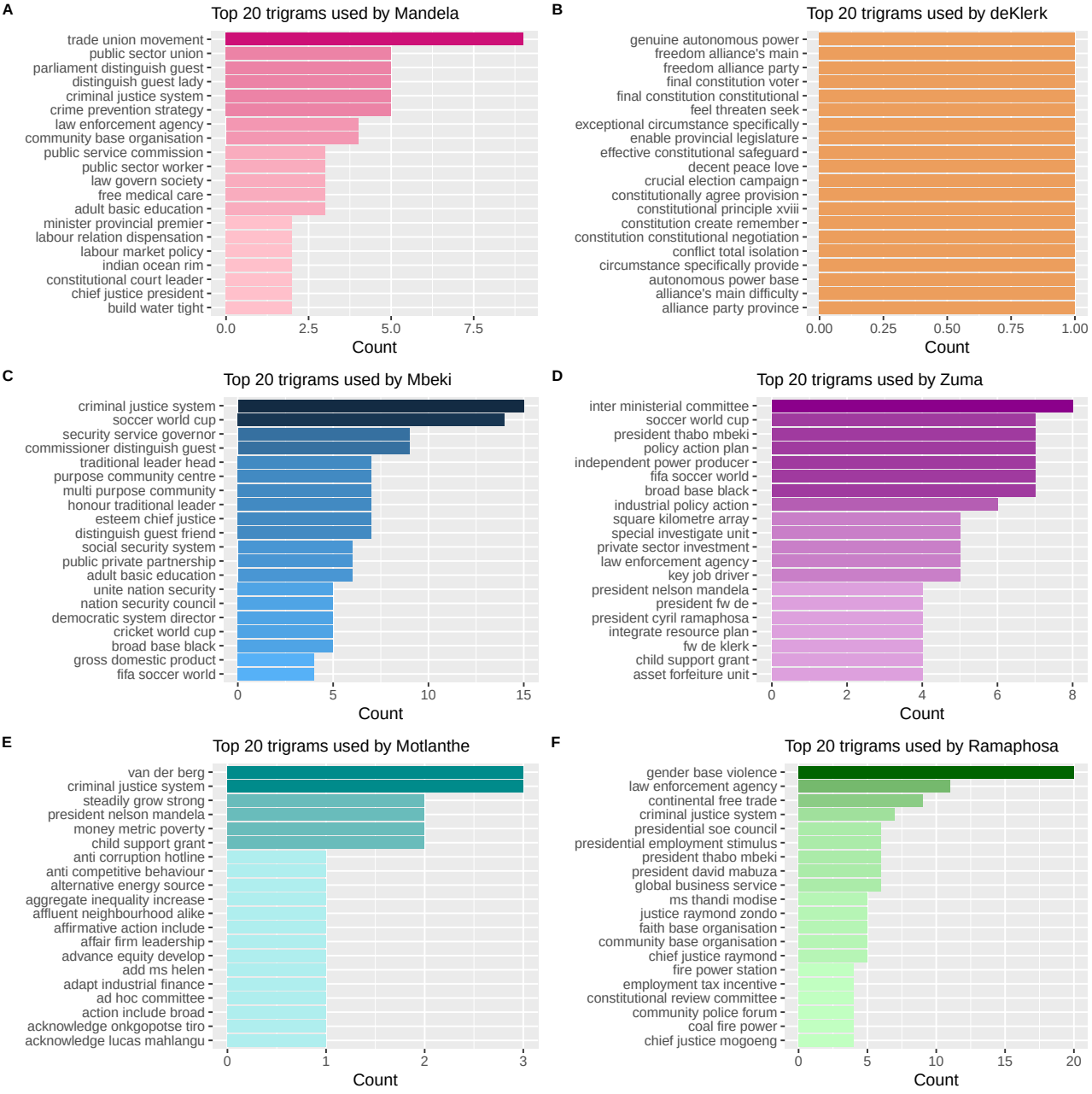

Figure 3: Top 20 trio of words (trigrams) used by each president. **A** relates to Mandela's trigrams, **B** relates to deKlerk's trigrams, **C** relates to Mbeki's trigrams, **D** relates to Zuma's trigrams, **E** relates to Motlanthe's trigrams, and **F** relates to Ramaphosa's trigrams.

3.2 Sentiment Analysis

Figure 4...

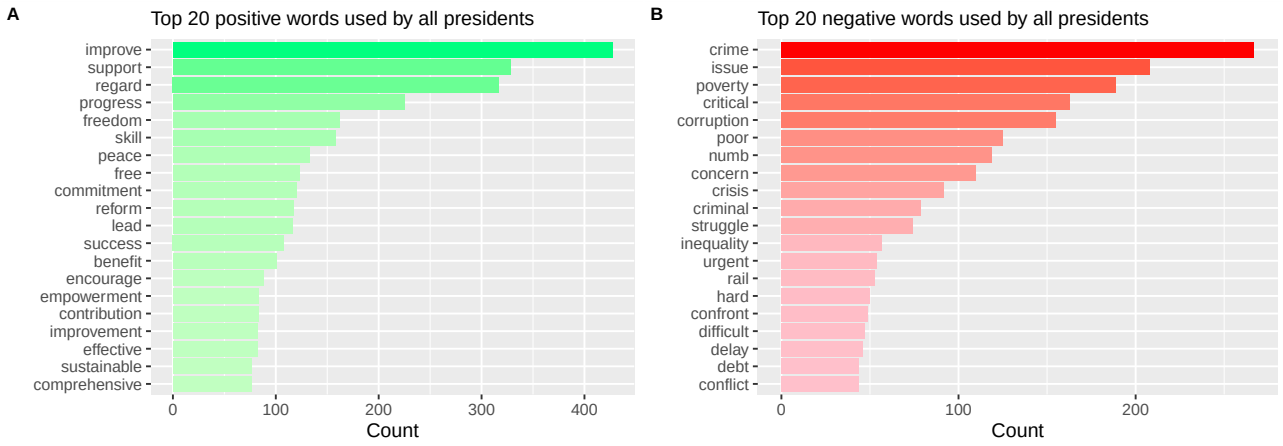

Figure 4: Top 20 positive (**A**) and negative (**B**) words used by all presidents.

Figure 5...

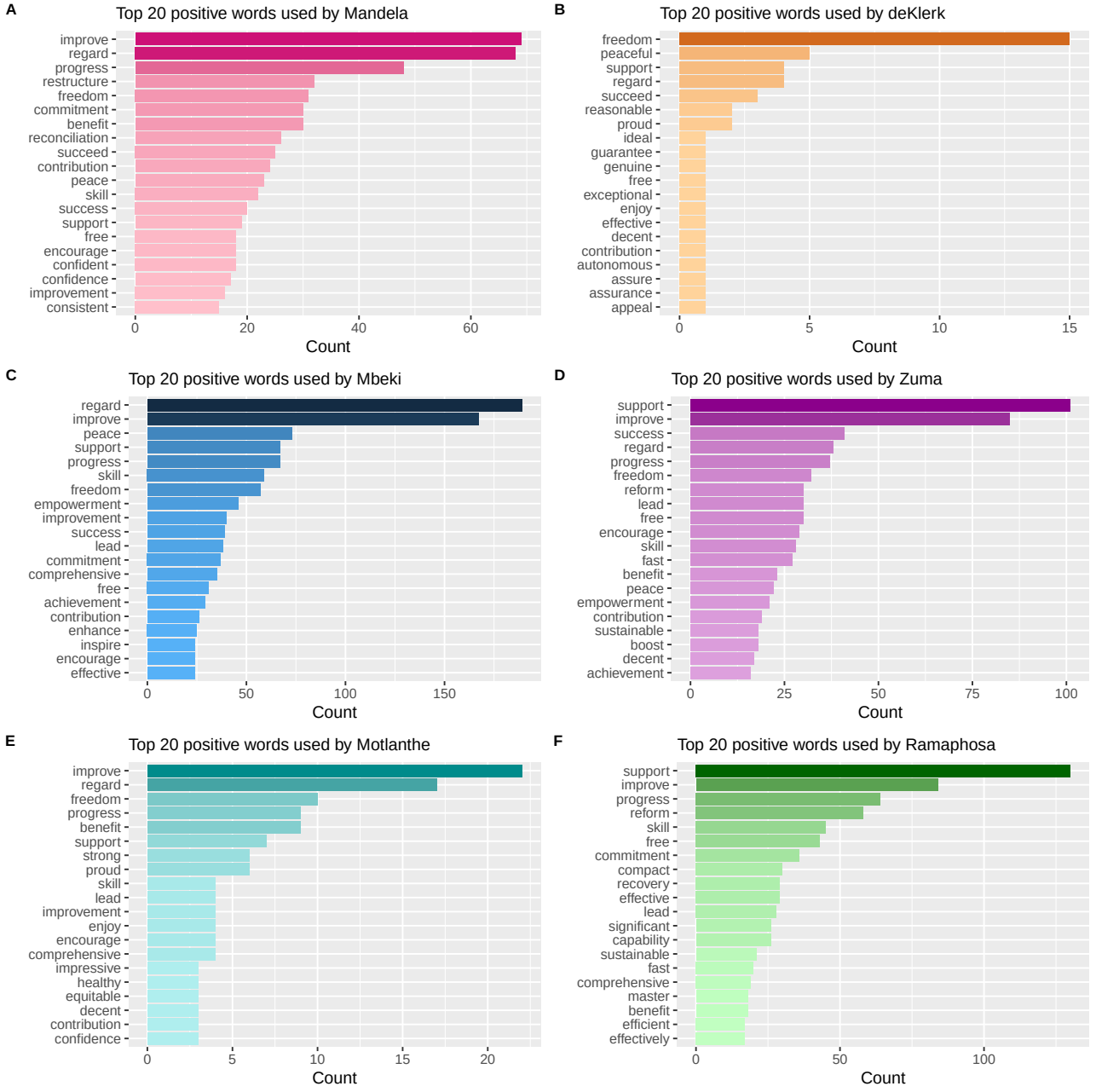

Figure 5: Top 20 positive words used by each president. **A** relates to Mandela's positive words, **B** relates to deKlerk's positive words, **C** relates to Mbeki's positive words, **D** relates to Zuma's positive words, **E** relates to Motlanthe's positive words, and **F** relates to Ramaphosa's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

Figure 6...

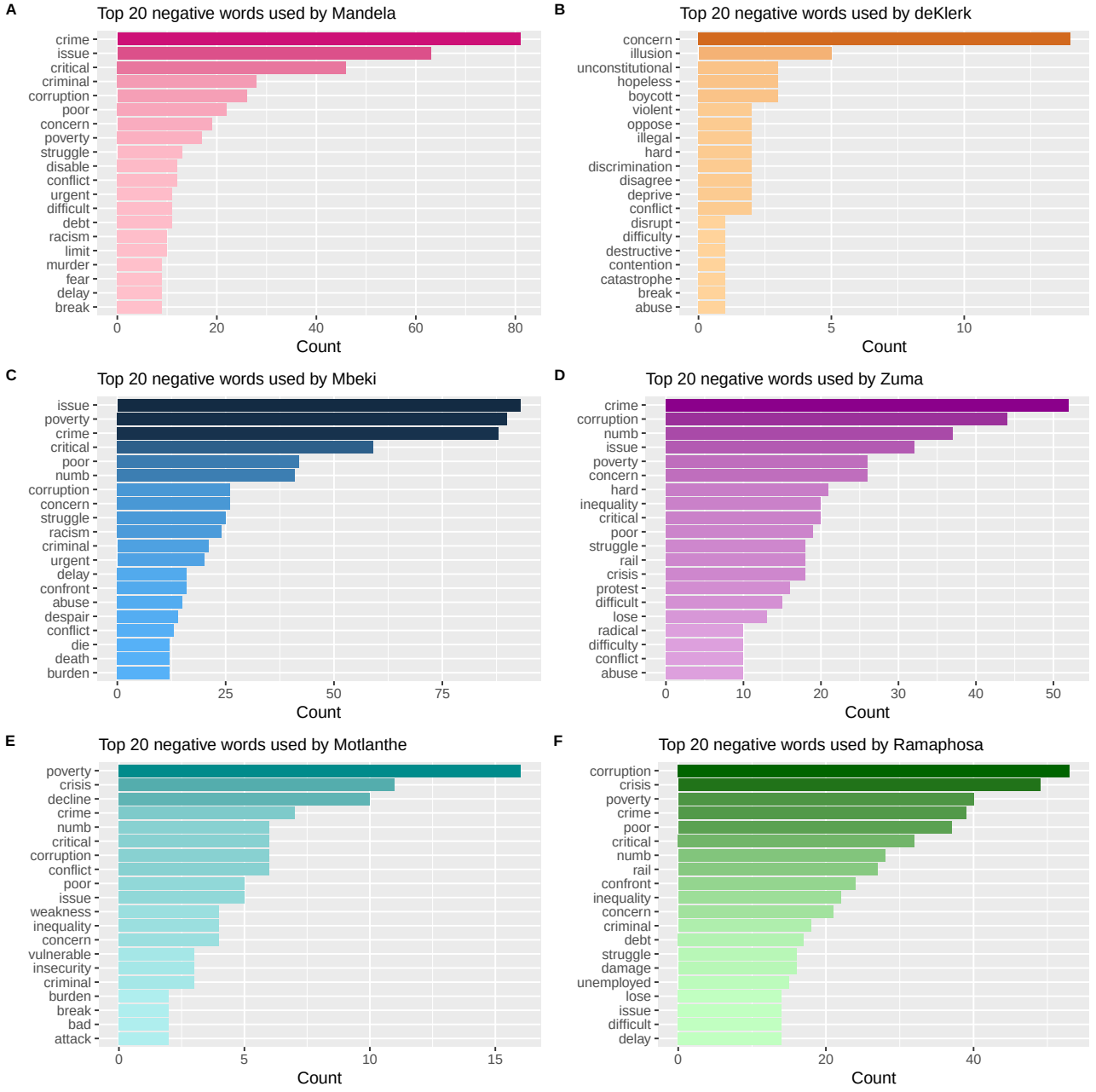

Figure 6: Top 20 negative words used by each president. **A** relates to Mandela's negative words, **B** relates to deKlerk's negative words, **C** relates to Mbeki's negative words, **D** relates to Zuma's negative words, **E** relates to Motlanthe's negative words, and **F** relates to Ramaphosa's negative words.

Figure 7...

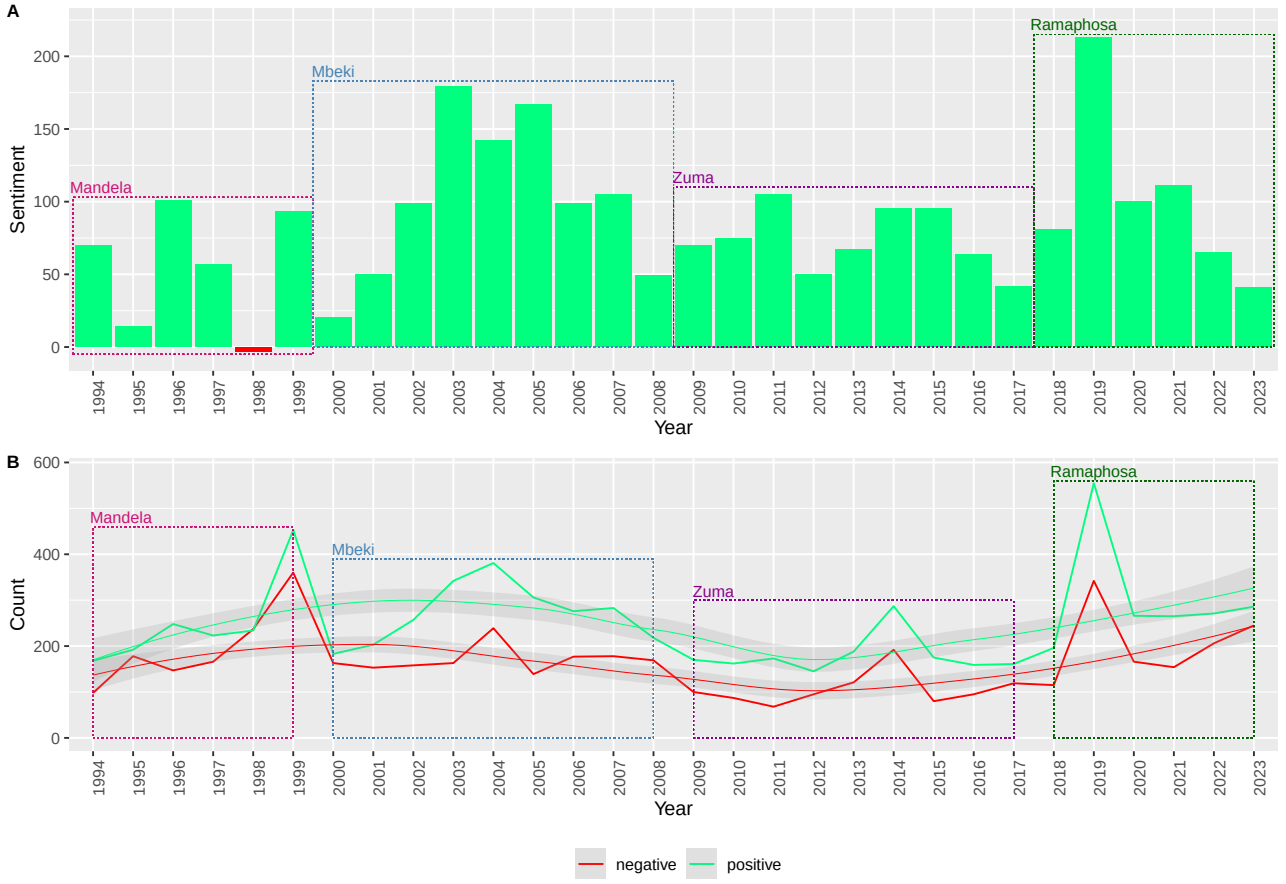

Figure 7: Positive and negative words sentiment over time (excluding 1-term presidents such as deKlerk and Motlanthe). **A** corresponds to how many more positive than negative words (or vice versa) computed over the years [1994-2023]. So, the y-axis for **A** was calculated as *Sentiment = positive words - negative words*. **B** refers to the actual count of positive and negative words 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.

Plot **A** of Figure 8 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.

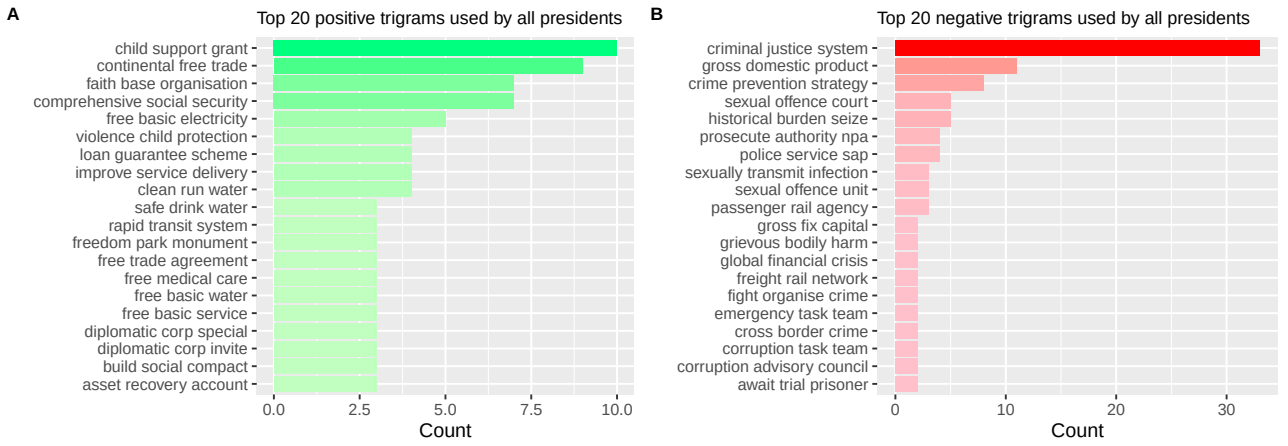

Figure 8: Top 20 positive (**A**) and negative (**B**) trio of words (trigrams) used by all presidents.

Figure 9 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.

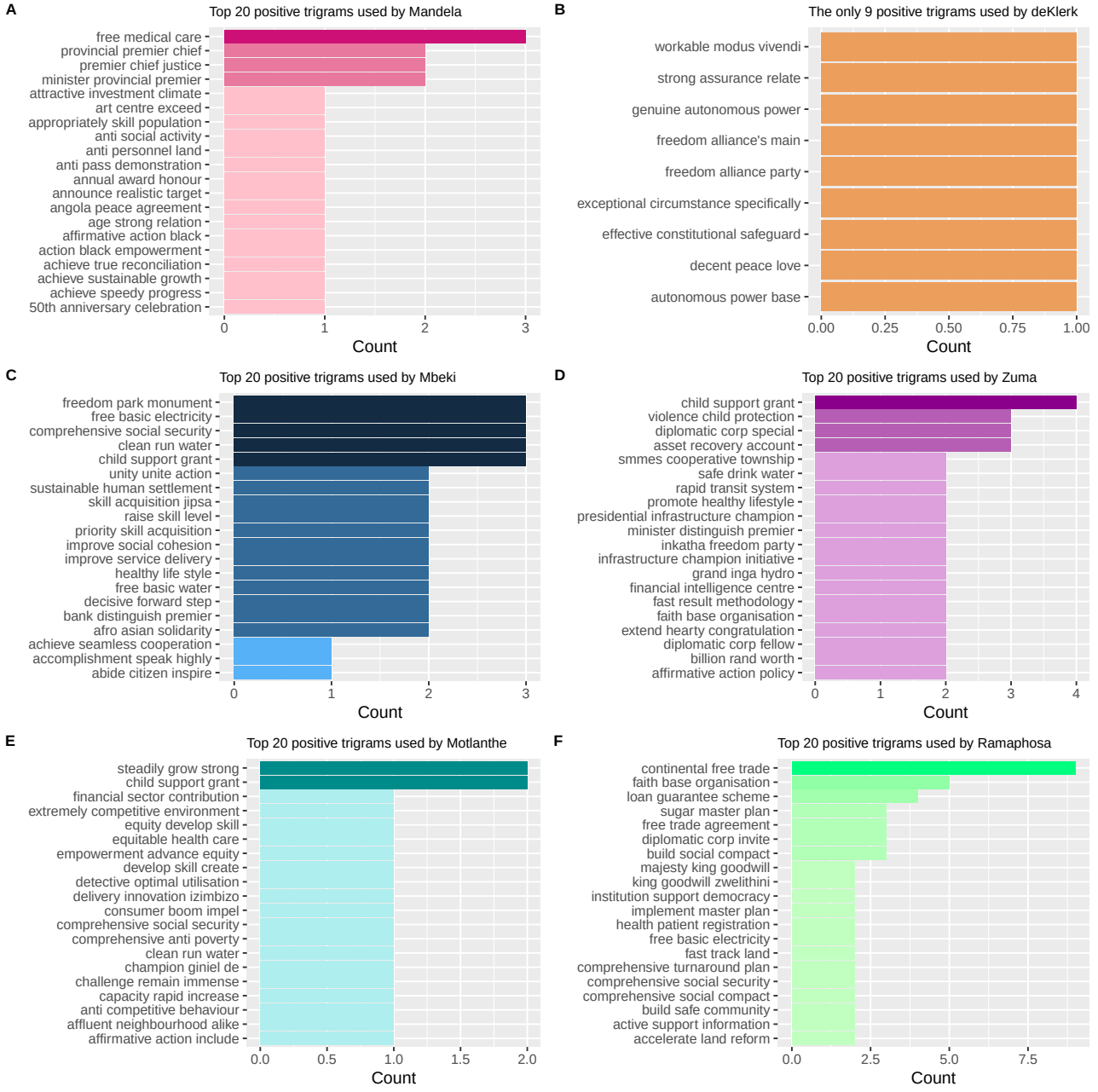

Figure 9: Top 20 positive trio of words (trigrams) used by each president. **A** relates to Mandela's positive trigrams, **B** relates to deKlerk's positive trigrams, **C** relates to Mbeki's positive trigrams, **D** relates to Zuma's positive trigrams, **E** relates to Motlanthe's positive trigrams, and **F** relates to Ramaphosa's positive trigrams.

Figure 10 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. Mot-

lanthe'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.

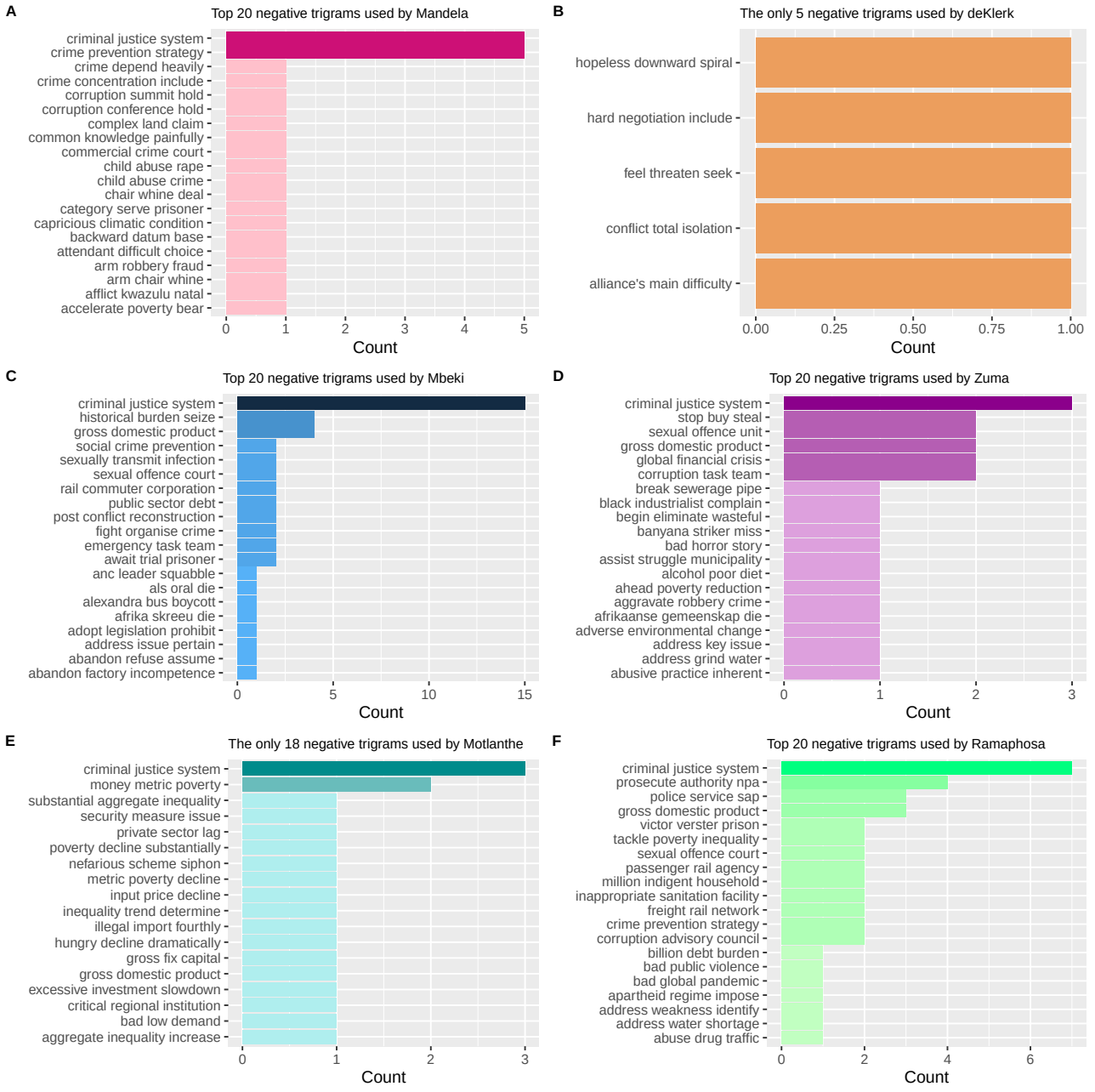

Figure 10: Top 20 negative trio of words (trigrams) used by each president. **A** relates to Mandela's negative trigrams, **B** relates to deKlerk's negative trigrams, **C** relates to Mbeki's negative trigrams, **D** relates to Zuma's negative trigrams, **E** relates to Motlanthe's negative trigrams, and **F** relates to Ramaphosa's negative trigrams.

So, when we investigated the sentiment over the years from Figure 11, 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.

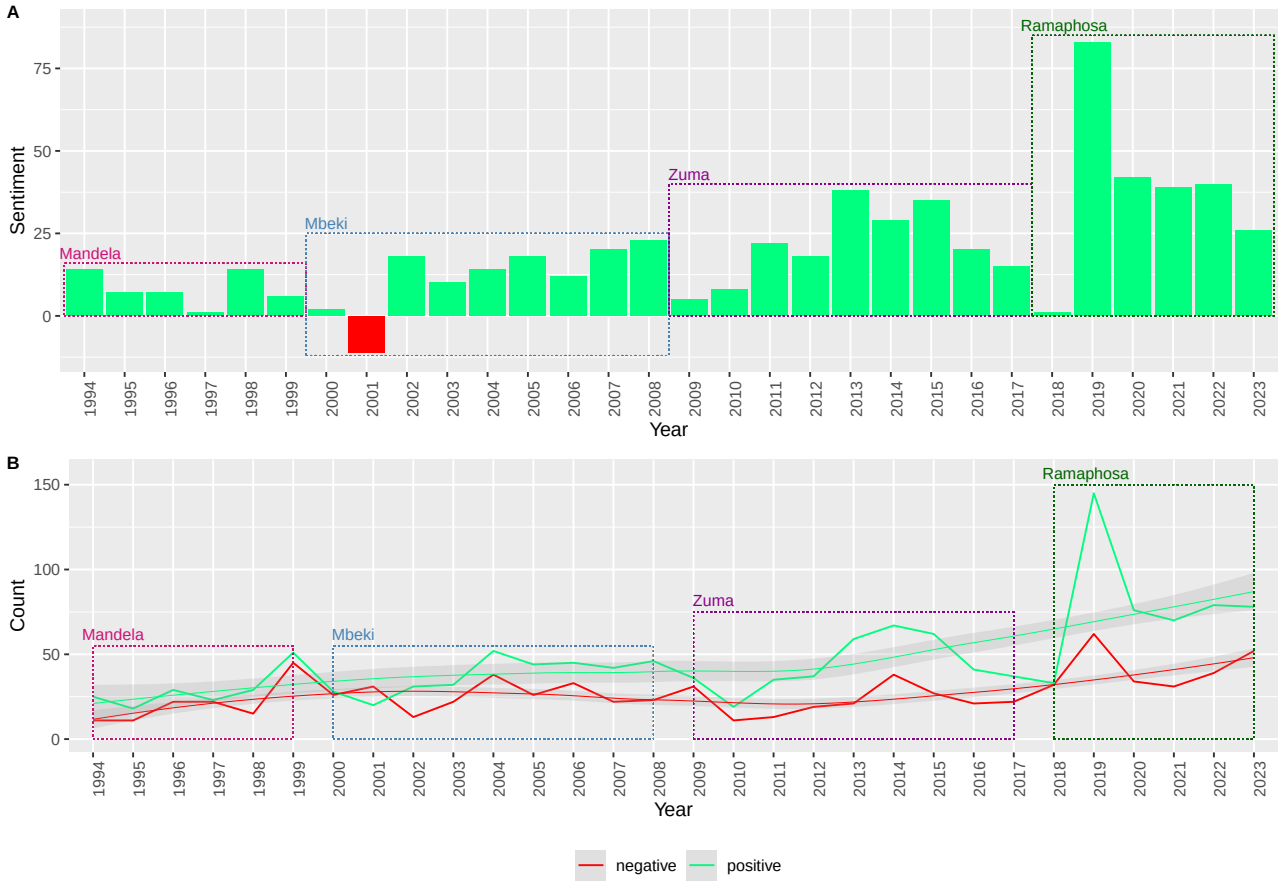

Figure 11: Positive and negative trio of words (trigrams) sentiment over time (excluding 1-term presidents such as deKlerk and Motlanthe). **A** corresponds to how many more positive than negative trigrams (or vice versa) computed over the years [1994-2023]. So, the y-axis for **A** was calculated as *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.

3.3 Topic Modelling

Steve's section

4 Conclusion

.....

References

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.

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.