NLP of UNGA 78th Session Statements

Analysing statements from African countries during the 78th United Nations General Assembly Debate through text mining, employing Natural Language Processing techniques and predefined algorithms.

Table of Contents

[List of Tables 4](#_Toc151843399)

[List of Figures 4](#_Toc151843400)

[1 Manipulating Text 5](#_Toc151843401)

[1.1 Frequency Tokens 5](#_Toc151843402)

[1.2 Universal part of speech (POS) tag 8](#_Toc151843403)

[1.2.1 Nouns 9](#_Toc151843404)

[1.2.2 Verbs 10](#_Toc151843405)

[1.2.3 Adjectives 11](#_Toc151843406)

[1.2.4 Adverbs 12](#_Toc151843407)

[1.3 Summary 13](#_Toc151843408)

[2 Visualizing Words 14](#_Toc151843409)

[2.1 Frequent words 14](#_Toc151843410)

[2.1.1 Top most Frequent words per country 14](#_Toc151843411)

[2.1.2 Top most Frequent words as a whole 15](#_Toc151843412)

[2.2 word could 16](#_Toc151843413)

[2.3 Summary 17](#_Toc151843414)

[3 Key word Identification 18](#_Toc151843415)

[3.1 Using Rake method 19](#_Toc151843416)

[3.2 Pointwise mutual information 20](#_Toc151843417)

[3.3 Using POS method 21](#_Toc151843418)

[4 Analysing n-gram 22](#_Toc151843419)

[4.1 Nouns / adjectives used in same sentence 22](#_Toc151843420)

[4.2 Nouns / adjectives which follow one another 23](#_Toc151843421)

[4.3 Summary 24](#_Toc151843422)

[5 Sentiment Analysis 25](#_Toc151843423)

# List of Tables

[Table 1.1: descriptive statistics 6](#_Toc151824059)

# List of Figures

[Figure 1.1: Part of Speech distribution by country 8](#_Toc151824070)

[Figure 1.2: Top nouns by country 9](#_Toc151824071)

[Figure 1.3: Top Verbs by Country 10](#_Toc151824072)

[Figure 1.4: Top Adjectives by country 11](#_Toc151824073)

[Figure 1.5: Top Adverbs by Country 12](#_Toc151824074)

[Figure 2.1: Ten Most Frequent Words in Country Statement 15](#_Toc151824075)

[Figure 2.2:Top 20 Most Frequent words as a whole 15](#_Toc151824076)

[Figure 2.3: word could 16](#_Toc151824077)

[Figure 3.1:Key word identified by RAKE 19](#_Toc151824078)

[Figure 3.2: PMI (Pointwise Mutual Information) 20](#_Toc151824079)

[Figure 3.3: Part-of-Speech (PoS) 21](#_Toc151824080)

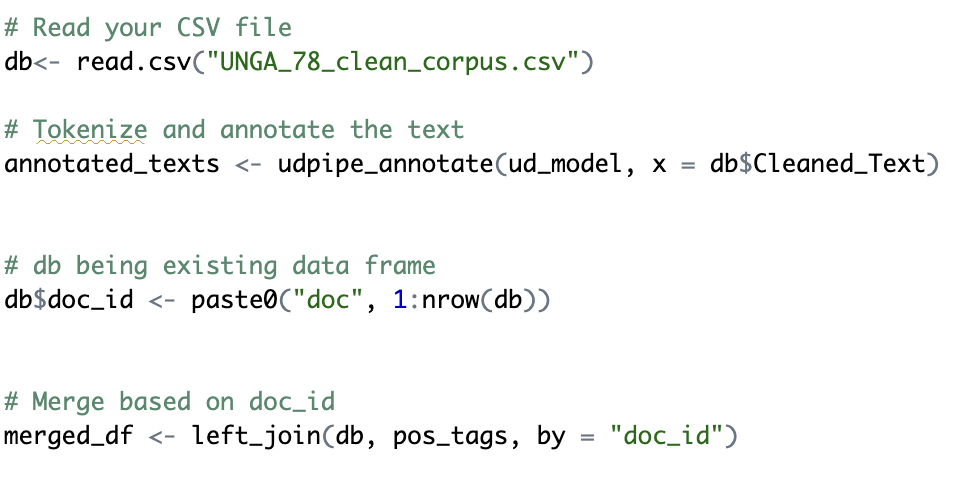
[Figure 4.1: Nouns / adjectives used in same sentence 22](#_Toc151824081)

[Figure 4.2: Nouns / adjectives which follow one another 23](#_Toc151824082)

[Figure 5.1: Speech Sentiment 25](#_Toc151824083)

# Manipulating Text

The process begins by loading the speeches' clean data. The data underwent cleaning, stemming, lemmatization, and categorization using the Universal Part-of-Speech (UPOS) system with the assistance of the udpipe R package.



## Frequency Tokens

To provide a concise and expeditious analysis of the speech, a tabular representation of word and sentence frequencies and the syntax (code) is presented herein.

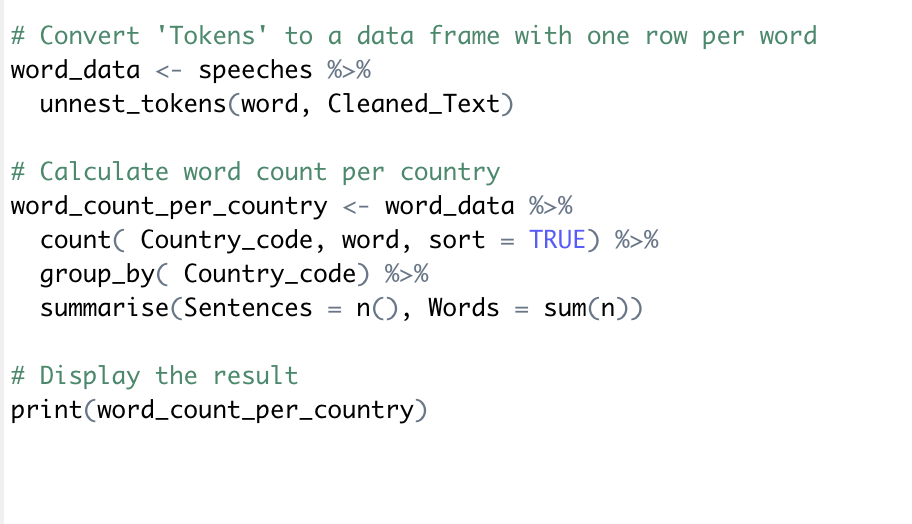


Table 1.1: descriptive statistics

|  |  |  |
| --- | --- | --- |
| Country | Sentences | Words |
| SIERRA LEONE | 971 | 1651 |
| ANGOLA | 855 | 1465 |
| KENYA | 902 | 1444 |
| EGYPT | 844 | 1419 |
| GAMBIA | 760 | 1336 |
| MAURITIUS | 784 | 1225 |
| MOZAMBIQUE | 738 | 1178 |
| ESWATINI' | 695 | 1150 |
| UGANDA | 607 | 1086 |
| LESOTHO | 674 | 1050 |
| GHANA | 642 | 1033 |
| SOUTH AFRICA | 545 | 1029 |
| NAMIBIA | 582 | 939 |
| SEYCHELLES' | 565 | 936 |
| TANZANIA | 608 | 931 |
| ETHIOPIA | 614 | 912 |
| NIGERIA | 624 | 902 |
| BOTSWANA | 593 | 888 |
| LIBERIA | 514 | 869 |
| ZIMBABWE' | 520 | 794 |
| CAPE VERDE | 463 | 777 |
| MALAWI | 458 | 775 |
| GUINEA-BISSAU | 336 | 512 |
| SOUTH SUDAN | 366 | 510 |
| ERITEA | 368 | 459 |
| RWANDA | 352 | 457 |

Based on the statistics shown in Table 1, it is apparent that Sierra Leone exhibited the greatest frequency of sentences and words in their speech, followed by Angola, Kenya, Egypt, Gambia, Mauritius, and Mozambique. The countries of Rwanda, Eritrea, South Sudan, and Guinea-Bissau presented the shortest sentences and words in their speech, as stated in given table.

## Universal part of speech (POS) tag

For a comprehensive list of the parts of speech (POS) tags and their corresponding definitions, please refer to this resource: <https://universaldependencies.org/u/pos/index.html>. The provided code examines the distribution of each category independently.

A screenshot of a computer program

Description automatically generated

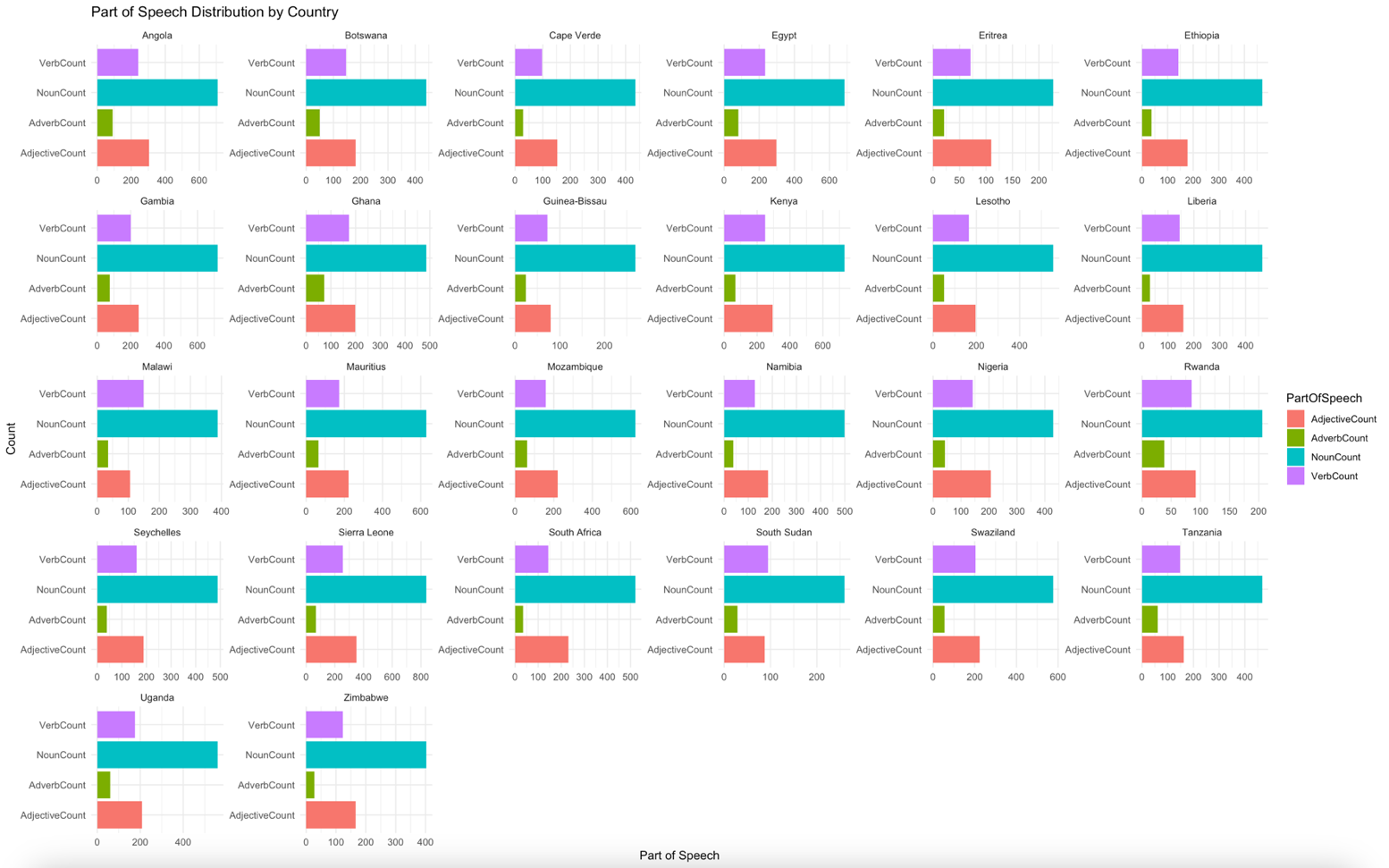


Figure 1.1: Part of Speech distribution by country

From figure 1.1, it is evident that Sierra Leone possesses the longest speech, surpassing all other countries in this regard. Subsequently, Figure 1.1 also illustrates the frequency distribution of each UPOS type. Most of the speeches primarily comprise of nouns, adverbs, verbs, and adjectives.

### Nouns

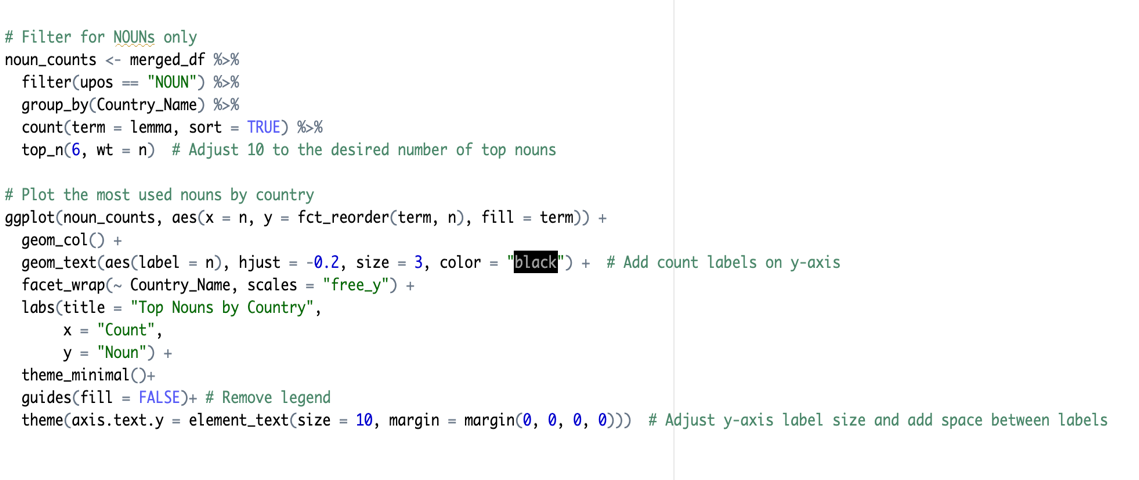
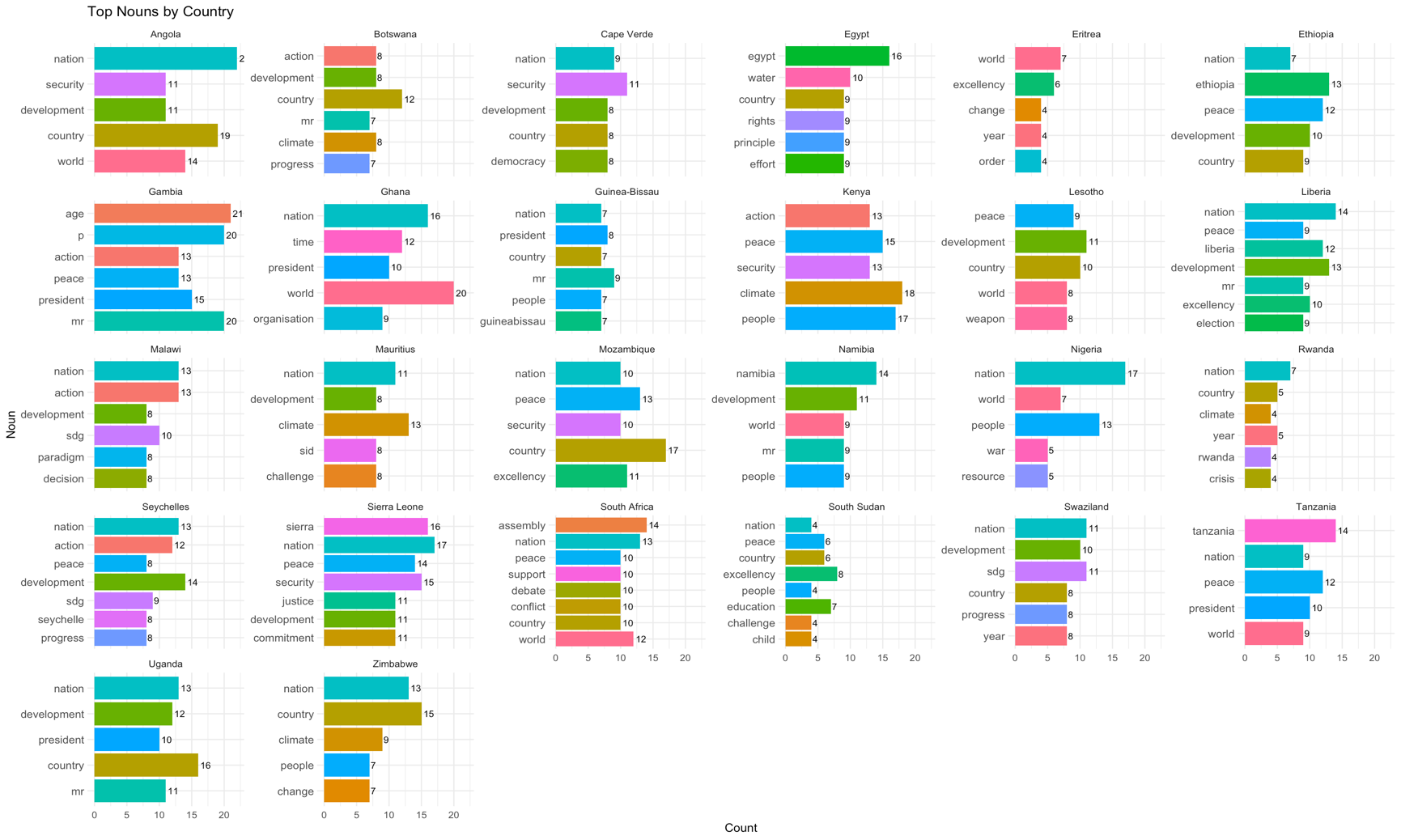


Figure 1.2: Top nouns by country

### Verbs

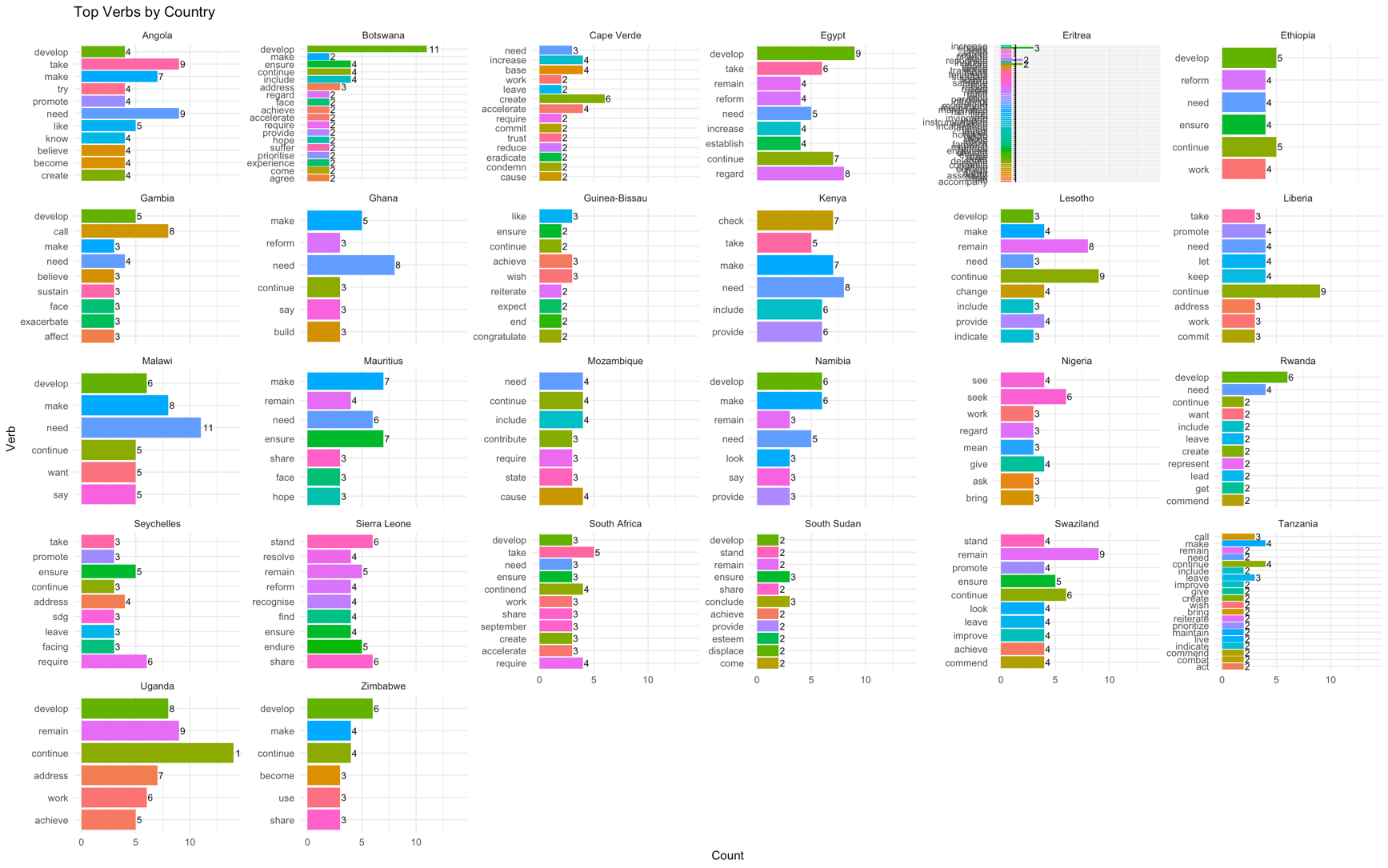


Figure 1.3: Top Verbs by Country

### Adjectives

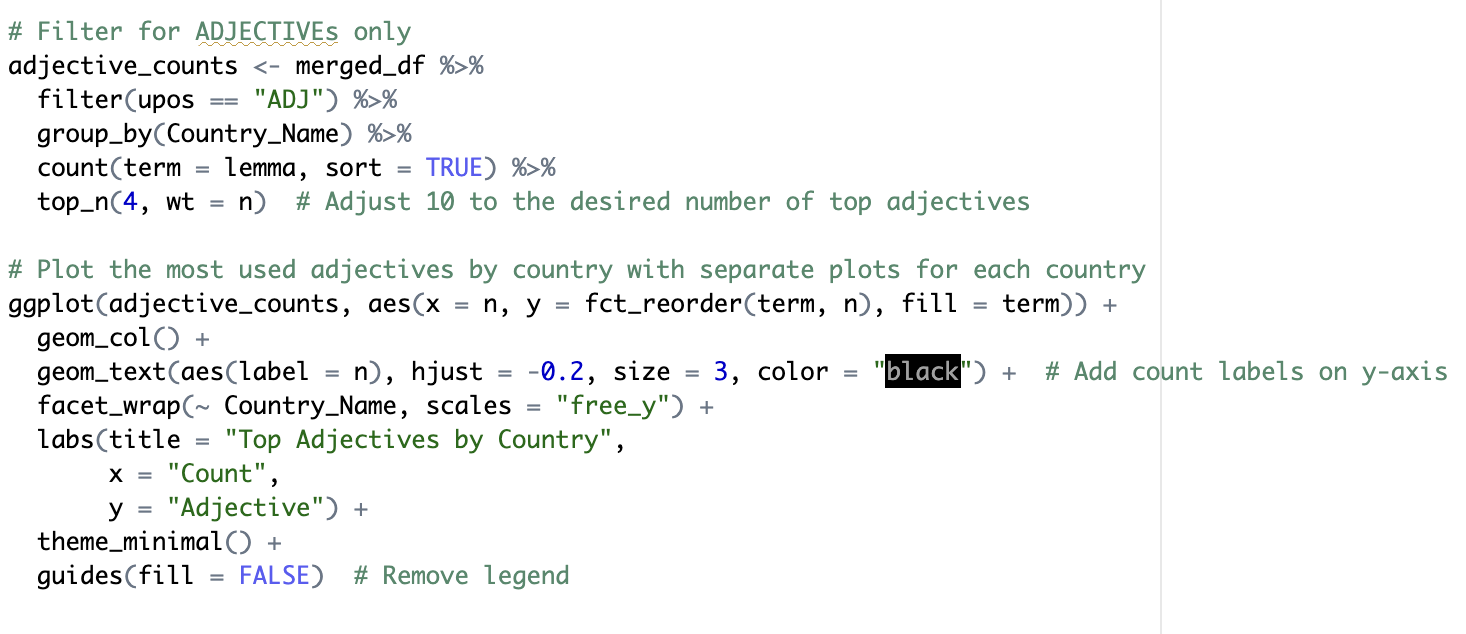


Figure 1.4: Top Adjectives by country

### Adverbs

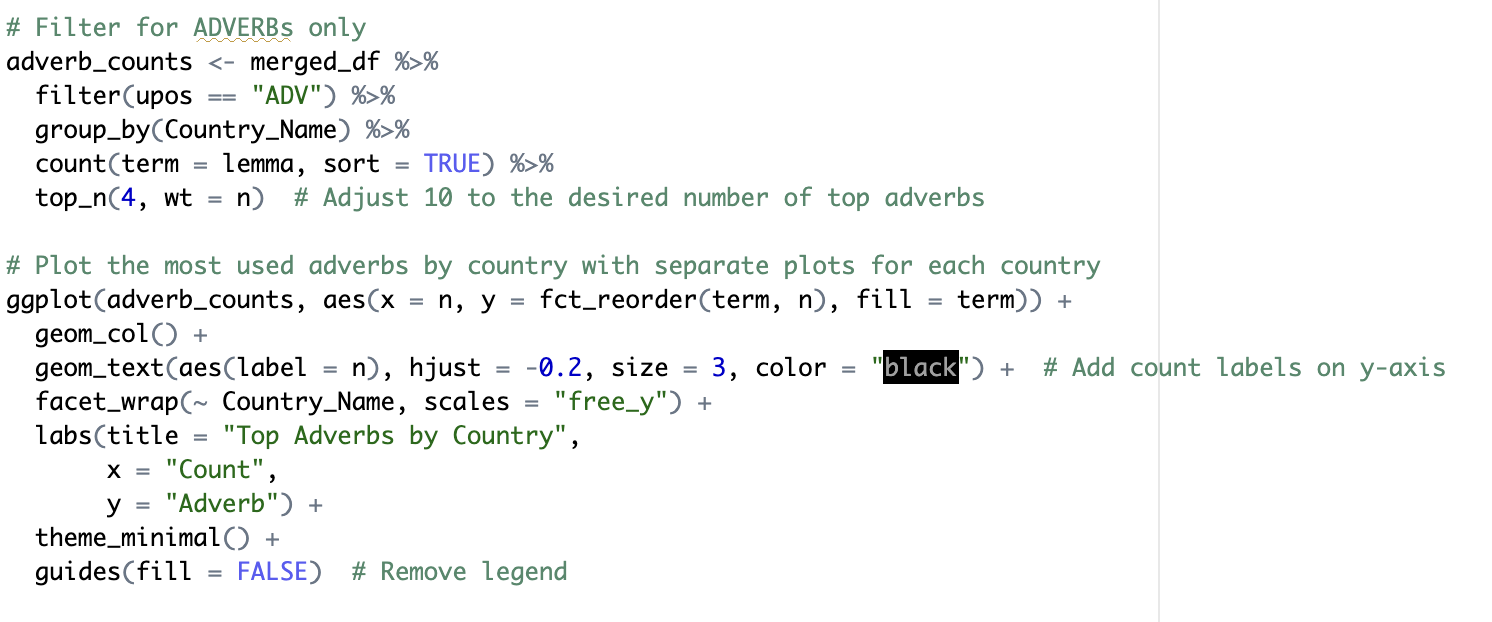
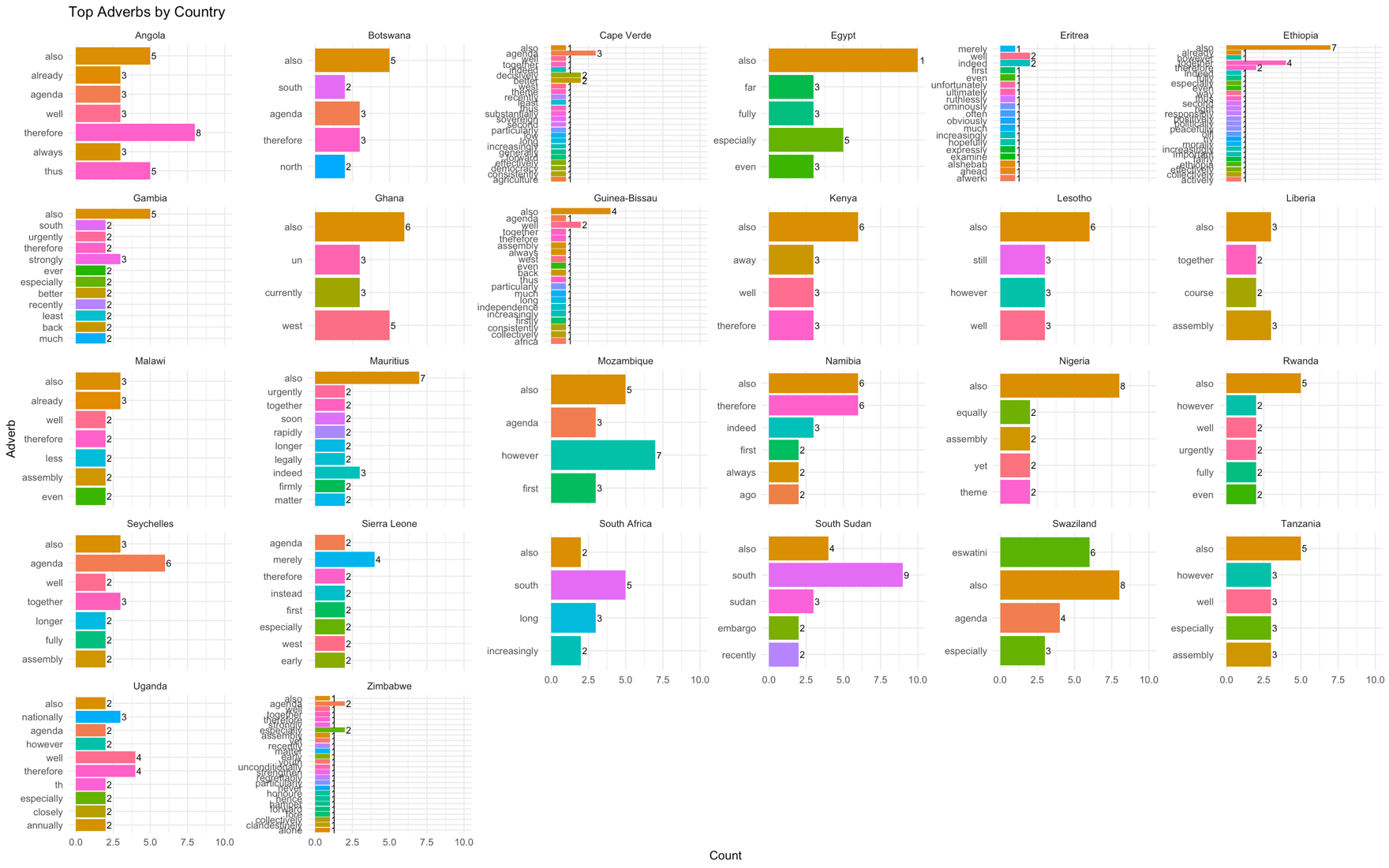


Figure 1.5: Top Adverbs by Country

## Summary

It is evident that Sierra Leone possesses the longest speech, surpassing all other countries in this regard. Subsequently, this sections also illustrates the frequency distribution of each UPOS type. Most of the speeches primarily comprise of nouns, adjectives, verbs, and adverbs. Based on the depicted picture, it can be observed that most national speeches exhibit a higher frequency of nouns and adjectives, followed by verbs, and finally adverbs. For a comprehensive list of the parts of speech (POS) tags and their corresponding definitions, please refer to this resource: <https://universaldependencies.org/u/pos/index.html> .

# Visualizing Words

The initial segment pertains to the frequency of words within each statement. This procedure calculates the words that exhibit the highest degree of exclusivity in a specific speech. This metric quantifies the degree of specificity exhibited by individual speeches in terms of their respective vocabularies. The second portion will incorporate a word cloud that encompasses all the aggregated statements, serving as an additional visual representation of the frequency of words used. The frequency of a word in the text is shown by its size.

## Frequent words

### Top most Frequent words per country

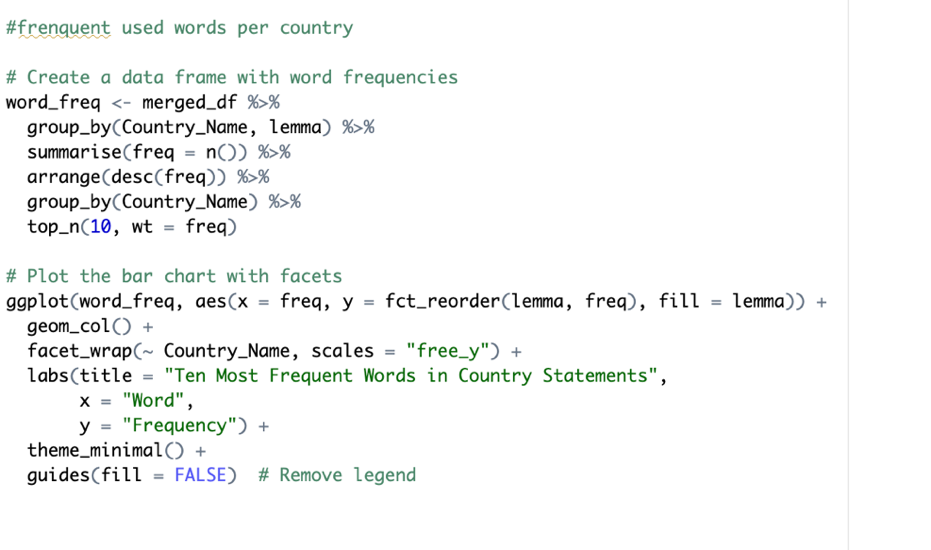
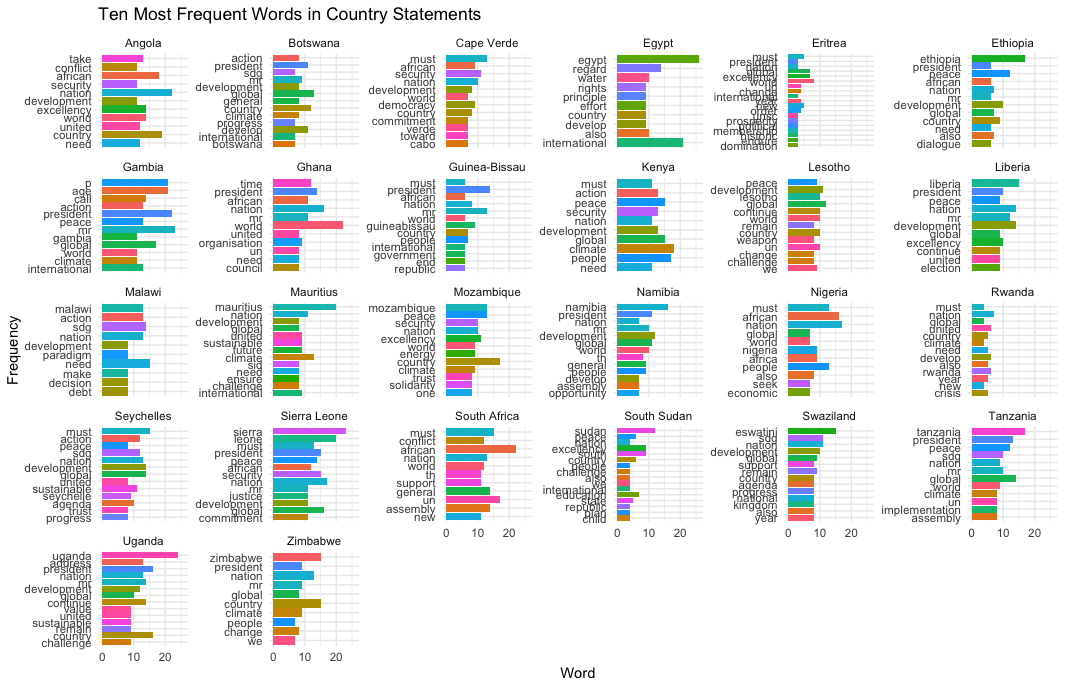
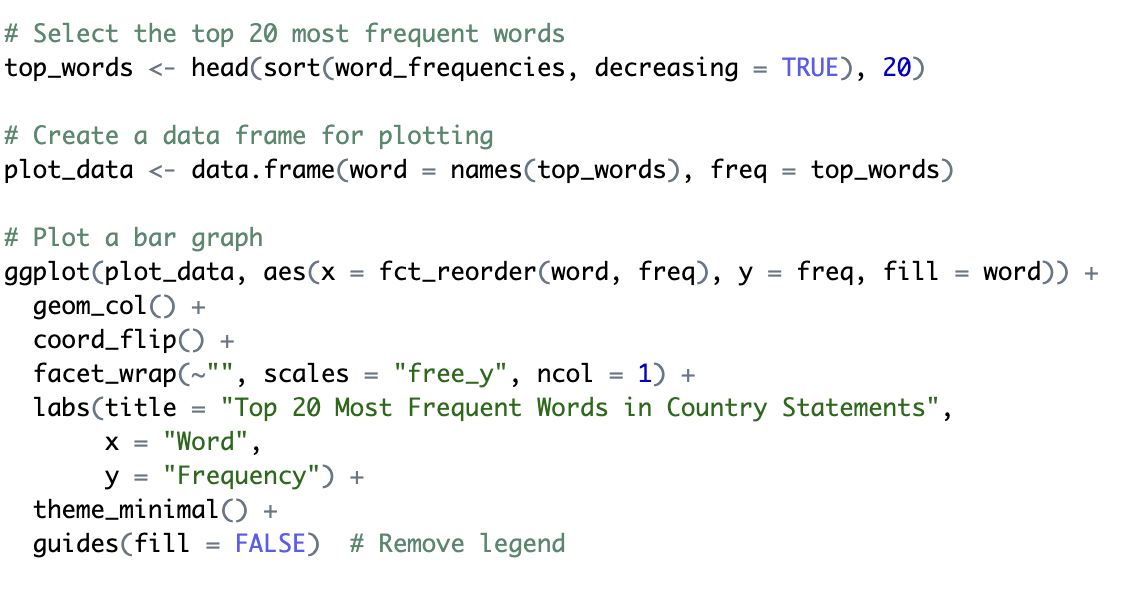


Figure 2.1: Ten Most Frequent Words in Country Statement

### Top most Frequent words as a whole



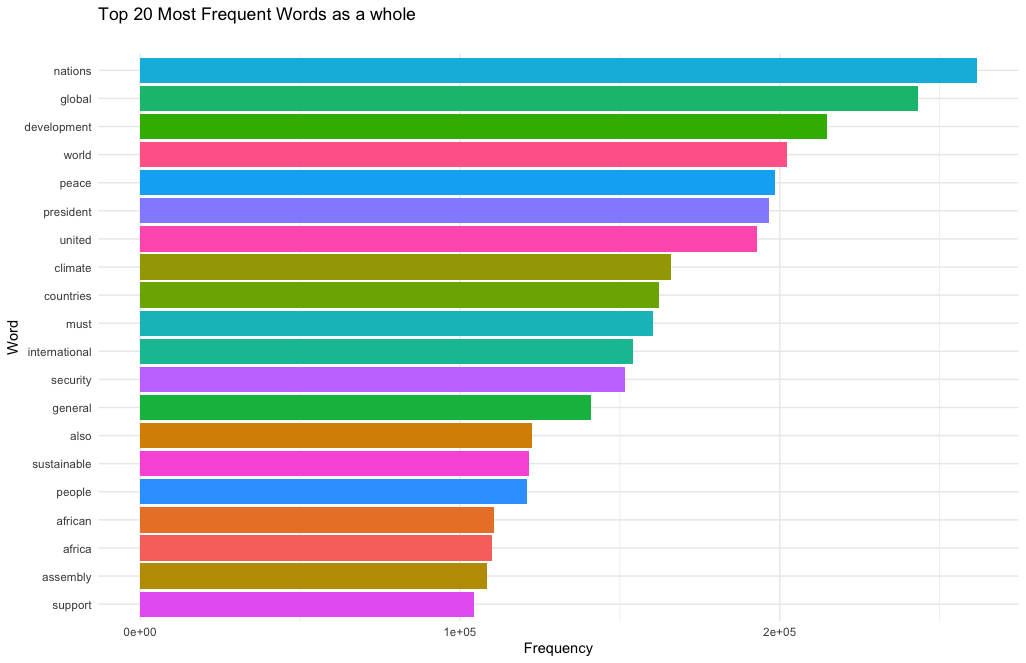


Figure 2.2:Top 20 Most Frequent words as a whole

Upon analyzing the top 20 most frequent words, it becomes evident that the words "nations," "global," "development," "word," "peace," "president," "united," and "climate" hold the highest frequency. However, the presence of these terms in isolation provides minimal assistance or guidance since they lack contextual information.

## word could

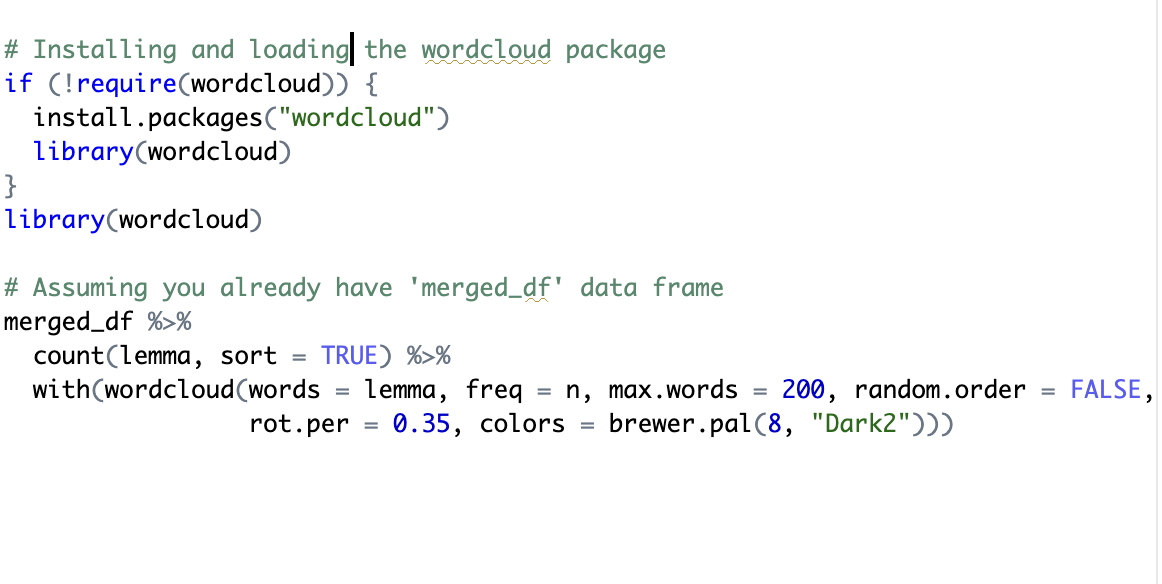


Figure 2.3: word could

## Summary

To summarize, the examination of word frequency, frequency statistics in the initial part yielded significant findings regarding the predominant usage of words in individual speeches. Nevertheless, it became apparent that specific terms acquired value alone when used in conjunction with others. To tackle this issue, the subsequent phase of the study placed emphasis on the identification and extraction of significant keyword combinations, recognizing the significance of contextual factors in facilitating a thorough comprehension of the claims. This methodology enriches the comprehensiveness of our analysis by encompassing subtle nuances that may not be readily discernible through the sole utilization of isolated word frequency metrics.

# Key word Identification

Frequency statistics of words are revealing, but one may find words which only make sense in combination with other words. Hence the goal of finding and extracting keywords which are a combination of words. The udpipe R package provides three method to identify keywords in text :

* RAKE (Rapid Automatic Keyword Extraction)
* Collocation ordering using Pointwise Mutual Information
* Parts of Speech phrase sequence detection

Both RAKE and PoS techniques are used to generate rankings of common keywords across all combined speeches. Using different algorithms, for the same purpose, are a useful ways of testing if different models perform in an expected, comparable way:

## Using Rake method

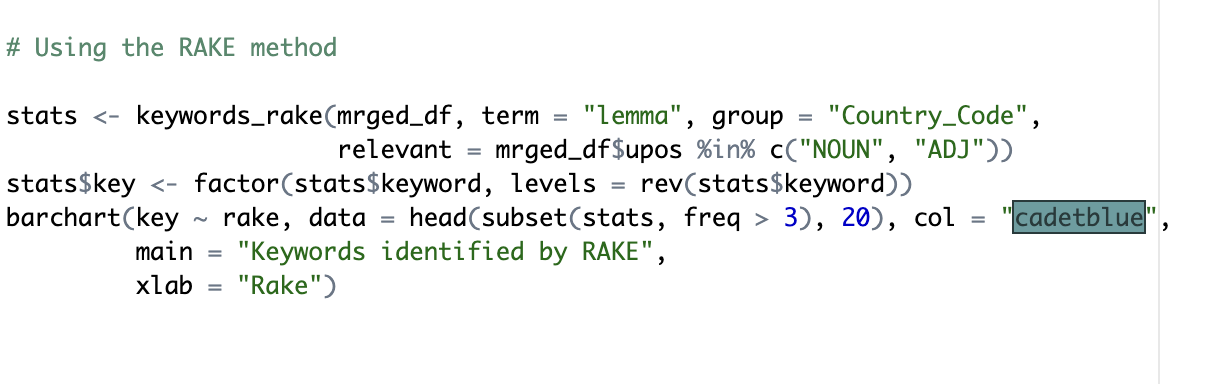
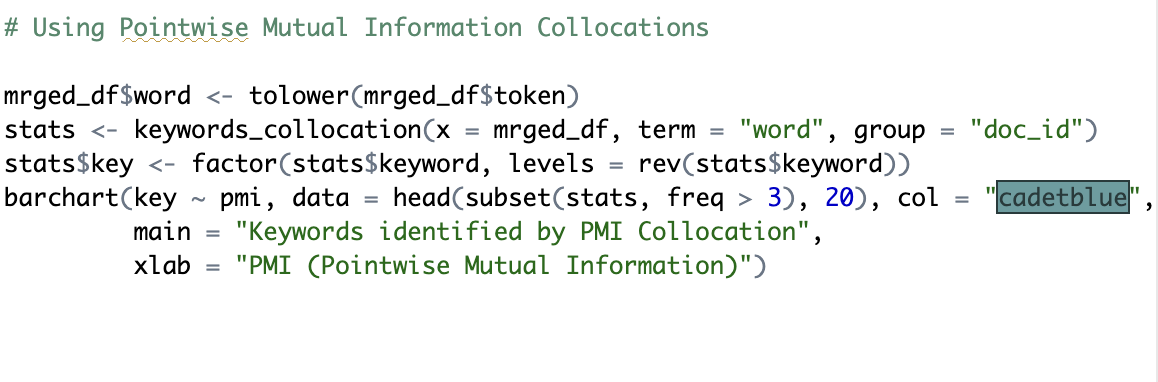




Figure 3.1:Key word identified by RAKE

The RAKE algorithm was employed to identify the prevailing keywords, which encompassed terms such as "lady" and "gentlemen," "United Nations," "climate change," "sustainable development," "international law," "global community," and "climate summit."

## Pointwise mutual information



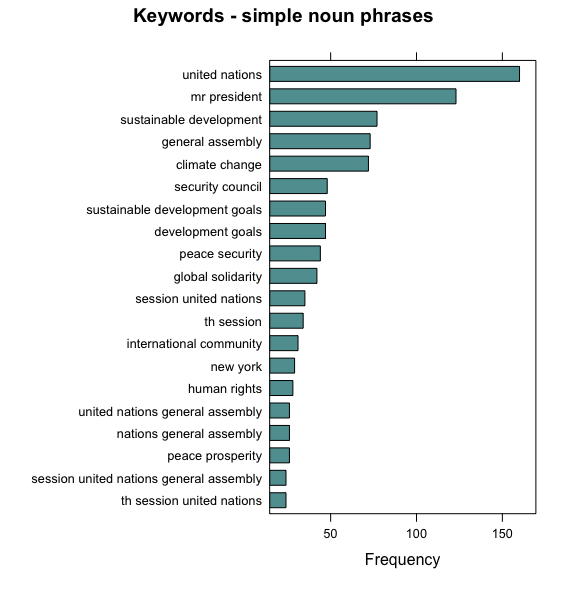
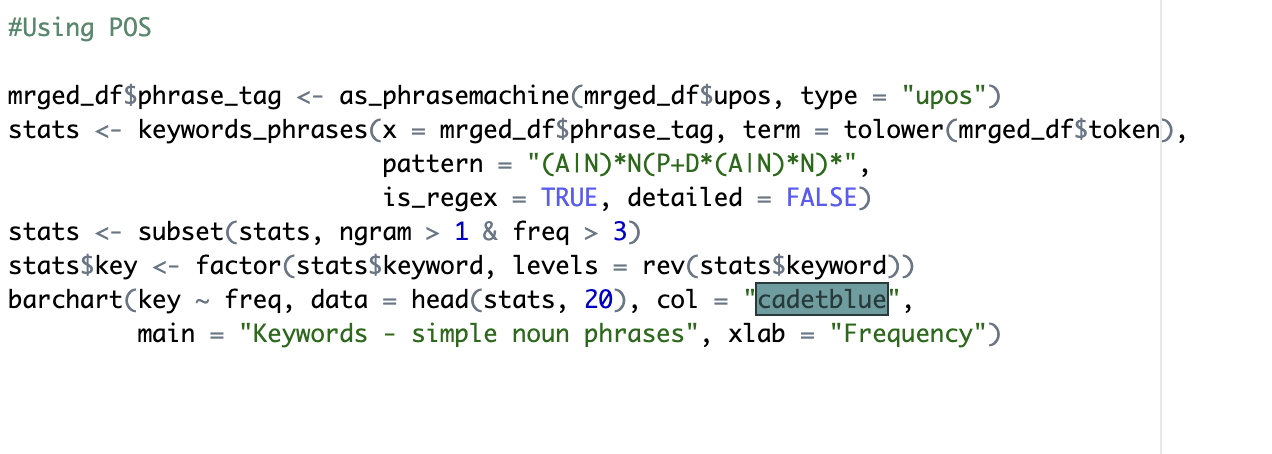


Figure 3.2: PMI (Pointwise Mutual Information)

Utilizing the PMI (Pointwise Mutual Information) framework, the prevailing keywords identified encompass the United Nations, the President, sustainable development, the General Assembly, climate change, the Security Council, sustainable development goals, development goals, and peace security.

## Using POS method



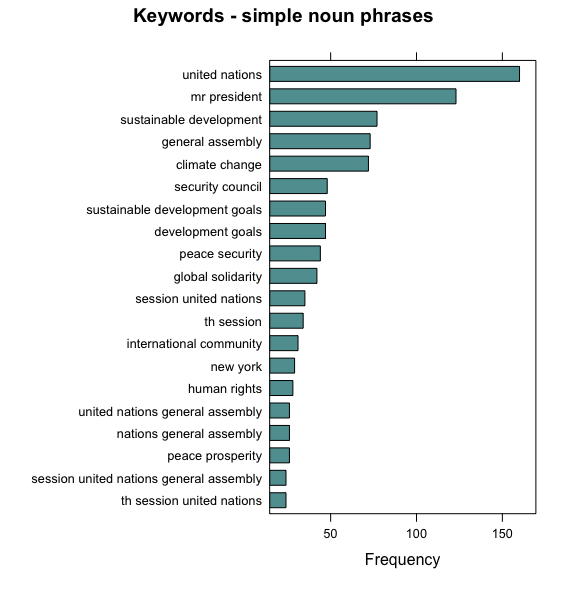


Figure 3.3: Part-of-Speech (PoS)

The results obtained with the Part-of-Speech (PoS) method exhibit similarities with those obtained through the Pointwise Mutual Information (PMI) method. The most frequently occurring terms identified are "United Nations," "Mr. President," "sustainable development," "General Assembly," "climate change," "Security Council," "sustainable development goals," "development goals," and "peace security."

# Analysing n-gram

An n-gram refers to a consecutive sequence of n words extracted from a given text. For instance, a bigram is a combination of two words, where the value of n is equal to 2. This analysis provides an initial examination of the occurrence frequencies of the most commonly observed bigram (n=2) and trigram (n=3).

## Nouns / adjectives used in same sentence

A screenshot of a computer program

Description automatically generated

A diagram of lines and words

Description automatically generated with medium confidence

Figure 4.1: Nouns / adjectives used in same sentence

## Nouns / adjectives which follow one another

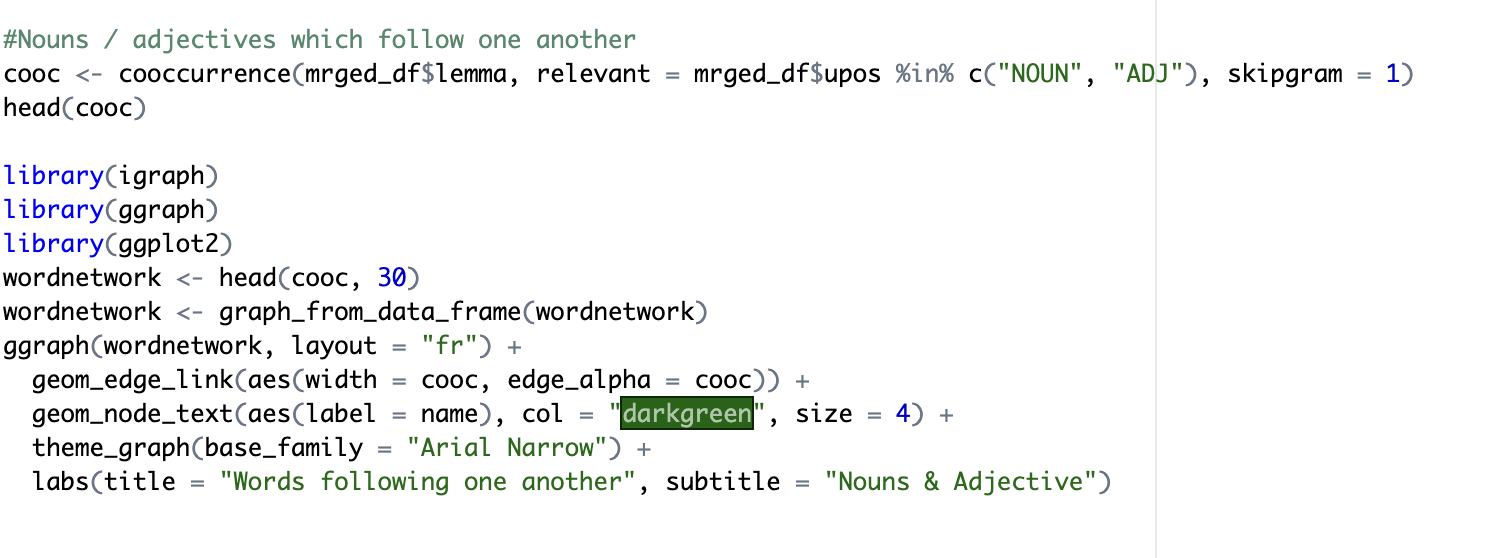
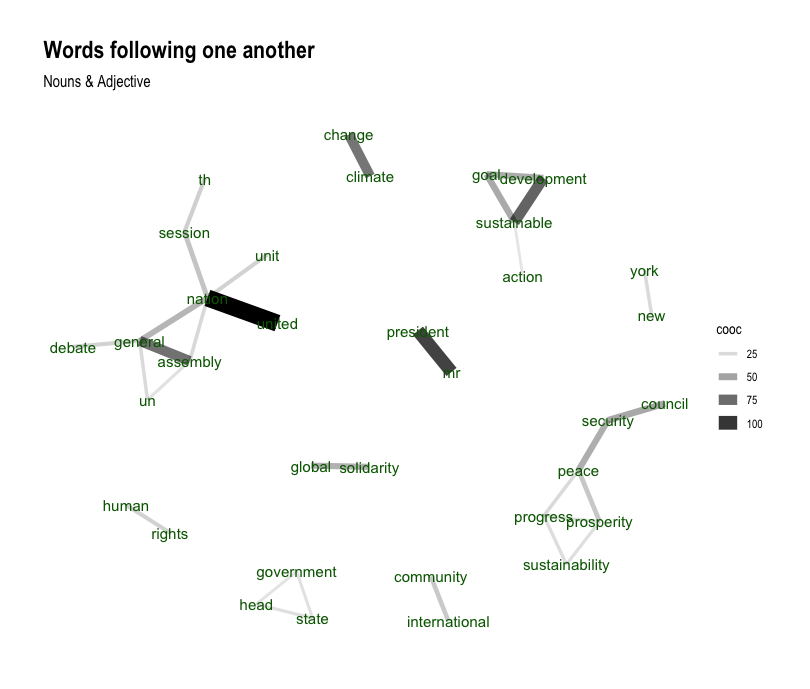


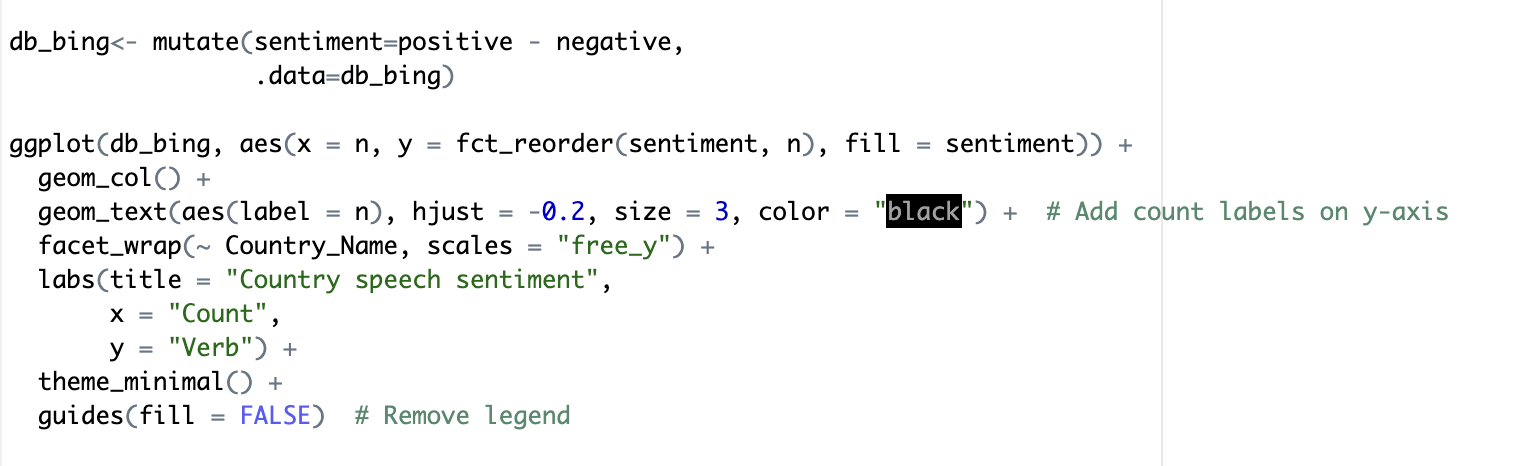
Figure 4.2: Nouns / adjectives which follow one another

## Summary

In contrast to examining individual characteristic words, the analysis now shifts focus to characteristic n-grams per document. This approach visualizes combinations of words that are most representative of each country's statement. The resulting network graph illustrates common co-occurrences, reaffirming findings from previous analyses (RAKE, PMI, PoS). Notably, terms like 'President,' 'Sustainable development,' 'Progress,' 'Prosperity,' 'Peace,' 'The government,' 'Global solidarity,' and variations of 'United Nation' consistently emerge. This aligns with the earlier frequency analysis, highlighting the persistence of key thematic elements across different analytical perspectives.

# Sentiment Analysis

From this analysis we can tell the sentiment of each speech as delivered by each country.



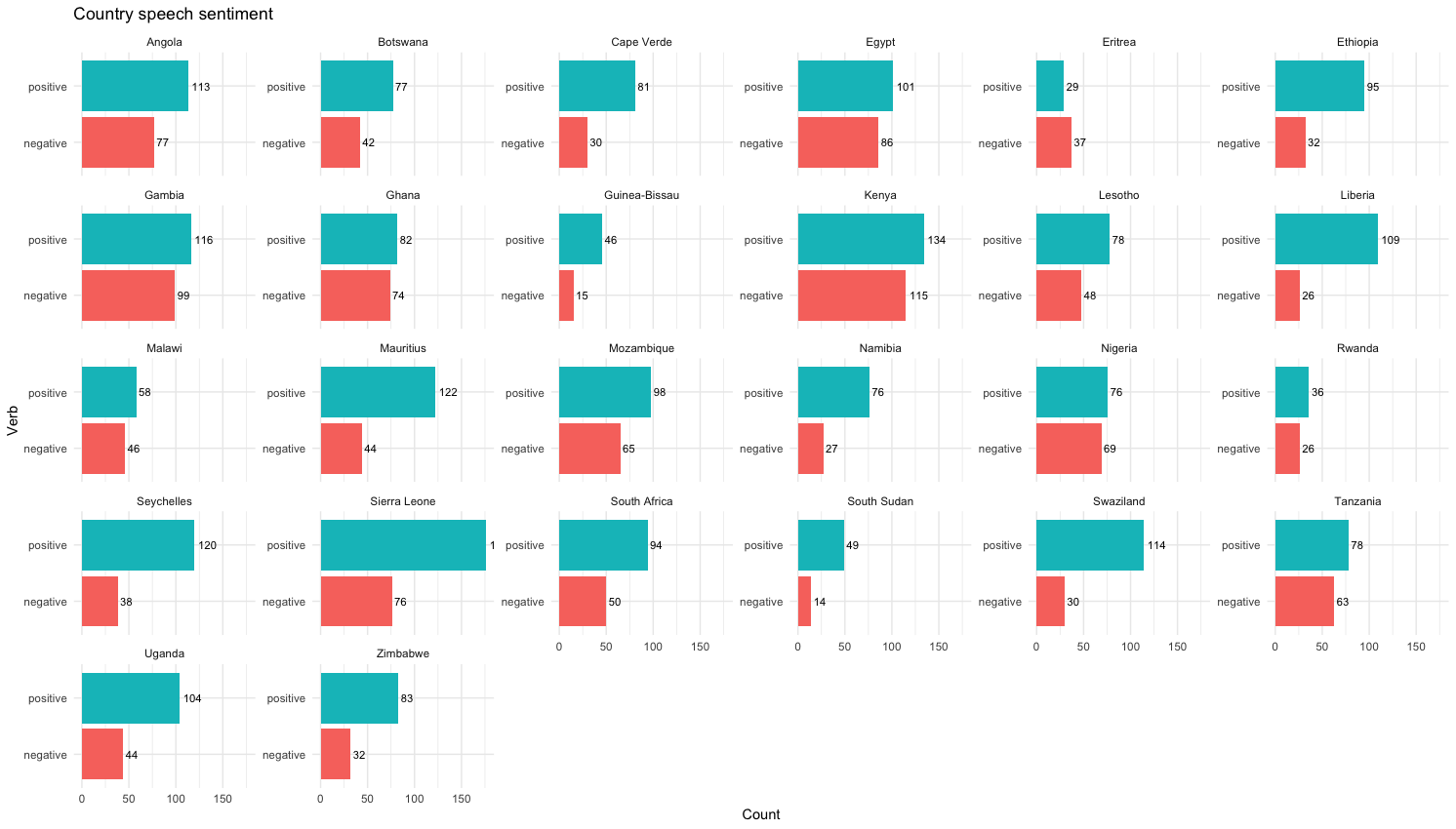


Figure 5.1: Speech Sentiment

Upon observation of the aforementioned figure, it becomes evident that the speeches, on average, had a favorable emotion. Examining Eritrea's speech is of significant importance, as it stands as the sole speech expressing a negative viewpoint.