# Assignment 2: South African Media Agency Twitter: Sentiment & Topic Modeling

21013527 (K. van Antwerpen), 21670897 (T. Luyt), 20073445 (F. Cilliers)

## Contents

# 1. Background

South Africa has seen an increase in COVID-19 recently and that has been labelled the third wave. The statistic of a third wave has moved the country into another level 4 lockdown. The news on statistics and general COVID-19 related happenings in South Africa are usually reported by their media agencies. The media agencies use many platforms online to repost their articles such as websites and mobile apps. Twitter is a platform where they can post their headline with a link to the article. Using sentiment analysis and topic modelling, we will compare how sentiment and topics have changed over time and how they compare to other media agencies.

# 2. Analysis Process

Twitter posts will be extracted using the "rtweet" package for the programming language "R". By using the get_timeline method in the package we can extract the latest 3200 Tweets from a specific user without premium. In this case, the users will be a selection of top South African media agencies. 3200 Tweets will give us 2 months' worth of data per media agency. Relevant COVID-19 Tweets will be extracted from that data. We use VADER to conduct sentiment analysis on the Tweets extracted. We choose VADER over sentimentR, as VADER has been academically proven to provide a more accurate sentiment on Tweets. VADER will give us a positive, negative, neutral, and compound metric on the Tweet. Compound is the Tweets overall sentiment. We then conducted topic modeling using LDA. Most of our process comes from working through Text Mining with R

## 2.1. Reasons for Media Agencies Choosen

Justify what makesa new source good. ADD FIGURES *News24:* Recognized by APP Annie (App Annie is the standard in app analytics and app market data) as the most known South African internet media source. *Times Live:* Claim to be South Africa's second-biggest news website, published by Arena Holdings (Times Live website). No evidence found to disprove this claim. In top 10 of most visited publication websites for South Africa. *Daily Maverick:* Boasts free, fair, and fearless reporting. *eNCA:* In top 10 of most visited publication websites for South Africa. *SABC News:* National news company with government ties. Reaches a wide variety of viewers in different languages. The company is both state owned and a public broadcaster company.

# 3. Giving Context to our Data and Tidying.

We first clean the original Tweets to remove unique News based language. Media agencies often lead their Tweet about an article with the category it belongs to, eg. OPINION, BUSINESS, WATCH.

Some tweets fetched date further back, as seen in figure 4.1. The 3200 tweet pull per user causes this. It tells us that agencies like News24 and SABC News Tweet more daily than Daily Maverick.


Figure 3.1: Post Count by Day

Next, by modeling bigrams and trigrams for our dataset, we get a better understanding of what topic is being discussed with each word, as shown in figure 3.2. Trigrams were modeled but were not necessary as bigrams provided enough information. We can then also see which sentiments are incorrectly labeled. “not good” gives better context of a negative sentiment, rather than it being incorrectly identified as positive good, as shown in

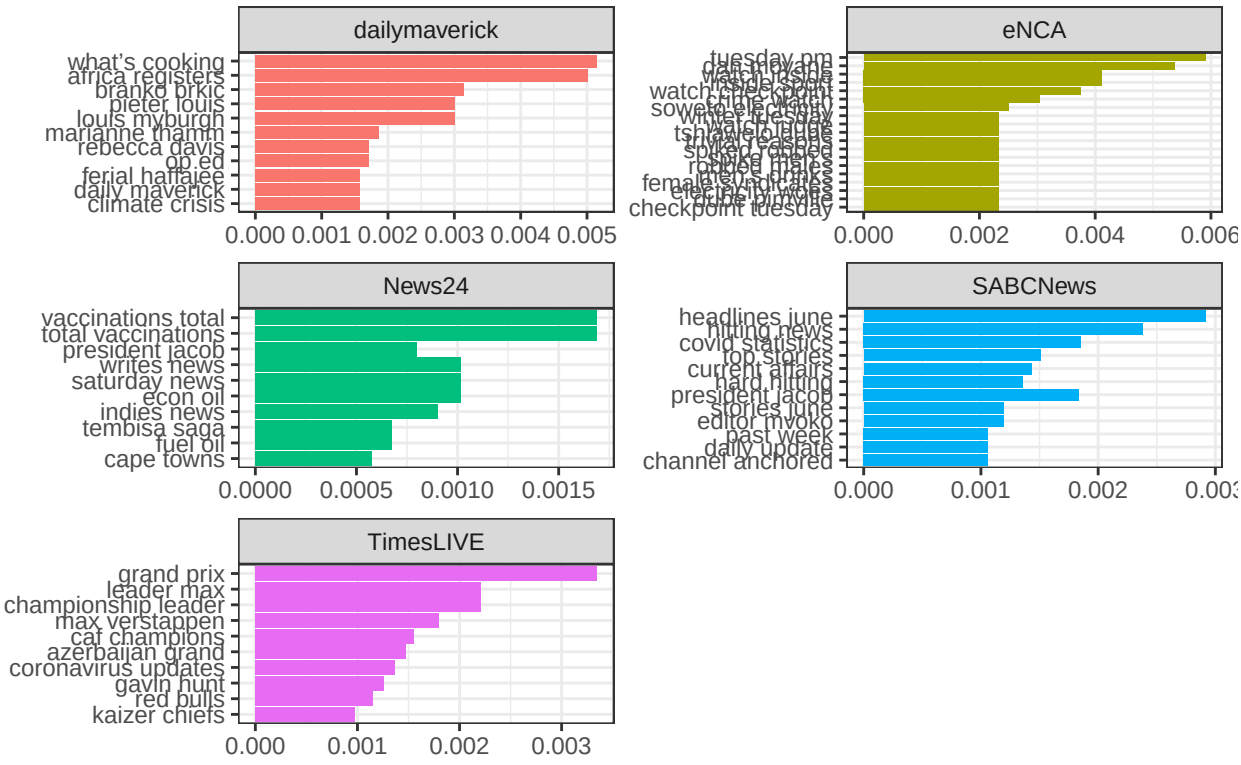



Figure 3.2: Bigrams

figure 3.3.

We give more weight to words that appear more often with the incorrect sentiment. The graph below (Figure 3.3) shows that 'no good' or 'not impressed' have the highest weight of being mislabeled as positive, and vice-versa for negative words like guilty. We remove these words to increase the accuracy of our sentiment analysis. We can now tidy our dataset based on tidyverse's stopwords collection, and our own negation word collection.

```
## i Using "','" as decimal and "'.'" as grouping mark. Use `read_delim()` for more control.

##
## -- Column specification --------------------------------------------------------
## cols(
##   TOKEN = col_character(),
##   `MEAN-SENTIMENT-RATING` = col_character(),
##   `STANDARD DEVIATION` = col_character(),
##   `RAW-HUMAN-SENTIMENT-RATINGS` = col_character()
## )
```

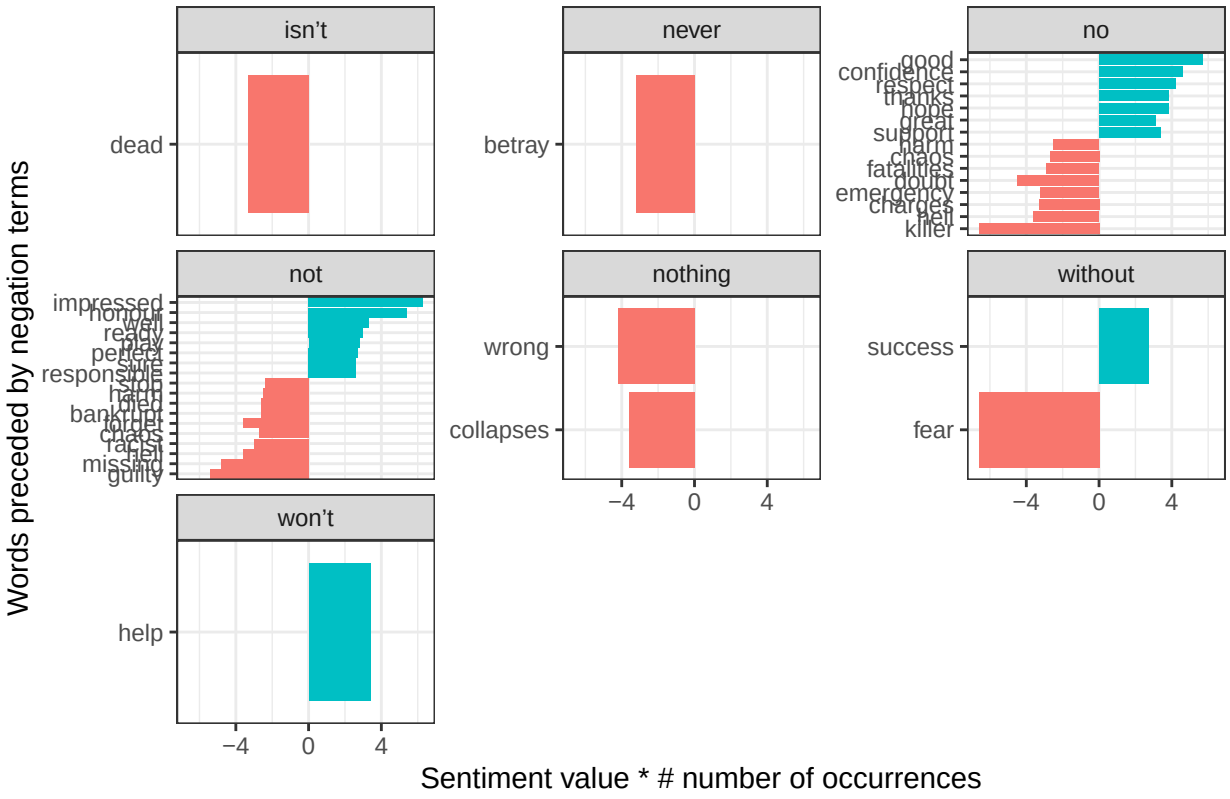



Figure 3.3: Negation Words

We tokenize the Tweets and remove stop words in tidytext, our own stopword dictionary, and negation word dictionary. We also use the "twitter" token to handle any left over @'s and URLS.

From our tidied Tweet dataset, we look for the top words that appear (Figure 3.4). It gives us a good idea of what topics are being discussed the most. We find that COVID-19 has been the main topic of discussion. President appears second as President Ramaphosa of South Africa usually addresses the nation regarding COVID-19 information. Additionally, Zuma also appears as he is mentioned as "former president Zuma" in most articles. Zuma appears more as his recent court avoidance and sentencing is being Tweeted. General words surrounding the COVID-19 topic as it is still the main pressure on the country, especially involving Gauteng's rise in infections.

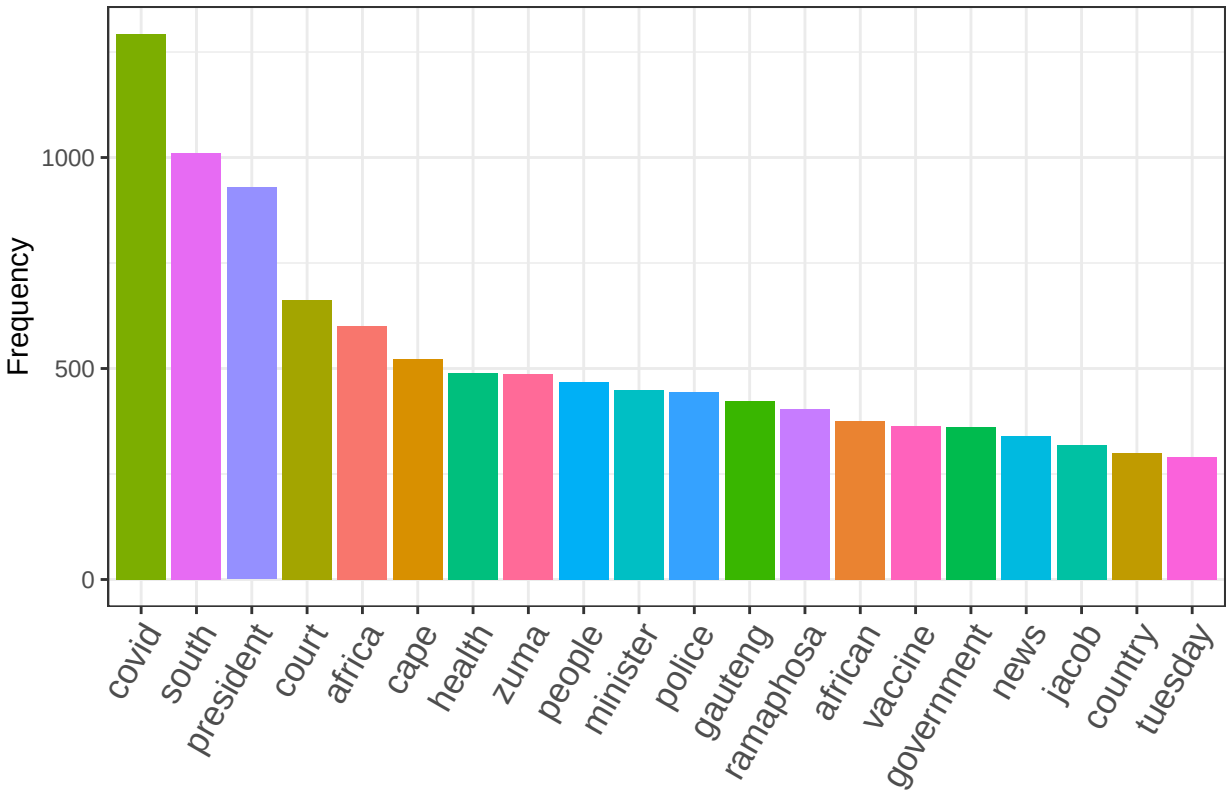



Figure 3.4: Most Frequent Media Agency Words

A tf-idf is then modeled to determine which words are the most important per media agency (Figure 3.5). We also model the word importance by week and determine which media agency has the most unique topics of the week. Daily Maverick is dominating the first four weeks as they are the only user with Tweets from that time.

```
## Selecting by tf_idf
```

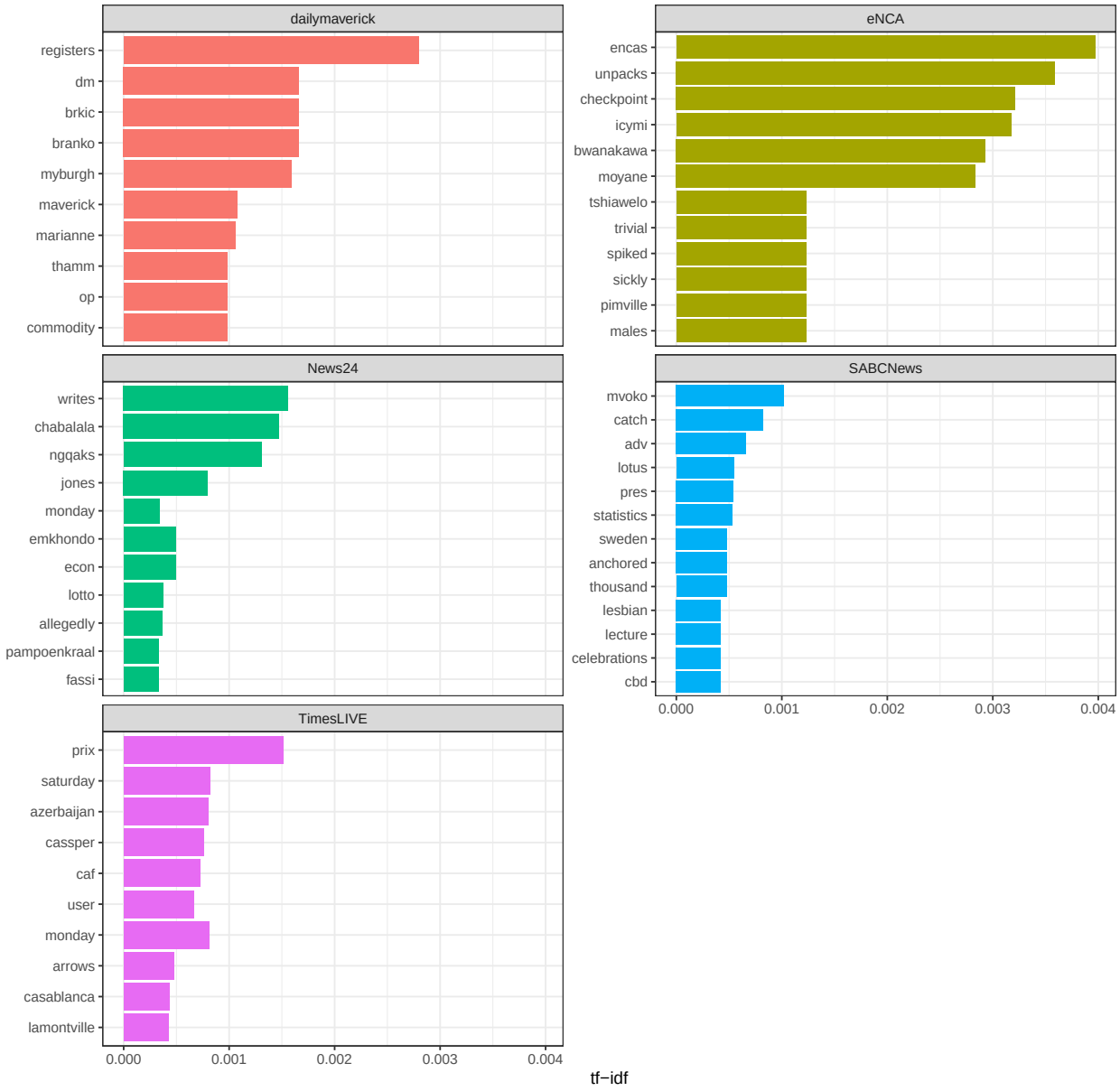



Figure 3.5: Highest tf−idf from each Media Agency

Tweet topics involved with the tf-idf's listed in the weeks collected (Figure 3.6):

- 18. Nazanin's struggles to breathe while Narenda mMdi's government abdicates all responsibility.
- 19. Newspaper is on sale now in and free to loyalty card holders so go grab a copy henni has movies for everyone so maybe ill see you there. extinction rebellion stop putting activists on trial it isn't in the public interest.
- 20. Covid ready for liftoff SA's vaccine mission impossible.
- 21. Springboks and Bafana Bafana face loss of players due to injury and covid. cabinet approves bill to strengthen sabcs finances management.
- 22. Expenditure uncovered through onfidential internal audit. exposed digital vibes bankrolled maintenance work paid to minister's son.

- 23. Concern hit sprinboks due to injury leading up to british and irish lions series. F1 grand prix endured four red flags due to a large amount of crashes around the track.
- 24. Danish footballer Christian Eriksen had a heart attack on the field in a recent Euros match. Inter Millan team mate, Lukaku, send out message of support. Youth day keynote address delivered by the president, cyril ramaphosa.
- 25. Dr. Mary Kawonga on hostipal capacity and Eskom's electric grid failures.
- 26. Student found death outside Walter Sisulu University. Rape survivor, Andile Gaelesiwe, has released a new book with her foundation.
- 27. The appearance of Jacob Zuma in the constitutional court. Jacob Zuma supporters move in a motorcar from Durban to Nkandla to offer support to the former president for contempt in the constitutional court, ahead of his arrest.

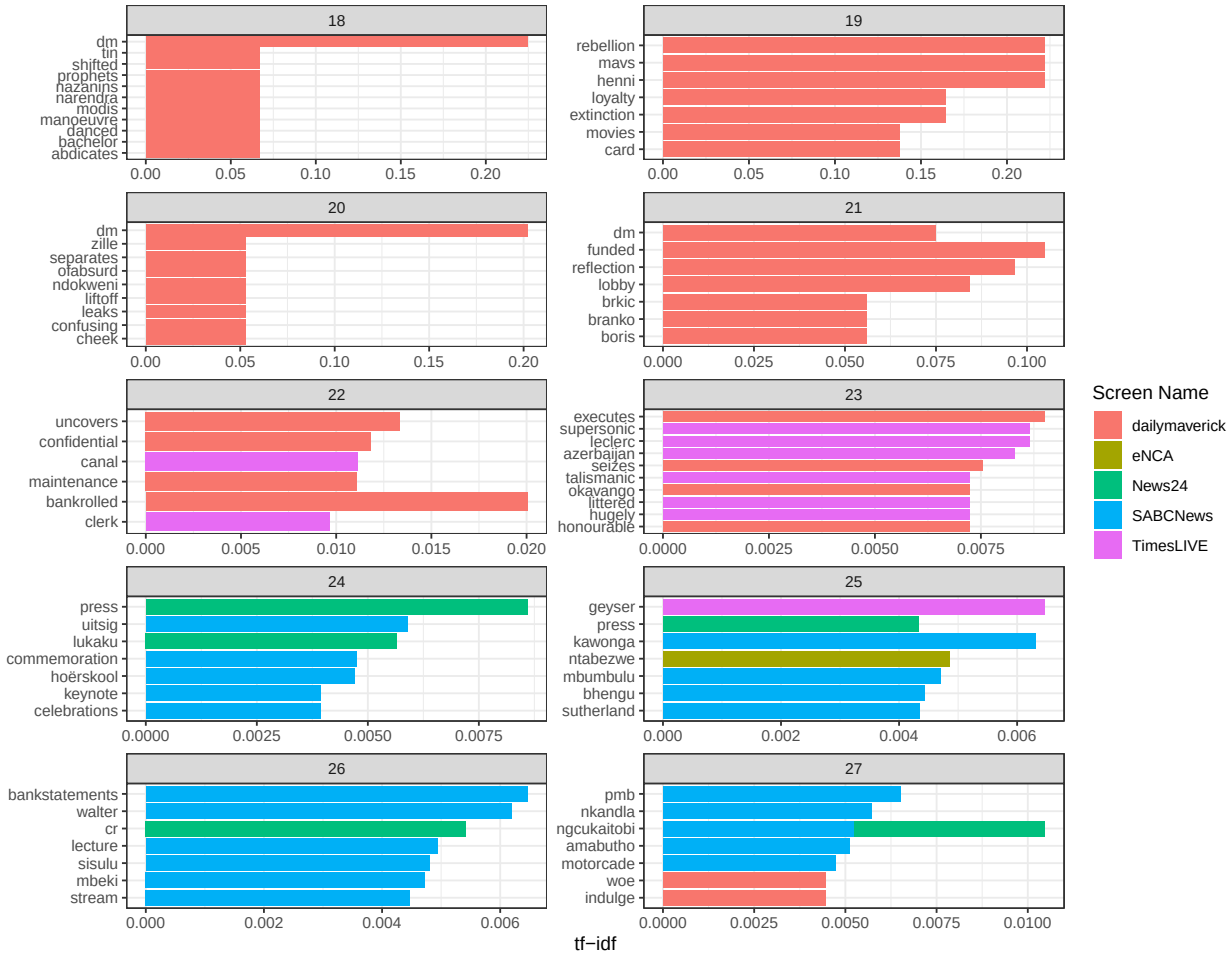



Figure 3.6: Highest tf−idf by Media Agency in Each Week

The figure (3.7) below shows which words are most likely to co-occur. We use this to understand the context of words in our topic modeling. The words shown are popular words in our dataset.

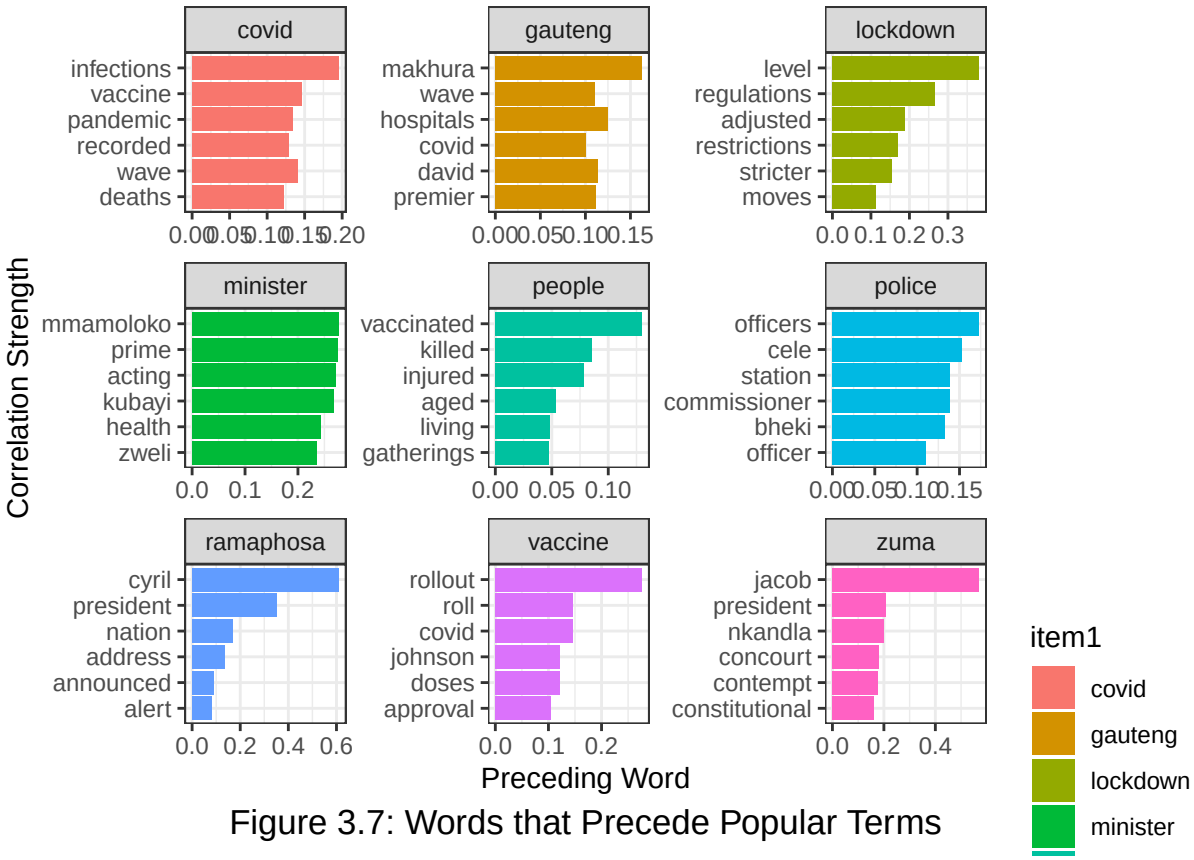



Figure 3.7: Words that Precede Popular Terms

# 4. Sentiment Analysis

## 4.1. Sentiment over time

The general sentiment over time is mostly negative (Figure 4.1). News articles use negative headlines to get a faster reaction from people when skimming through news. Positive News usually involve sport headlines.

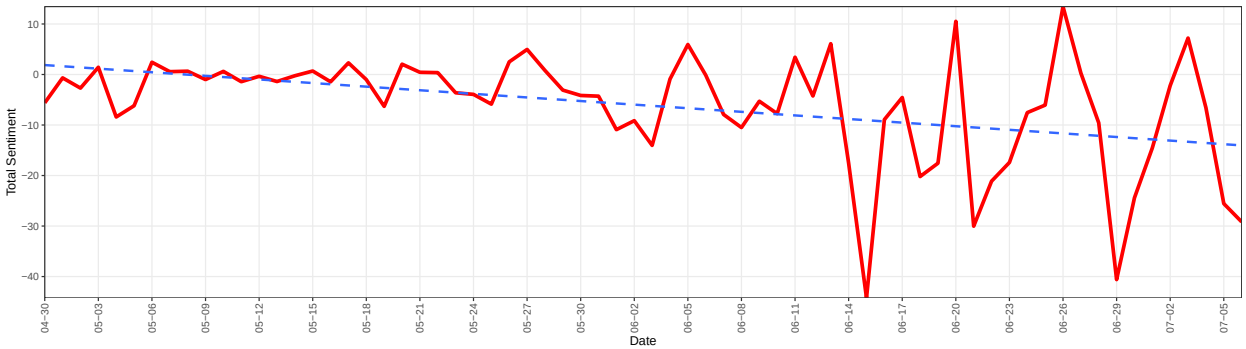



Figure 4.1: Sentiment Over Time by Media Agency

### 4.1.2. Most Negative Tweet

An illegal gold miner who was severely injured in a clash in which four other illegal miners were killed has been charged for their murders, Mpumalanga police said on Thursday.

### 4.1.3 Most Positive Tweet

Daily Maverick; Food for Mzansi get the nod at the Global Media Awards: @foodformzansi gets 3rd place in Ad Campaigns; honourable mention for Best Use of Audio @dailymaverick gets honourable mentions for Reader Engagement; Best Use of Print.

## 4.2. Sentiment Over Time per Agency

Daily Maverick, although decreasing, has the straightest regression line. It also has the least outlying total sentiment. We could conclude that this agency is the least bias to emotion. News24 has the most positive and positive daily news.

```
## `summarise()` has grouped output by 'created_at'. You can override using the `.groups` argument.
```

```
## `geom_smooth()` using formula 'y ~ x'
```

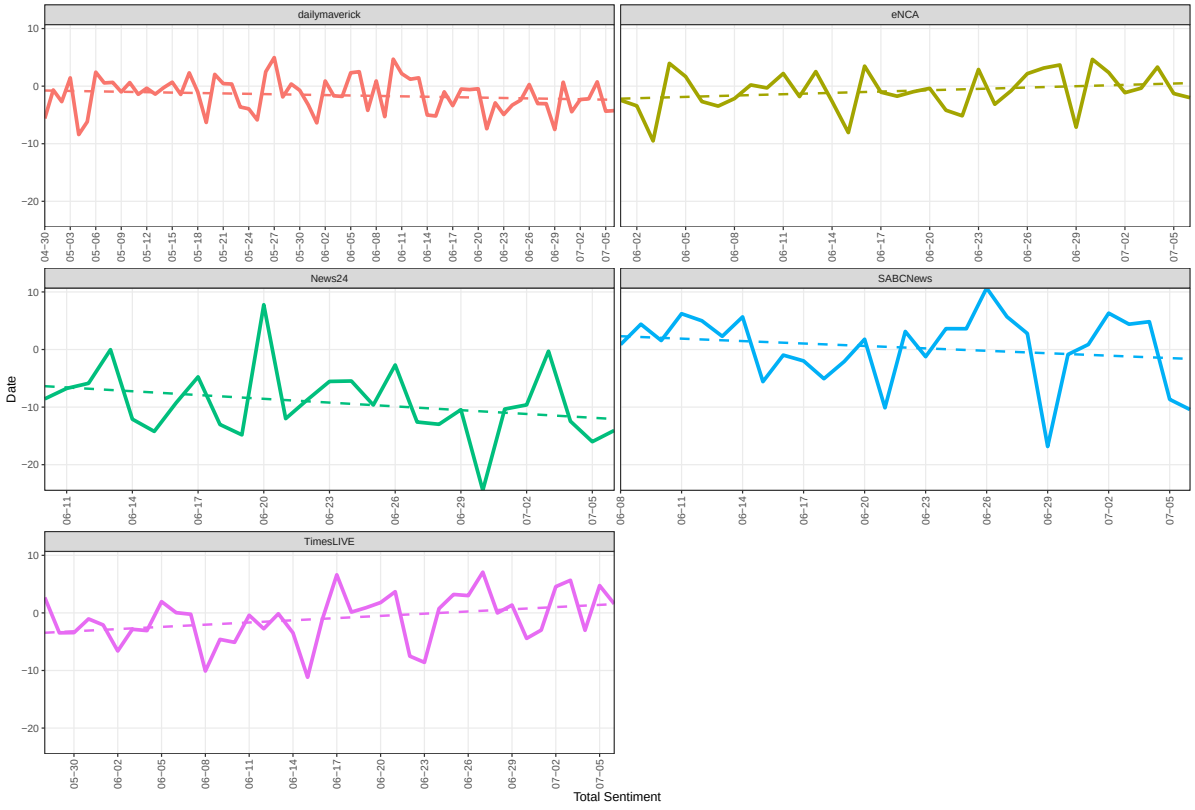
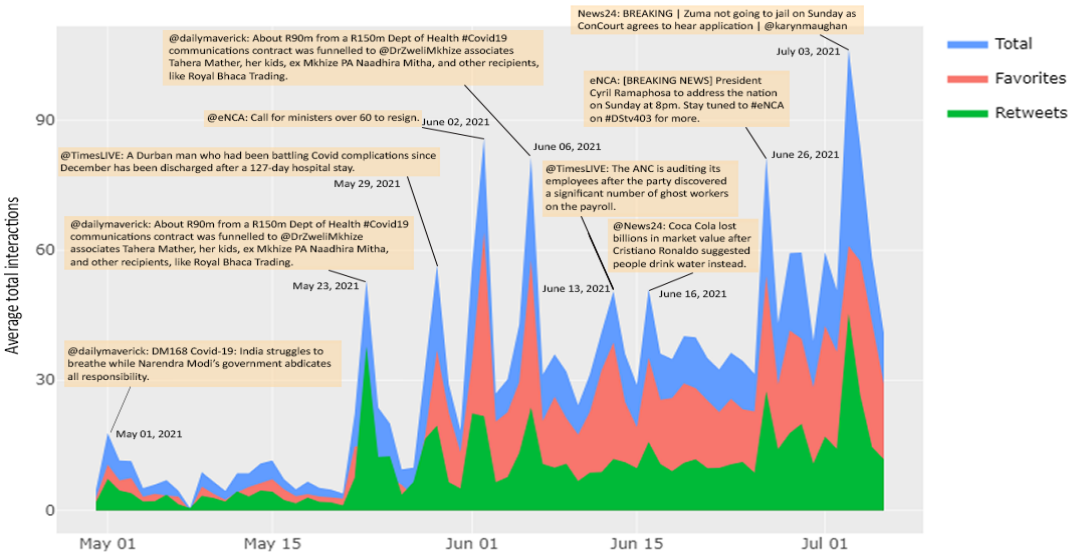



Figure 4.2: Sentiment Over Time by Media Agency

# 5. Interactions

The plot below maps the average total interactions by day. From the relevant peaks, we extract the top tweet in that day.




## Peak Tweets by User

*@eNCA:* Call for ministers over 60 to resign.

*@dailymaverick:* SCORPIO Floyd Shivambu's brother quietly pays back R4.55m, admits he received the VBS money gratuitously.

*@TimesLIVE:* Do you approve of Duduzane running for president?

*@News24:* Coca-Cola lost $4 billion in market value after Cristiano Ronaldo suggested people drink water instead.

*@SABCNews:* BREAKING NEWS: King of Eswatini has fled amid public violence in the country.

# 6. Topic Modelling

A topic model is a type of statistic model for discovering the abstract "topics" that occur in a collection of documents. We determine the best range k-value for LDA (Latent Dirichlet Allocation). The importance of the value of 'k' is to ensure you do not over estimate or underestimate the data. If you over estimate it, you will be left with topics that carry very little meaning and if you under estimate the data, you will lose out on topics that could have been useful to your research.

By looking at the lowest minimum and the highest maximum, we can determine the 'k'. This is how we have interpreted the graphs produced. Griffiths2004 is not informative in this situation and is therefore ignored.

This method used has been proved to produce the best results of LDA without subjectively tuning the 'k' value. Our result is 12 topics.

## 6.1. Gap k justification

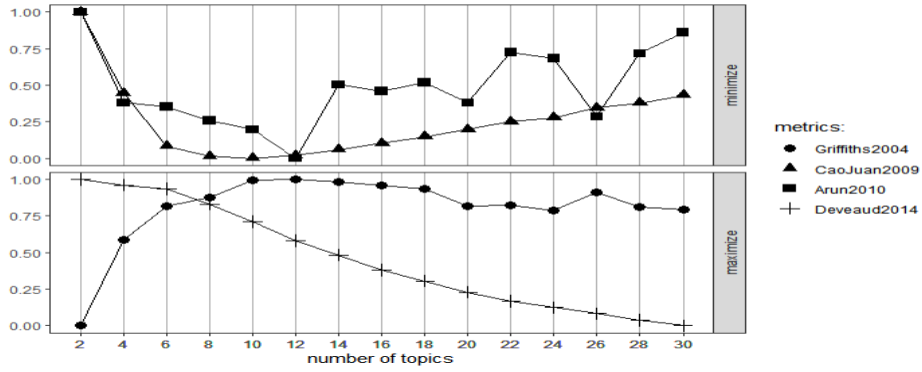



Figure 6.1: Justify K Value

## 6.2. Topics found

### 6.2.1. Beta

Our main topic influence is COVID-19, as this is what currently affects the country the most. Our words are mapped to a beta which shows the amount the word appears per topic.

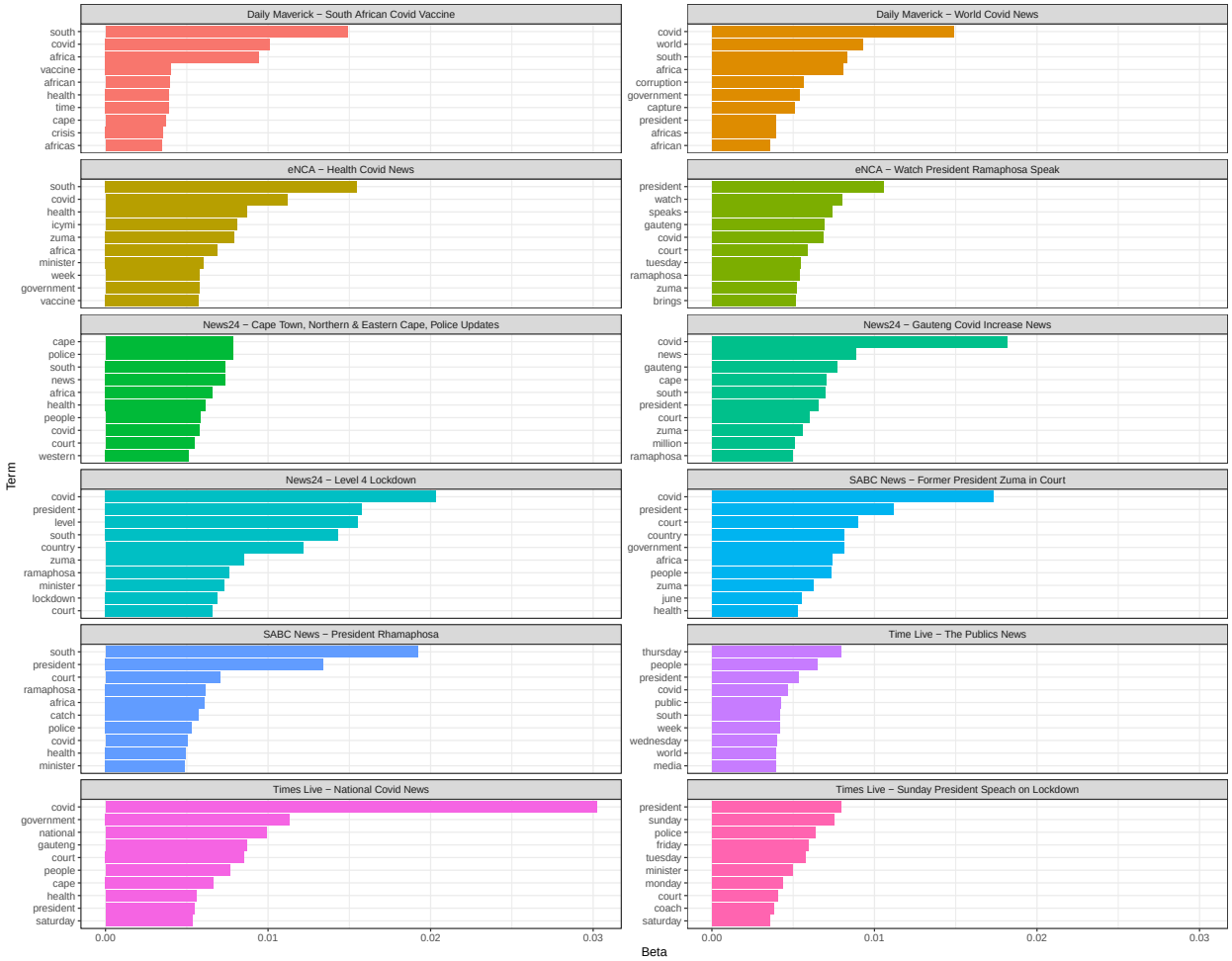
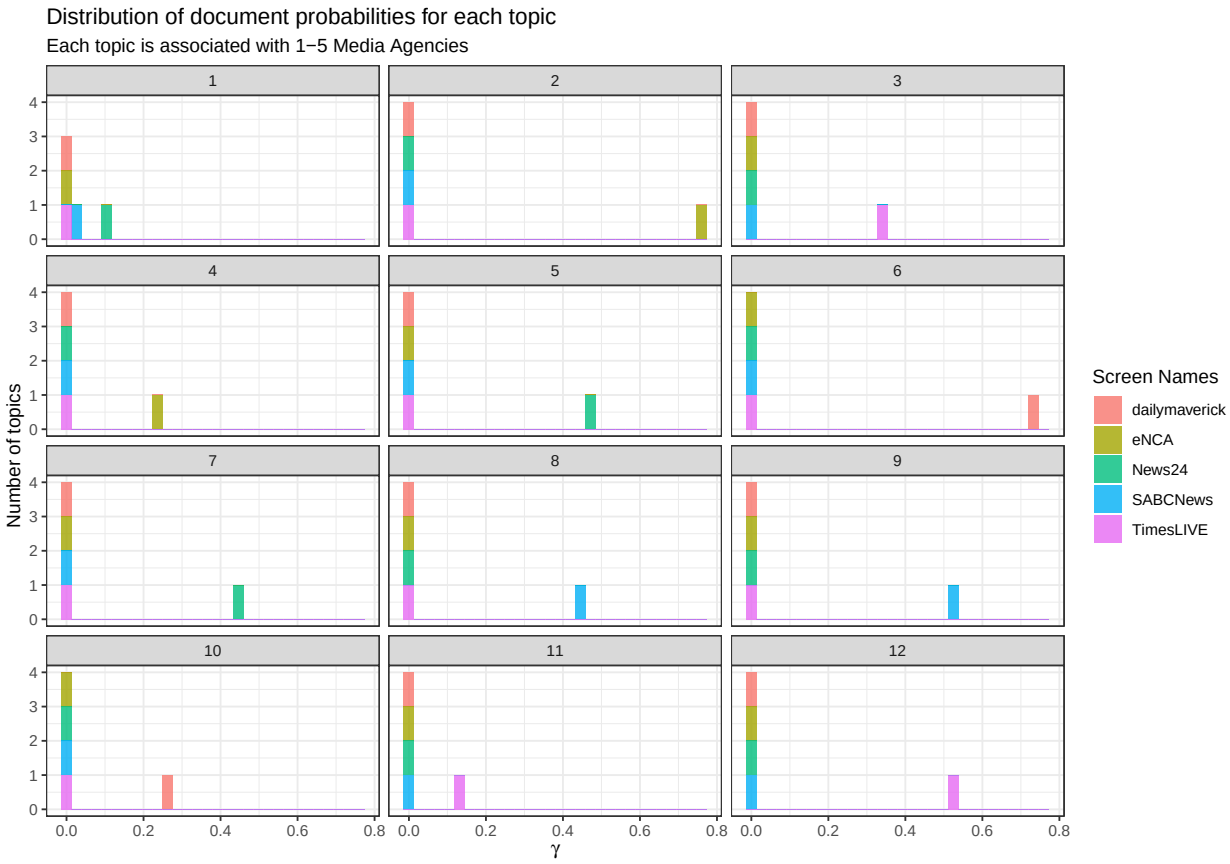



Figure 6.2: Modeled Topics (All Tweets)

### 6.2.2. Gamma

The gamma is added to the topics modeled. We can now see which agencies are strongly associated with each topic. The higher the gamma, the more strongly the specific agency associates with the topic.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




# 7. Additional Requirements

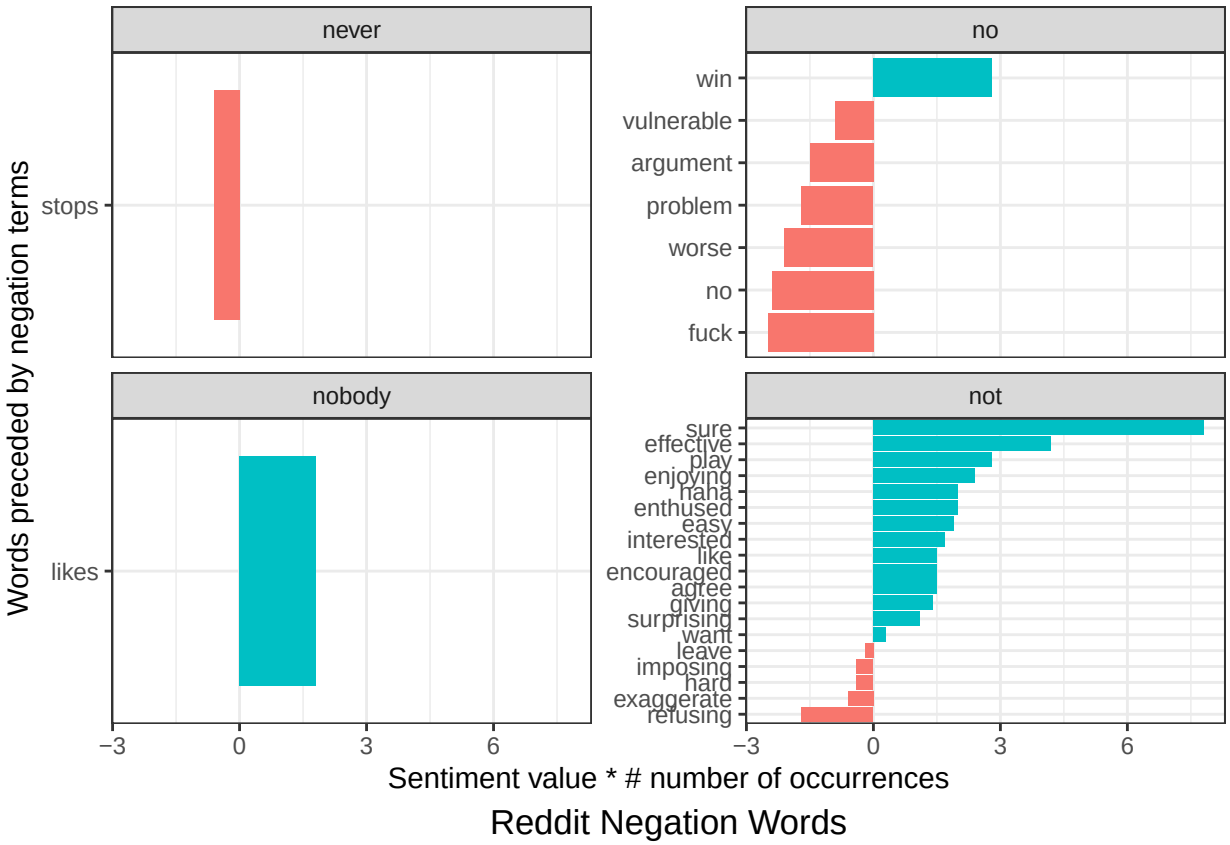
We choose Reddit as our other data source. Facebook was considered but API needing proof of identity with an ID document seemed excessive. The 'ReddicommentractoR' package is used to extract comment and post data. Search term 'Covid-19' is used. Other terms were not used as their post dates went further than 6 months. We most of the same analysis for Reddit data.

## 7.1. Reddit Comparing Comments From r/southafrica Covid-19 Posts.

We clean the data as we did before. Controversial and foul language is left in to not affect sentiment. The worst langauge is usually moderated within the Reddit communities before it is seen by thg public.

```
## i Using "','" as decimal and "'.'" as grouping mark. Use `read_delim()` for more control.

##
## -- Column specification --------------------------------------------------------
## cols(
##   TOKEN = col_character(),
##   `MEAN-SENTIMENT-RATING` = col_character(),
##   `STANDARD DEVIATION` = col_character(),
##   `RAW-HUMAN-SENTIMENT-RATINGS` = col_character()
## )
```




From our tidied Reddit dataset (Figure 7.2), we look for the top words that appear. This will give us a good idea of what topics are being discussed the most. We find that with COVID-19 comments being pulled, the comments discuss people's general interaction with the virus. The comments come from regular people and this shows what topics are discussed among people.

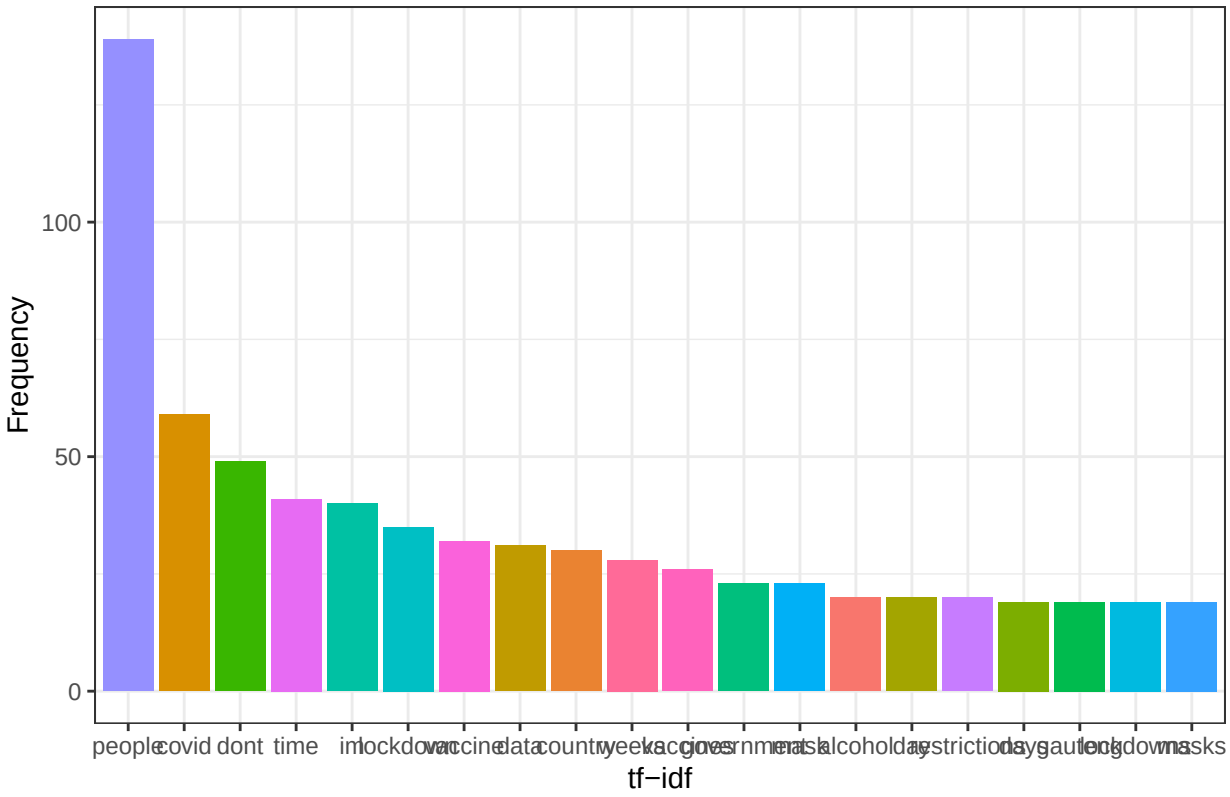



Figure 7.2: Most Frequent Reddit Comment Words

A tf-idf is then modeled to determine which words are the most important per media agency. We also model the word importance by week and determine which media agency has the most unique topics of the week. Daily Maverick is dominating the first four weeks as they are the only user with Tweets from that time.

```
## Selecting by tf_idf
```


Figure 7.3: Highest tf–idf from Title

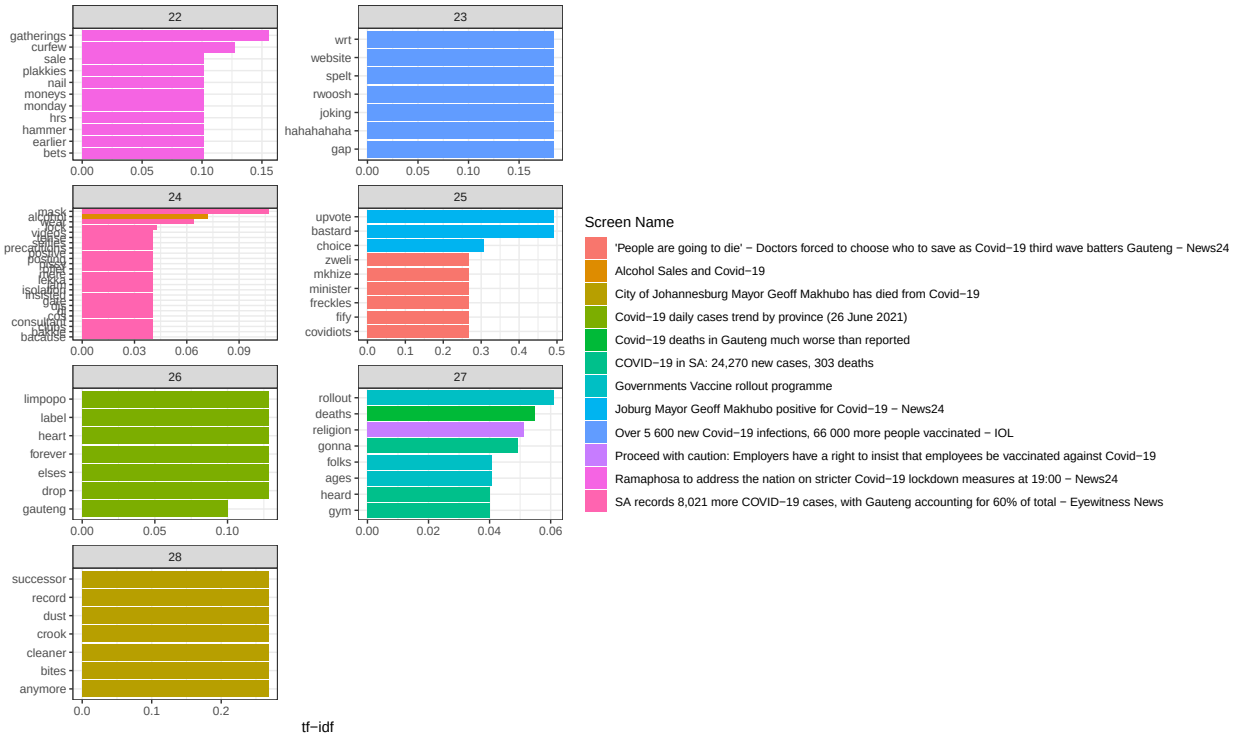



Figure 7.4: Highest tf−idf by Title in Each Week

The graph below shows which words are most likely to co-occur. We use this to understand the concomment of words in our topic modeling. The words shown are popular words in our dataset.

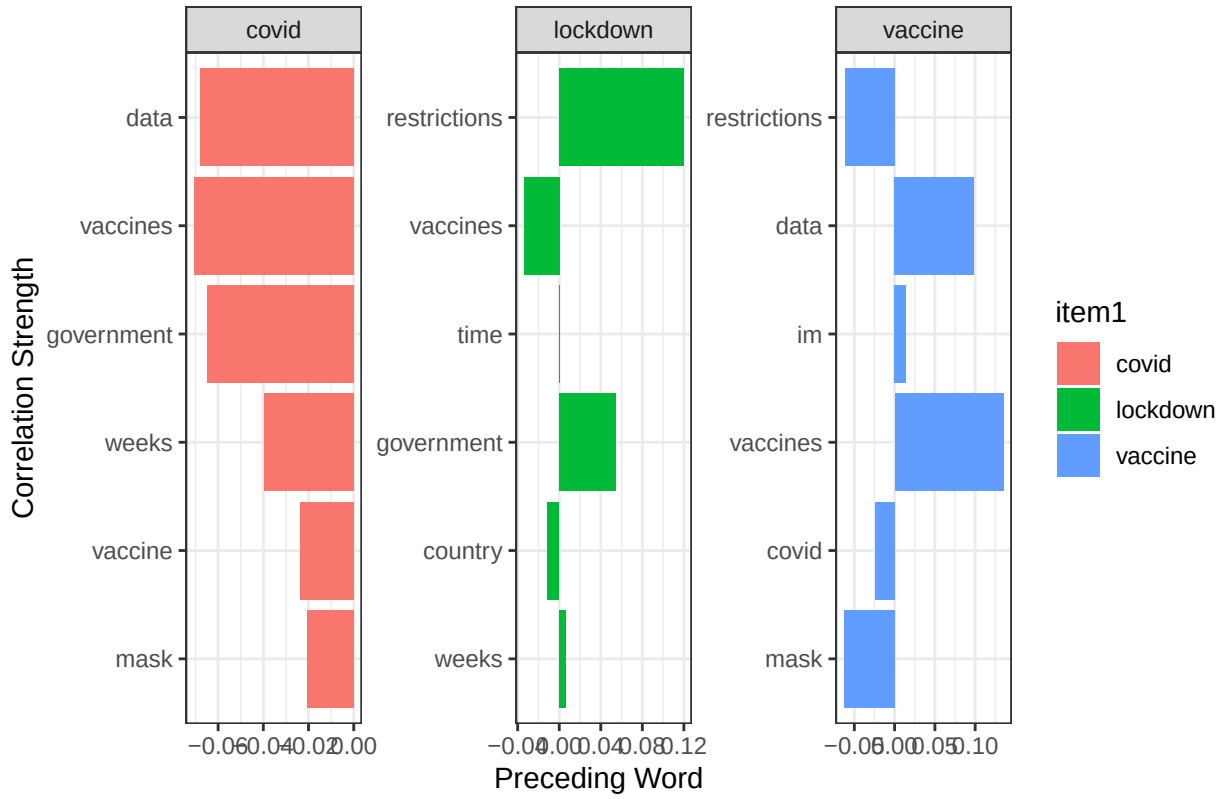



Figure 7.5: Words that Precede Popular Terms

```
## List of 1
##  $ legend.justification: num [1:2] -1 -2
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```

# 8. Reddit Sentiment Analysis

## 8.1. Sentiment over time

Reddit sentiment is sporadic. Everyone has their own opinion on a post title whether it is a positive or negative title.

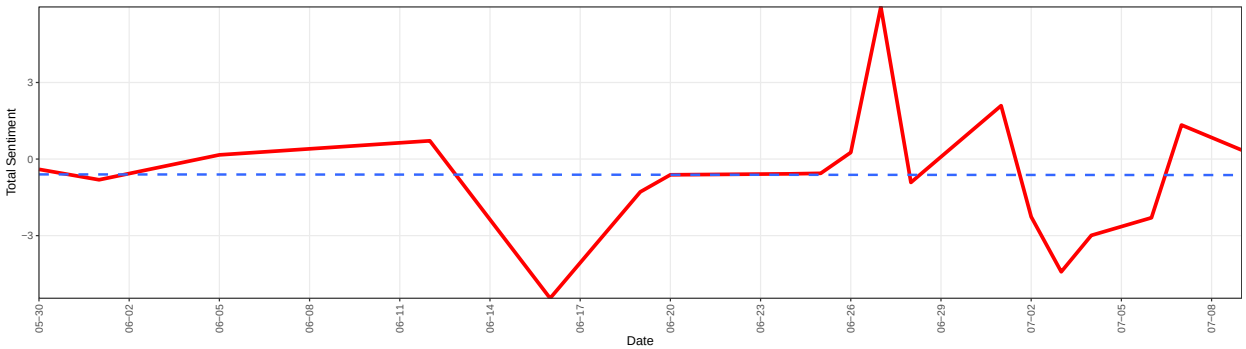



Figure 8.1: Sentiment Over Time by Media Agency

#### 8.1.1. Most Negative Comment

Lambpanties: it does but sheesh the poor people my dad rents property to a restuarant owner and the poor guy hasnt been able to pay his rent about or different months now one month my parents even pitched in to pay his staffs wages there is no way the poor guy is not going to go under from this cherry from hell on top he has covid right now to boot

#### 8.1.2. Most Positive Comment

babufrikhasaposse: but arts is a strong driver of an improved society you dont want to live in a society without arts and demanding other people meet a standard you set while pretending they dont contribute positively to society is a bit absurd not everyone is interested in science and the aim of education isnt economic value exclusively even if everything you said is true that doesnt make the comment i replied to mot just an asshole thing to say

# 9. Reddit Interactions

## 9.1. Highest Upvoted Comment from each Post

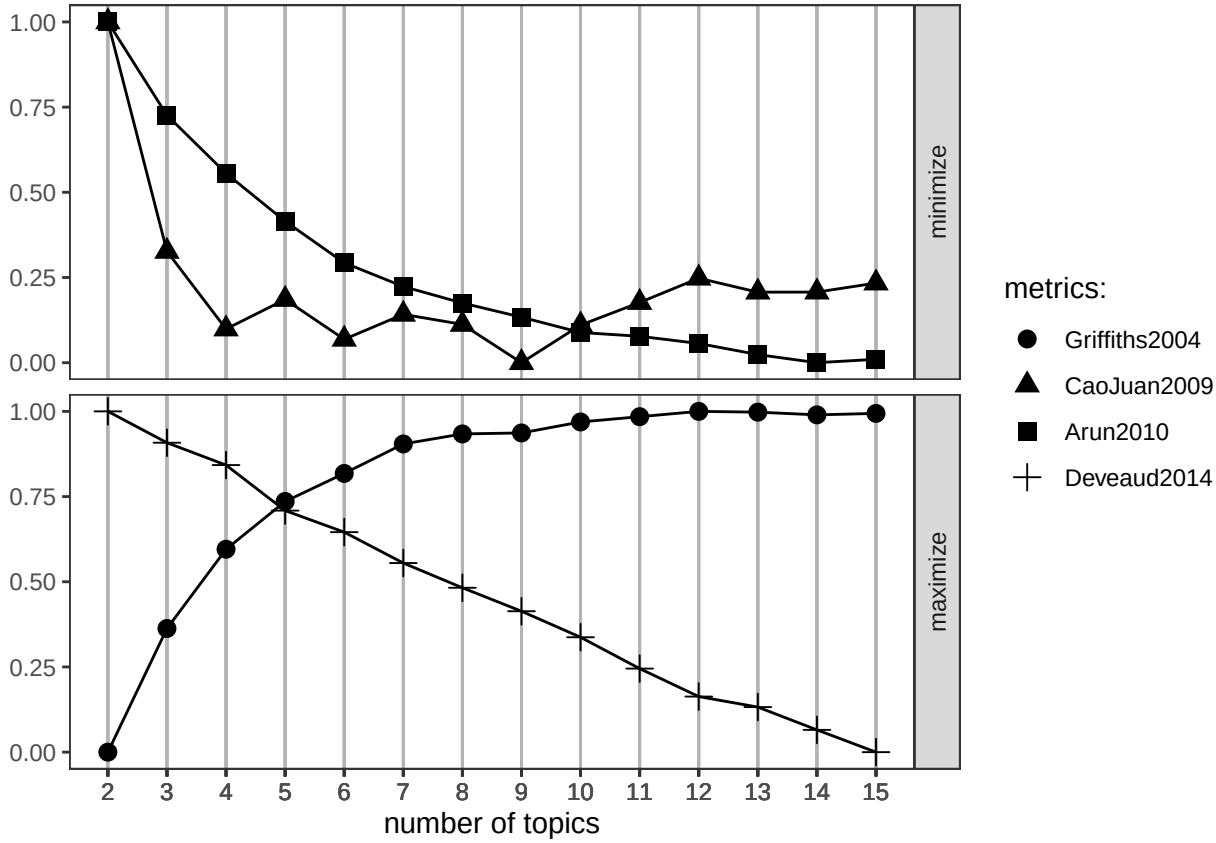
## 9.2. Controversial Comments

# 10. Reddit Topic Modelling

## 10.1. Gap k justification

Our result is 9-14 topics. From this we found 12 topics to be accurate.

```
## fit models... done.
## calculate metrics:
##   Griffiths2004... done.
##   CaoJuan2009... done.
##   Arun2010... done.
##   Deveaud2014... done.
```




## 10.2. Topics found

Topics of discussion in r/southafrica in the past 6 months.

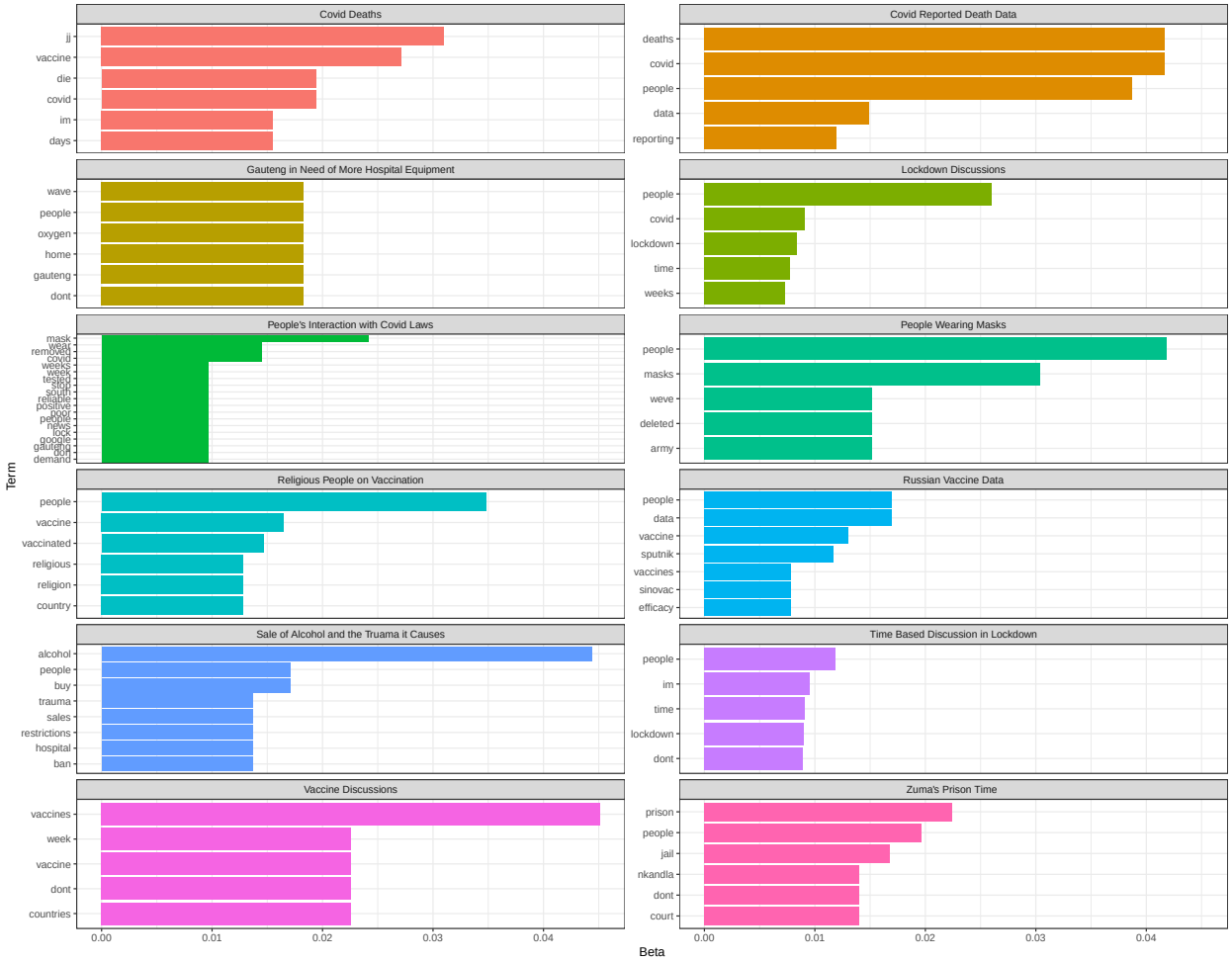



Figure 9.1: Modeled Topics (All Tweets)