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

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

Contents

GitHub

GitHub Repository Link

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. Media Agency Reasoning

After research on what makes a news source trustworth and unbias, we chose a mix of them.

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

3.1. Bigrams

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 3.3.

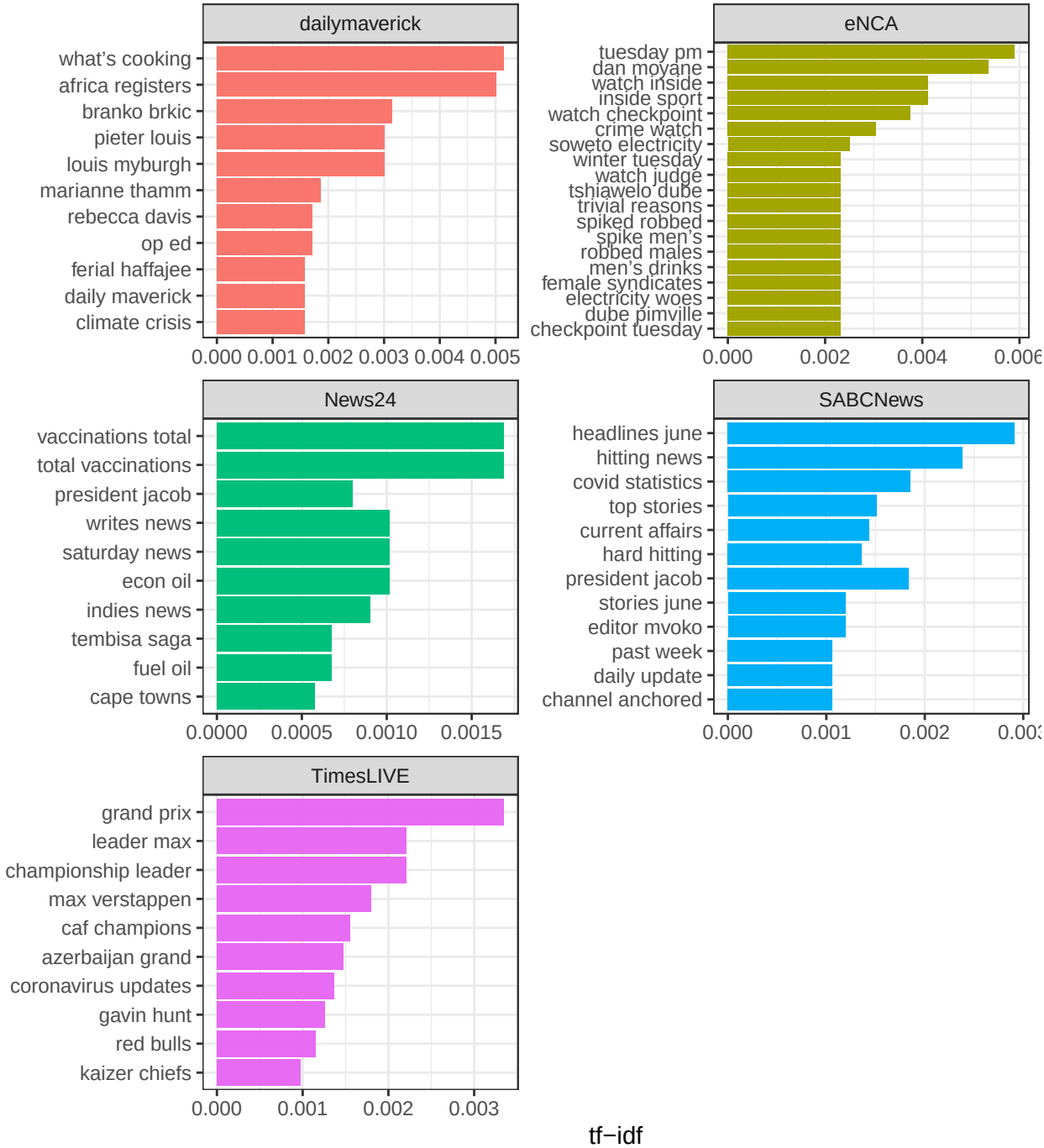


Figure 3.2: Bigrams

3.2. Negation Words

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.

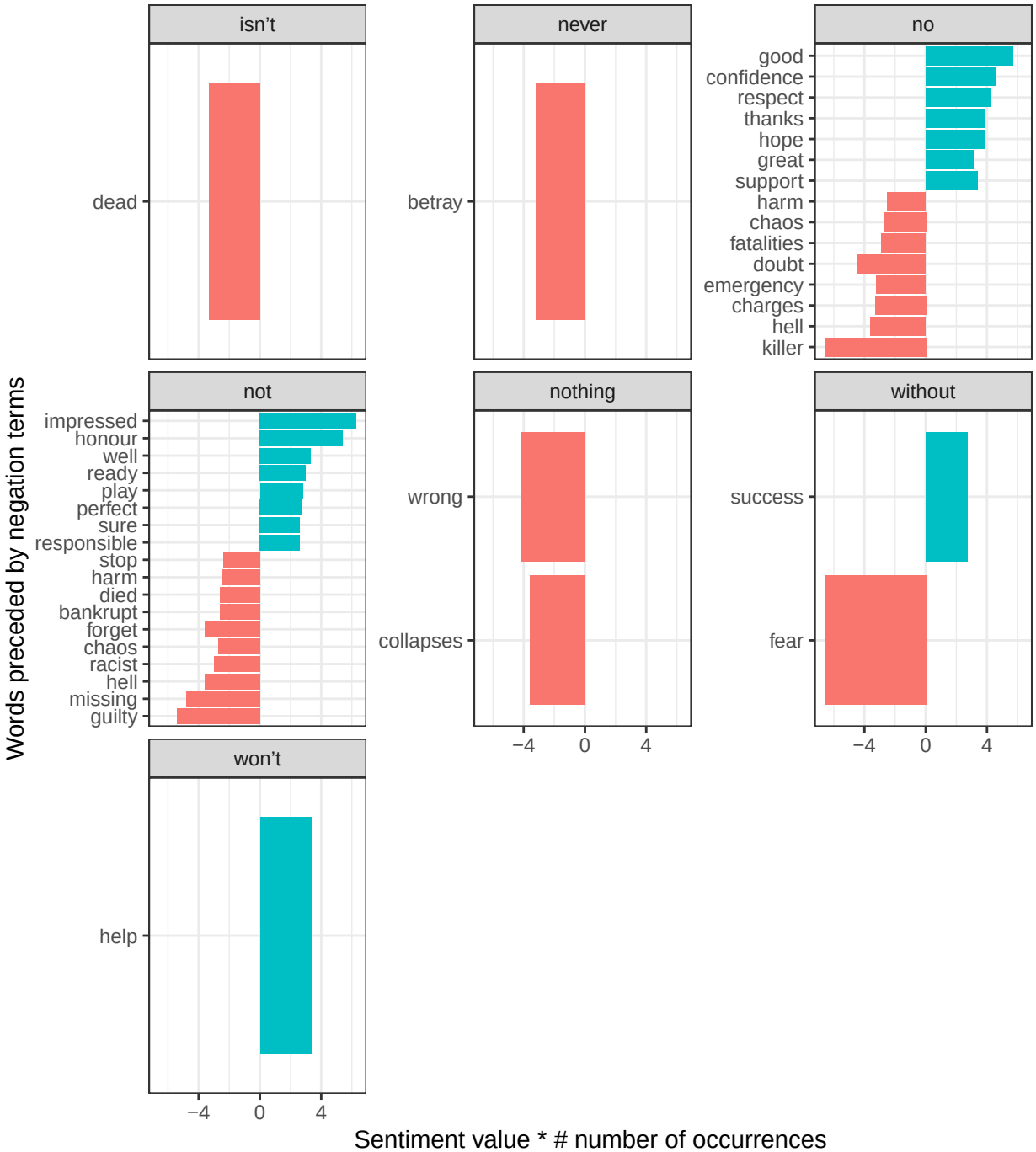


Figure 3.3: Negation Words

3.3. Top Words

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.

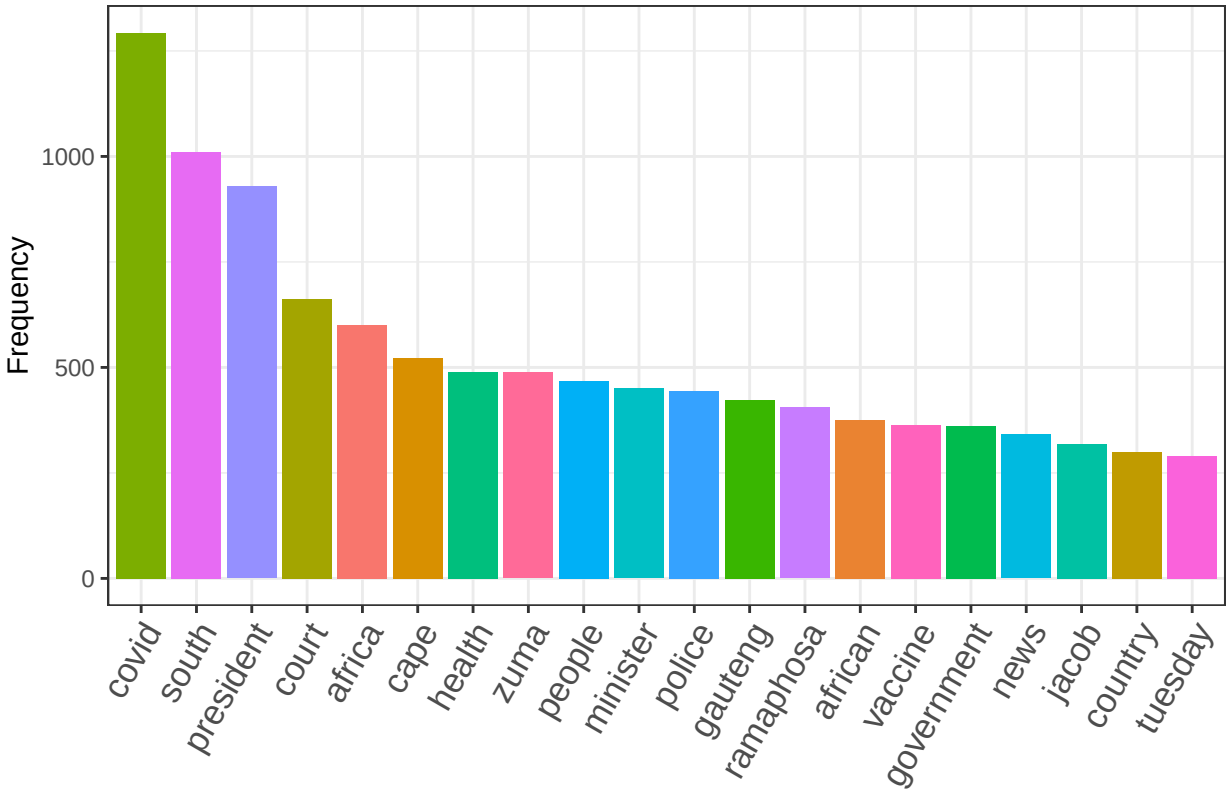


Figure 3.4: Most Frequent Media Agency Words

3.4. Analyzing Words and User Frequency

3.4.1. Term Frequency Distribution

What we see from figure 3.6 is each media agency's use of uncommon words. The more right-skewed the data is, the less unique the news articles are. Times Live is shown to have the most diversity.

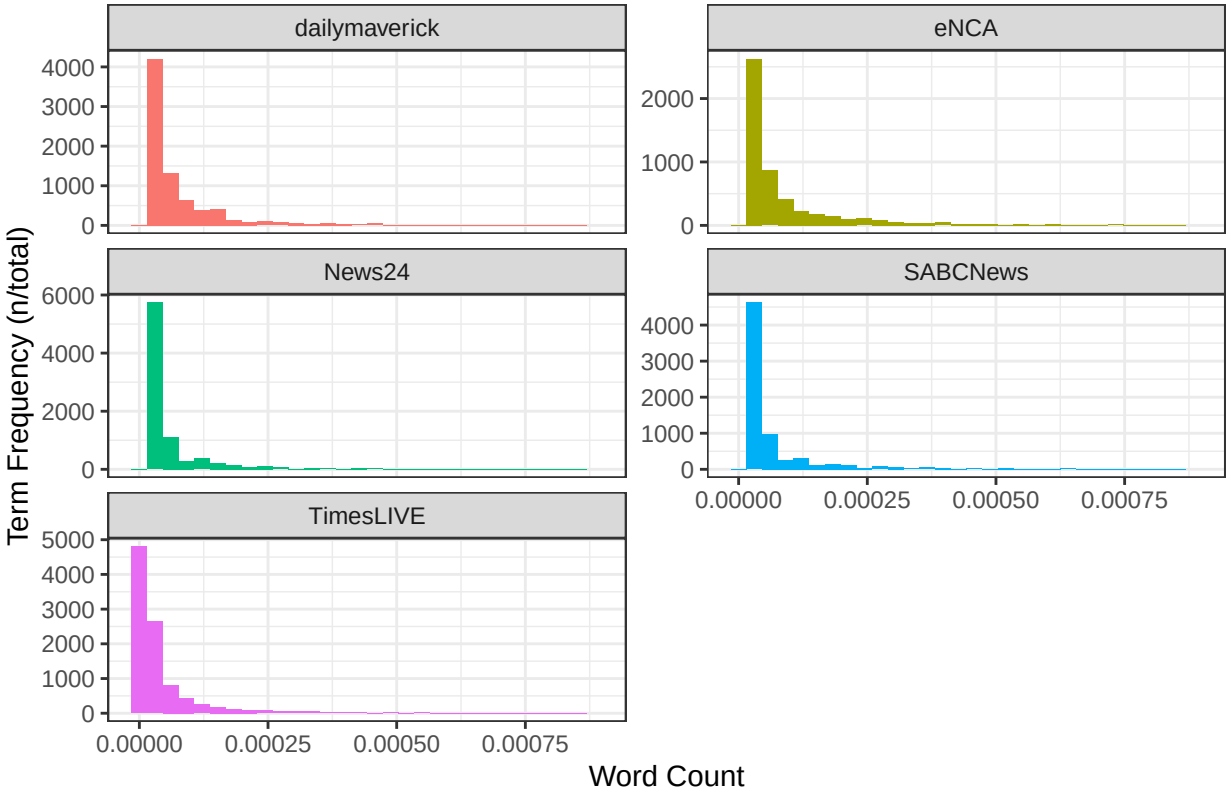


Figure 3.5: Distribution of Term Frequency

3.4.2. Zipf's Law

Our interpretation of Zipf's Law (Figure 3.6) on our dataset shows that the media agencies tend to use the same words often. News on different categories in the world will generally be of the same topic, but with different subjects. Eg. "Business News today for Amazon is x" vs. "Business News today for Microsoft is x".

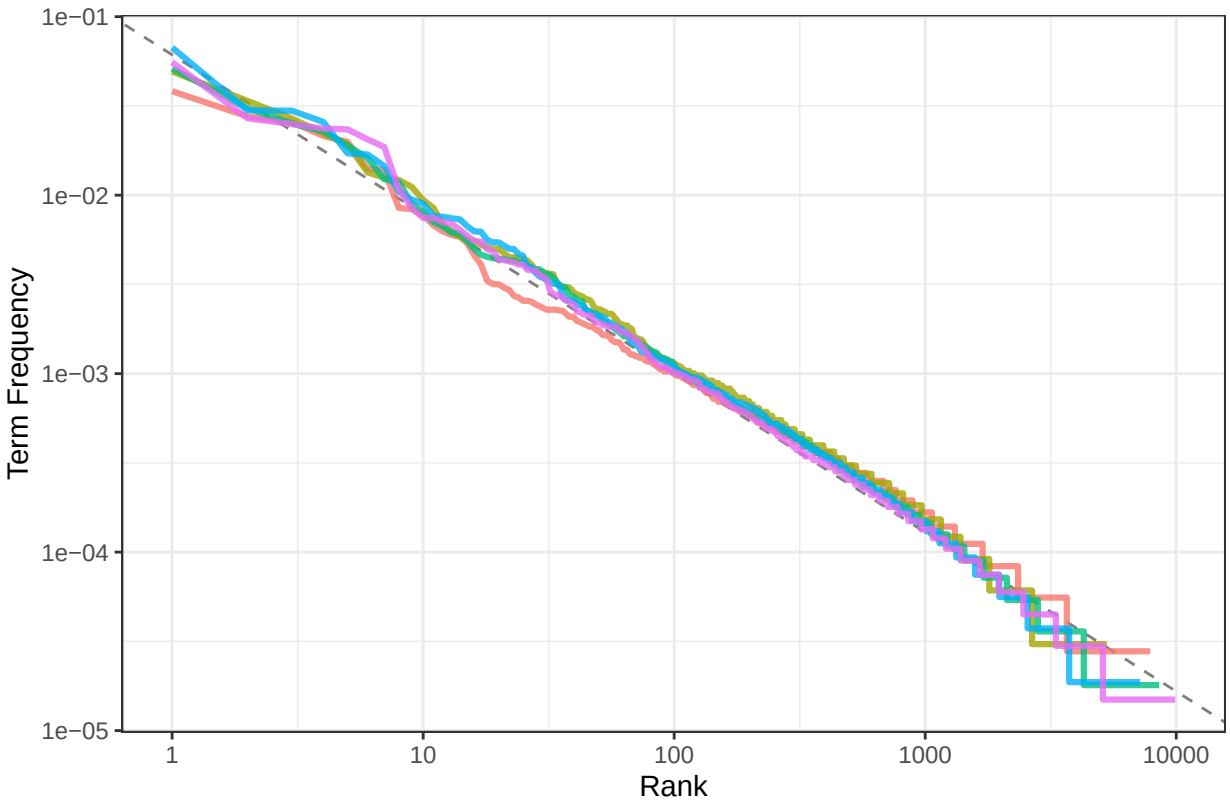


Figure 3.6: Zipf's Law with a Broken Power Law

3.5. TF-IDF

A tf-idf is then modeled to determine which words are the most important per media agency (Figure 3.7). 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.

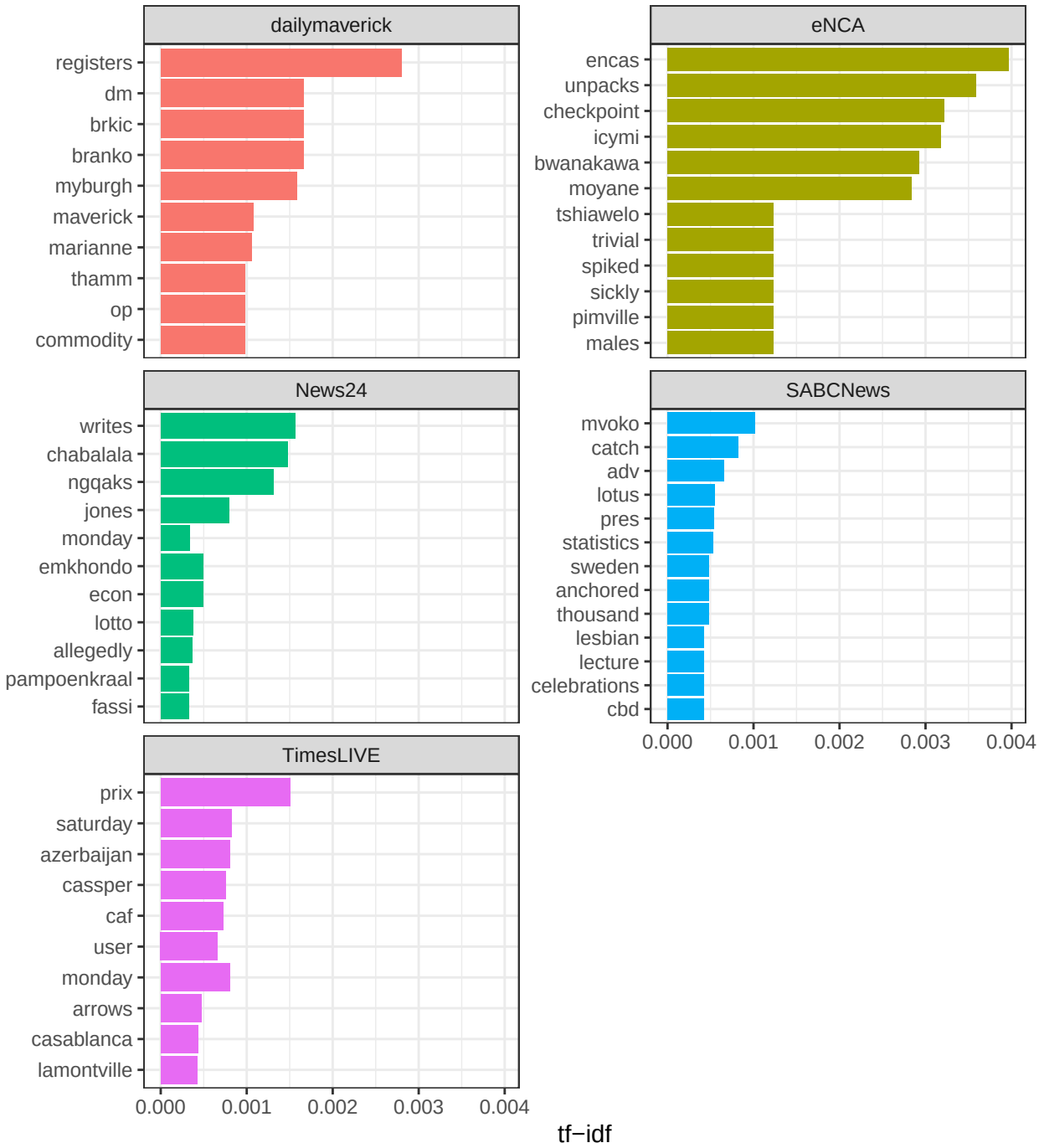


Figure 3.7: Highest tf−idf from each Media Agency

Tweet topics involved with the tf-idf's listed in the weeks collected (Figure 3.8):

- 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.

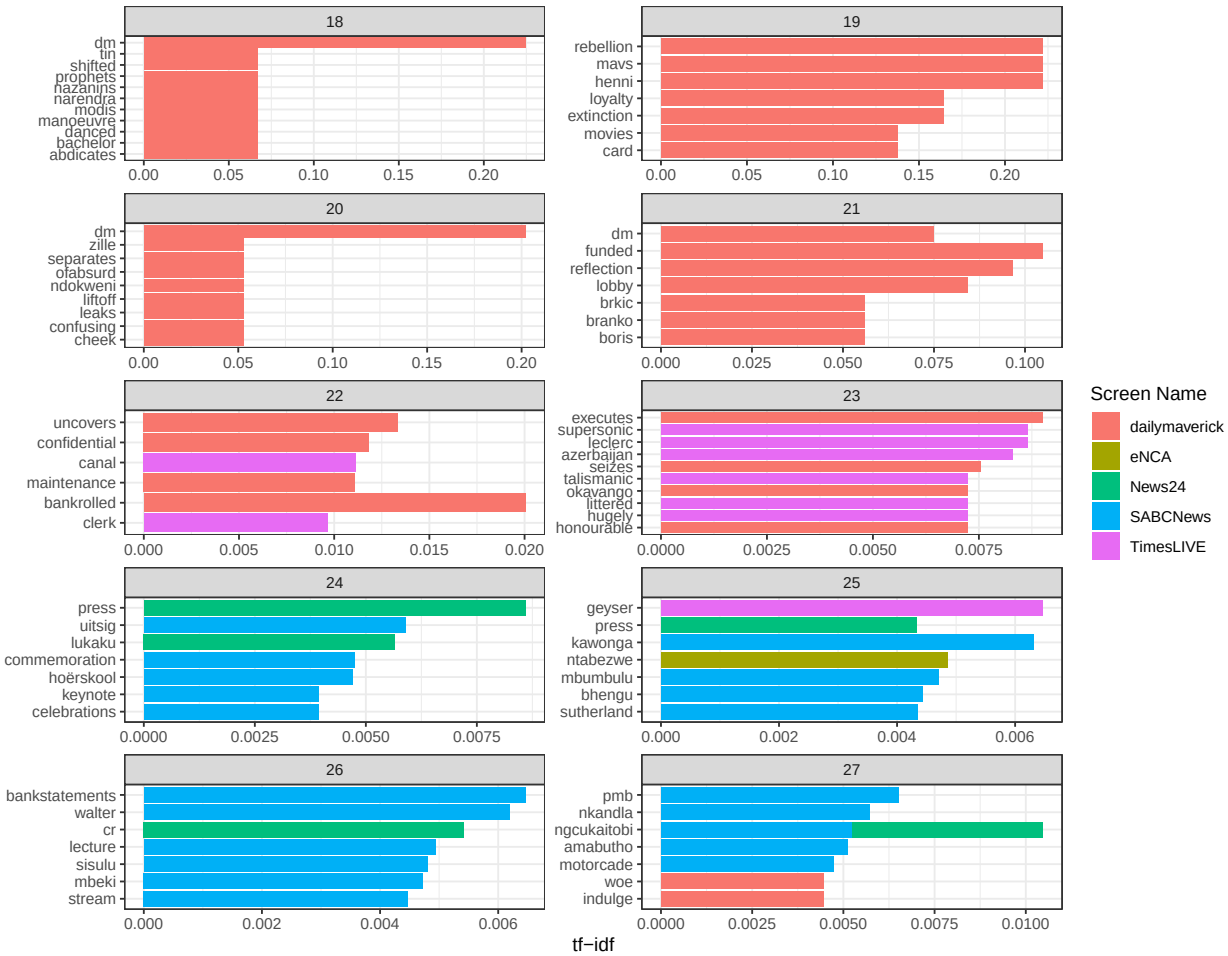


Figure 3.8: Highest tf−idf by Media Agency in Each Week

3.6. Word Correlation Among Users

The figure (3.9) 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.

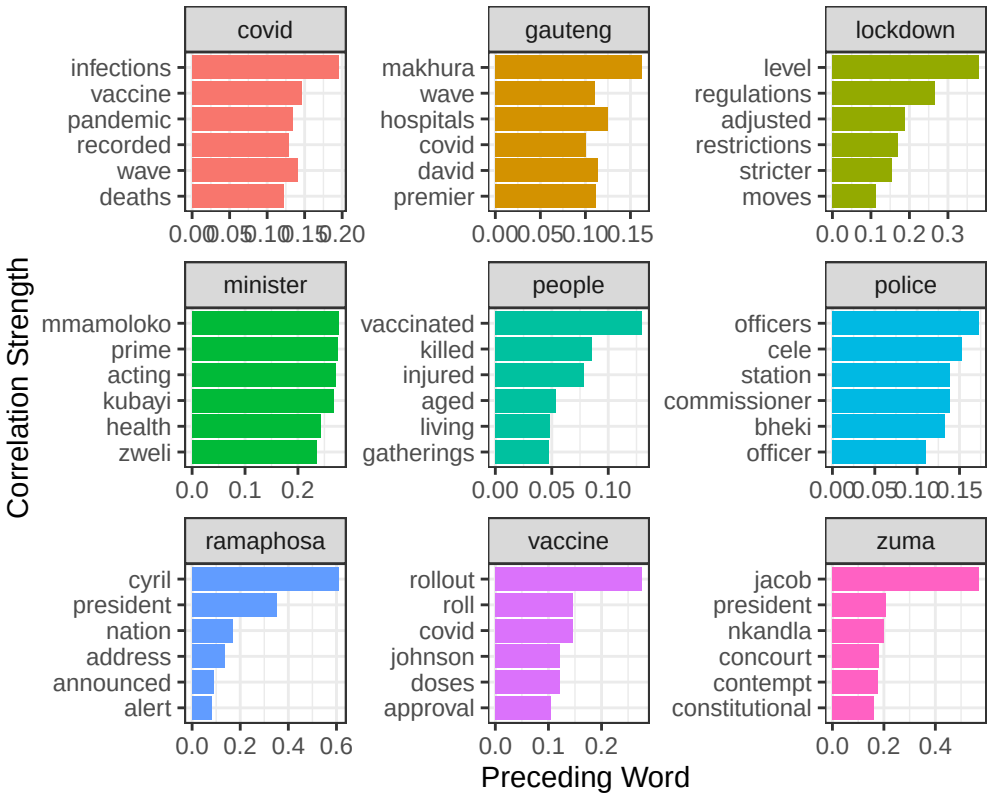


Figure 3.9: Words that Precede Popular Terms

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.

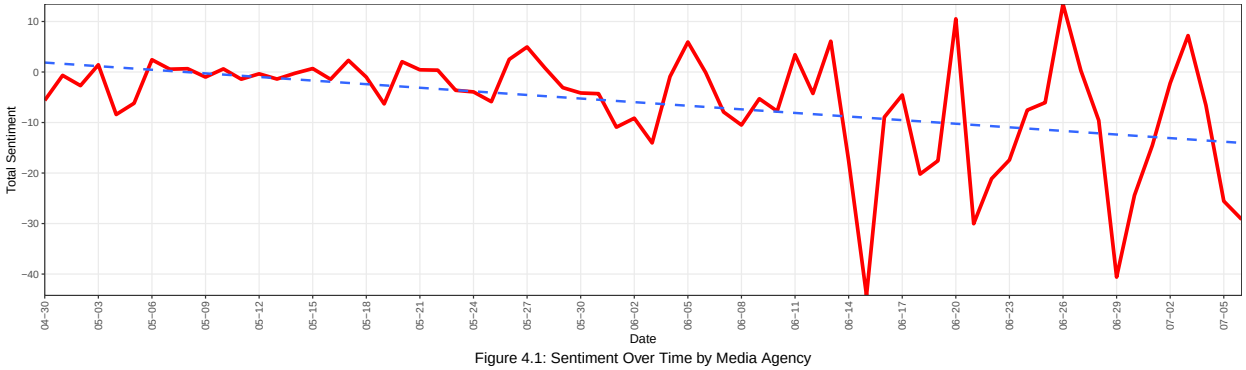


Figure 4.1: Sentiment Over Time by Media Agency

4.1.2. Most Negative Tweet

Times Live: 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 negative daily news.

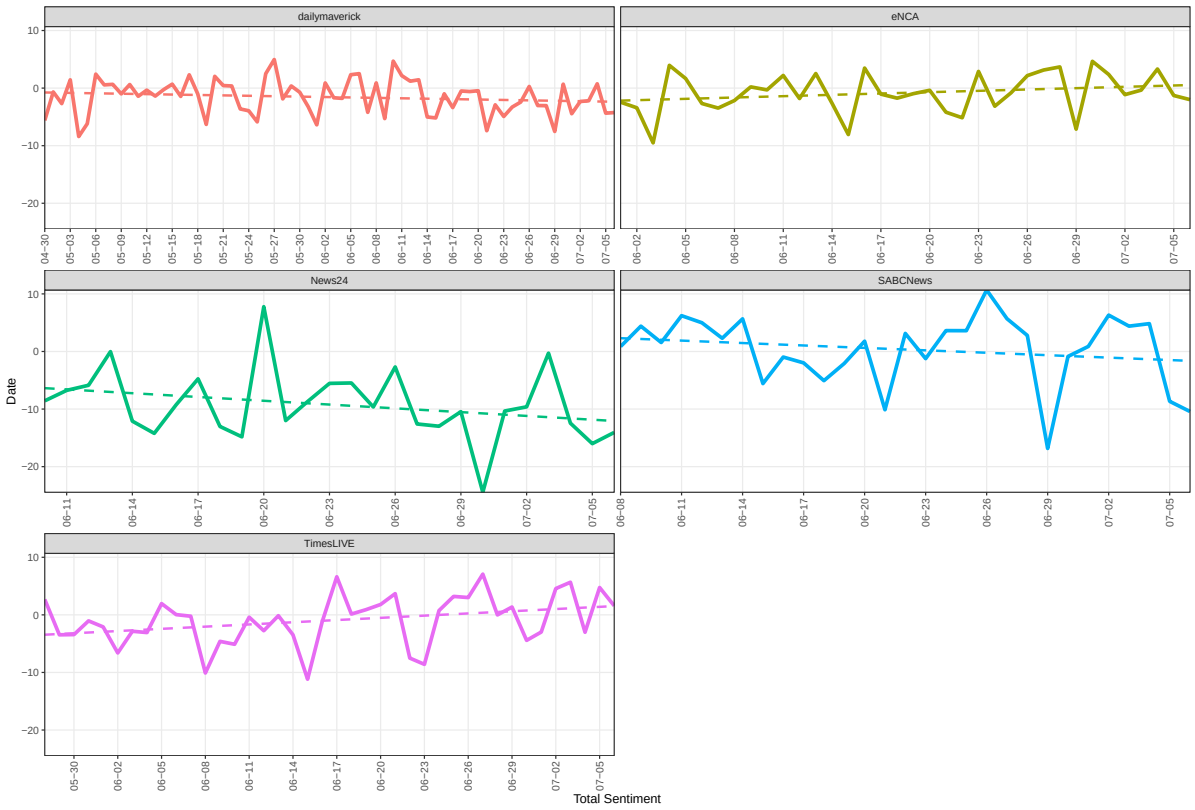
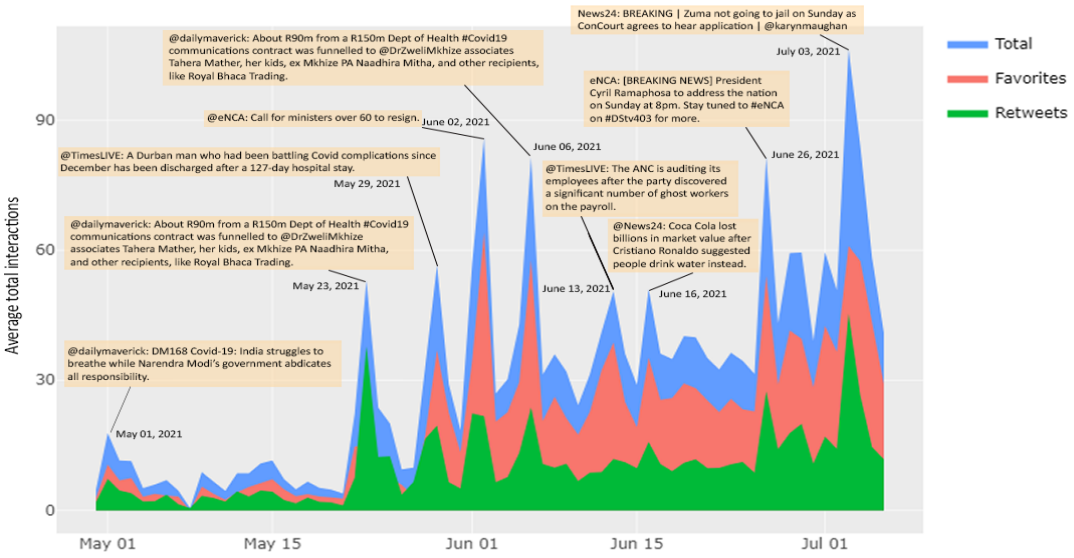


Figure 4.2: Sentiment Over Time by Media Agency

5. Interactions

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



5.1. 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 (Figure 6.1).

6.1. Gap k justification

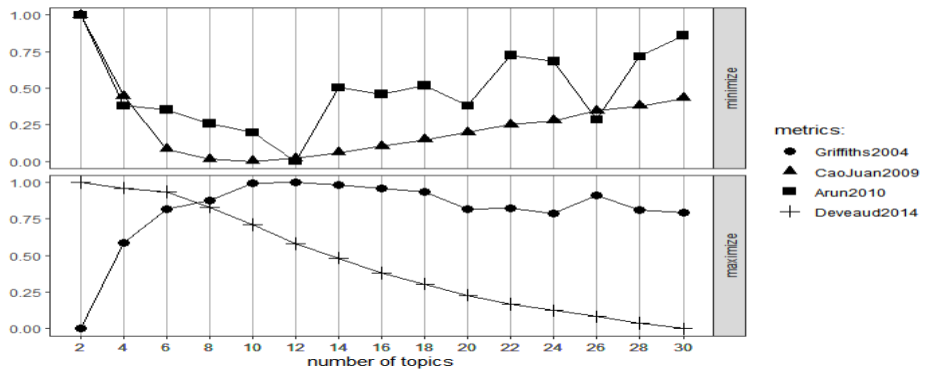


Figure 6.1: Justify K Value

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.

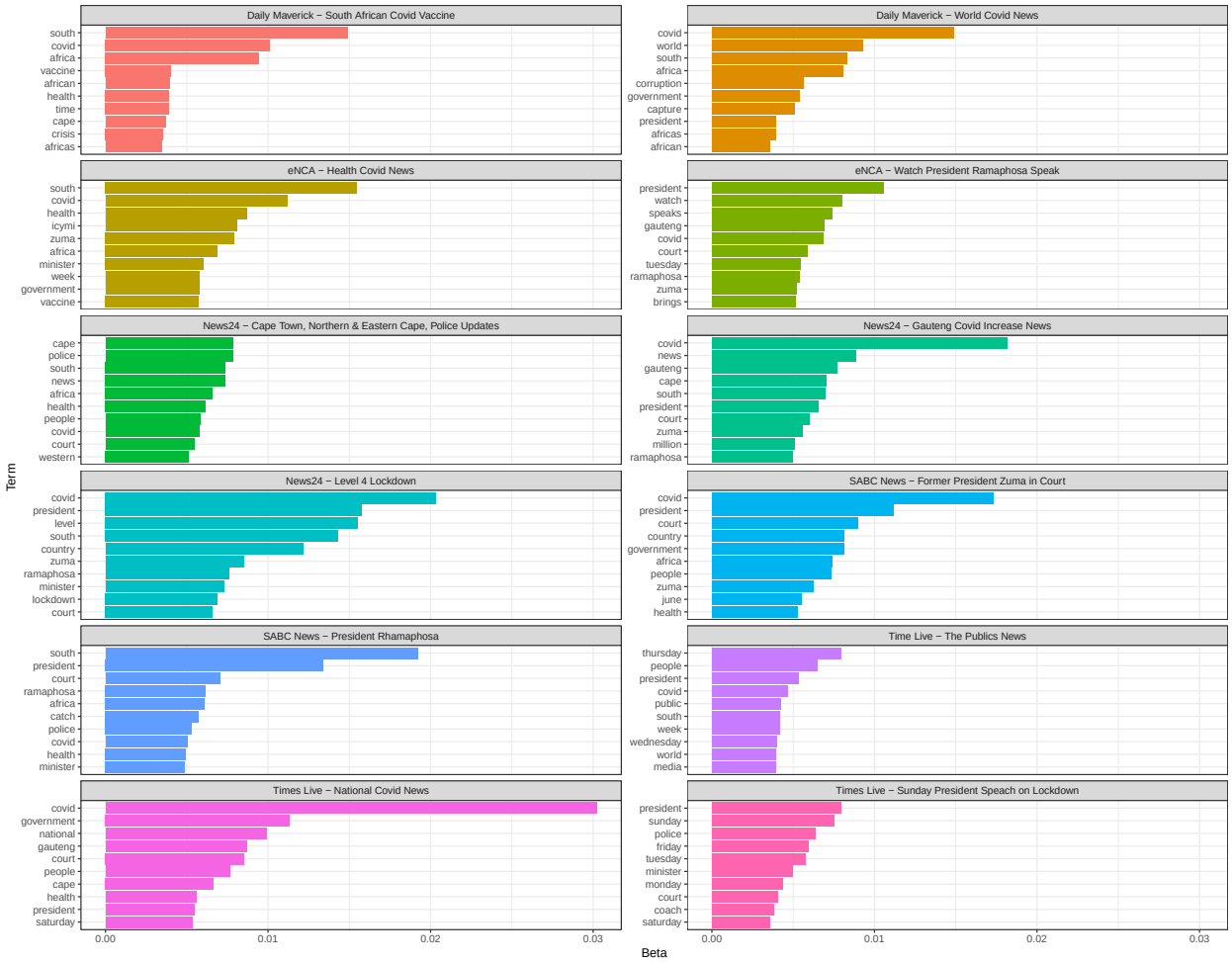
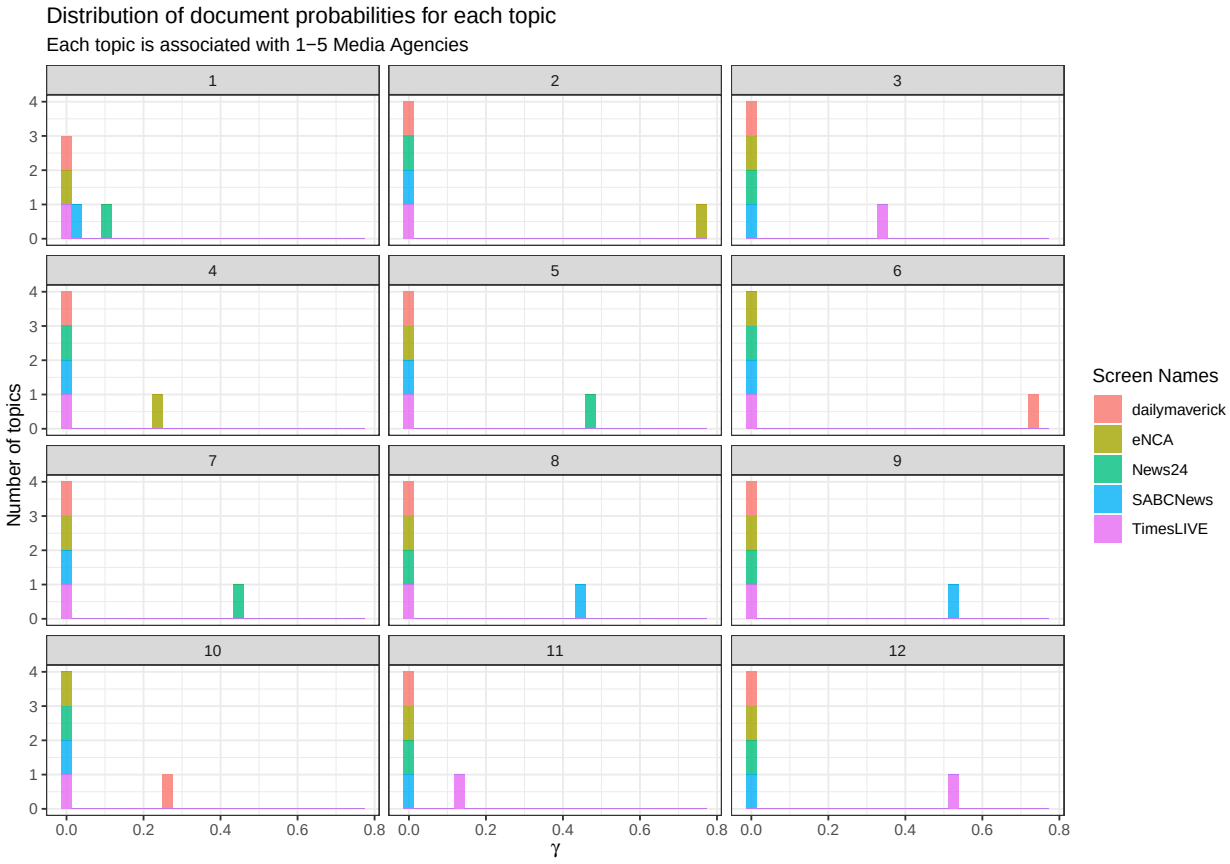


Figure 6.2: Modeled Topics (All Tweets)

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.



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 follow most of the same analysis for Reddit data as Twitter.

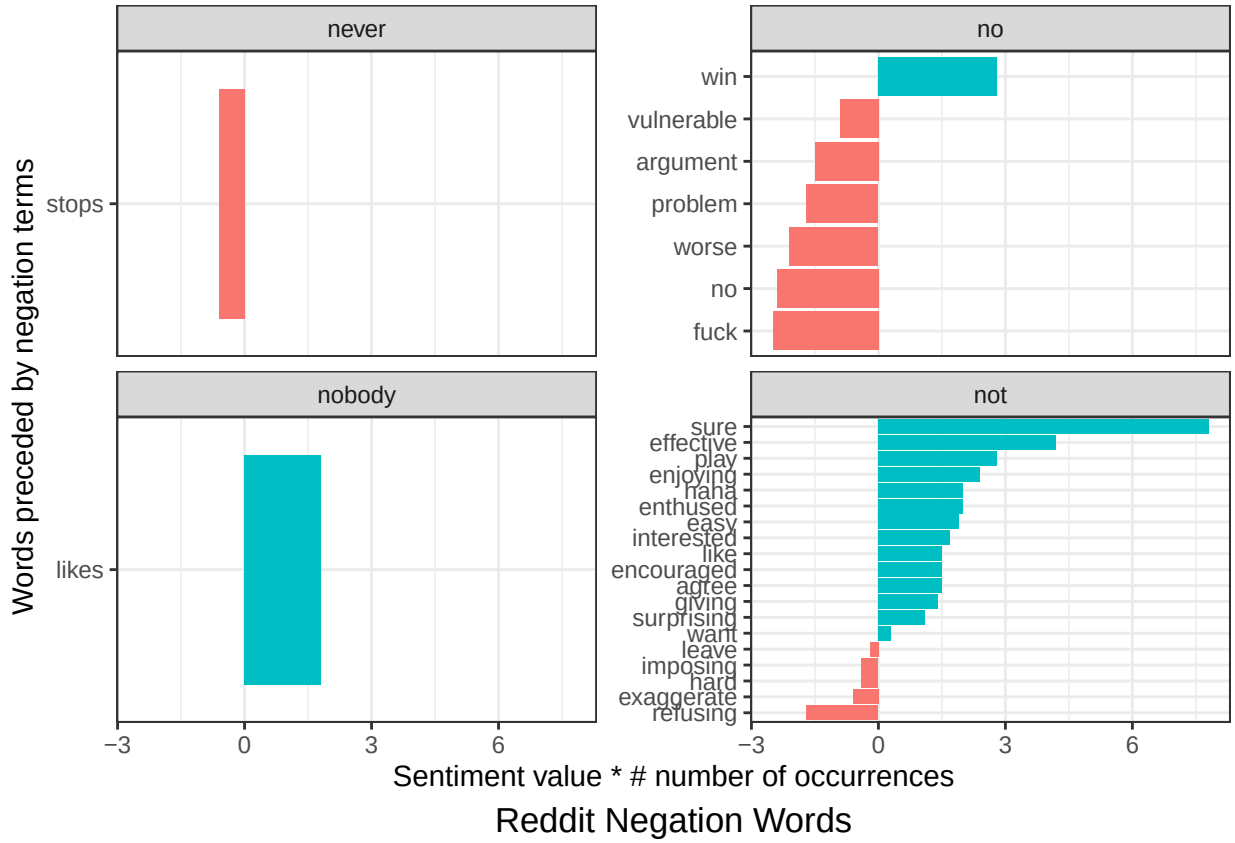
7.1. Reddit Comparing Comments From r/southafrica Covid-19 Posts.

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

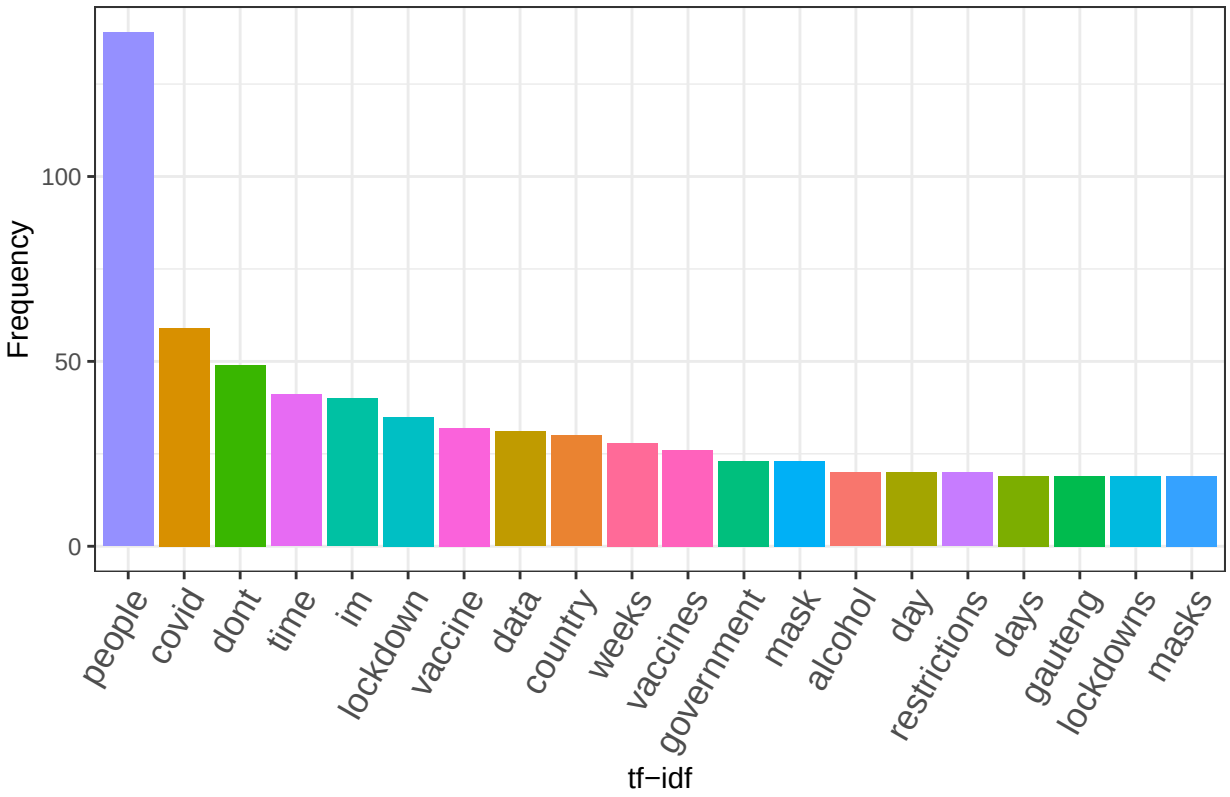


Figure 7.2: Most Frequent Reddit Comment Words

Our interpretation of Zipf's Law (Figure 7.3) on Reddit comments show that commenters use a lot of text slang, as they do not follow the broken power law of regular language.

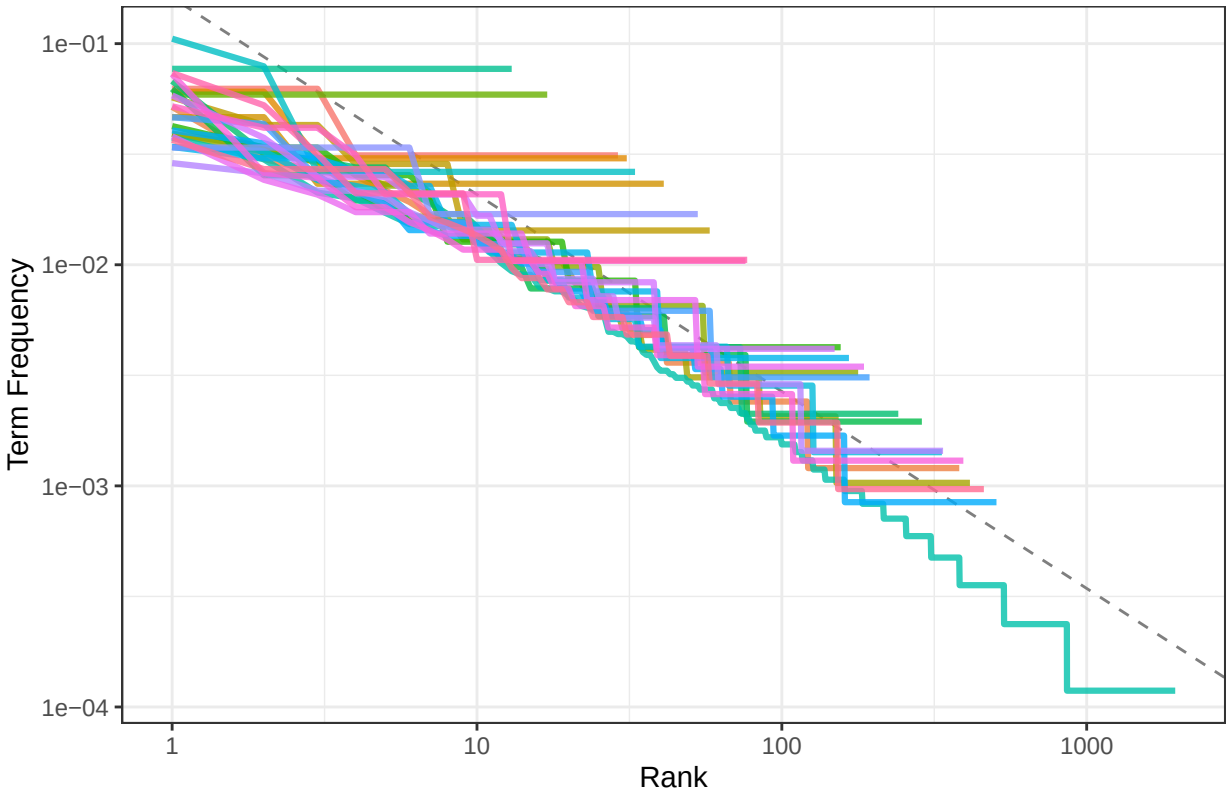


Figure 7.3: Reddit Zipf's Law with a Broken Power Law

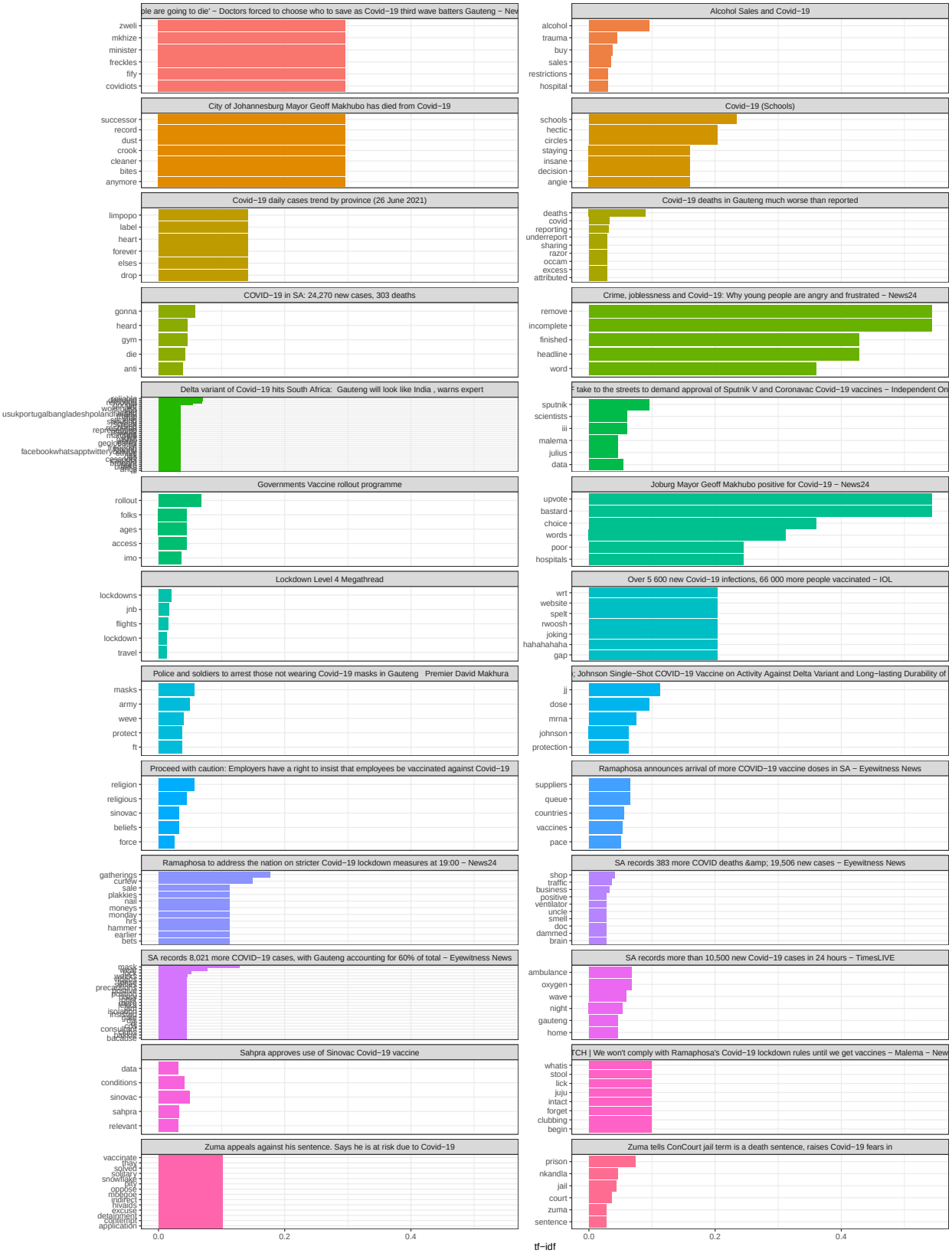


Figure 7.3: Highest tf−idf from Title

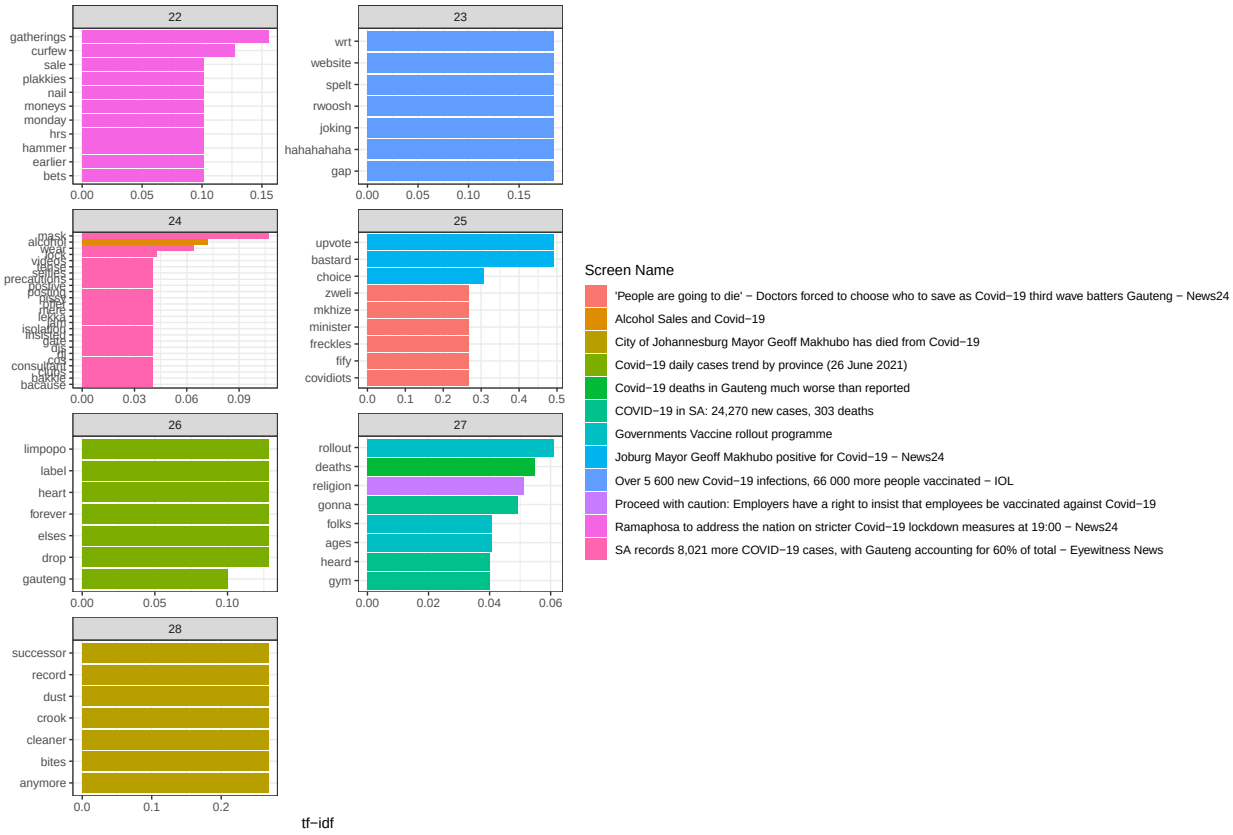


Figure 7.4: Highest tf−idf by Title in Each Week

The graph below shows which words are most likely to co-occur, or not. We use this to understand the context of words in our topic modeling. The words shown are popular words in our dataset.

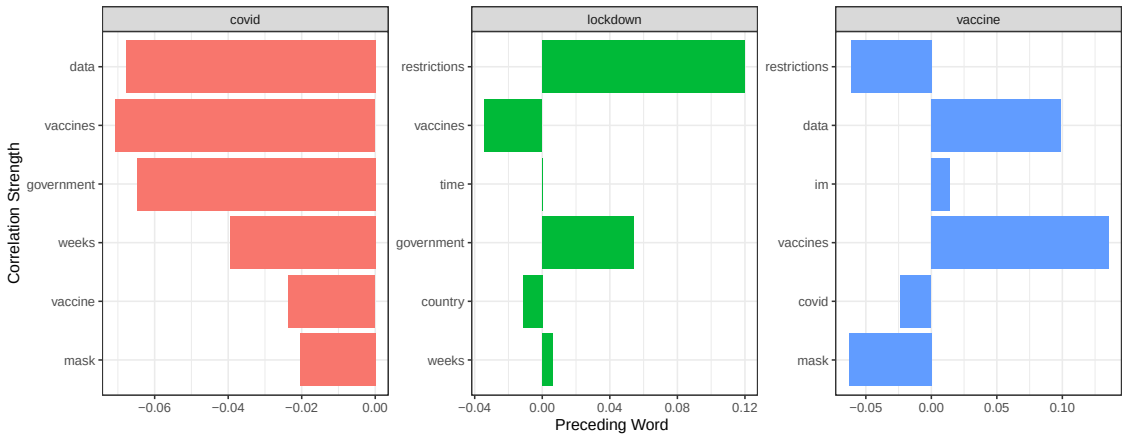


Figure 7.5: Words that Precede Popular Terms

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.

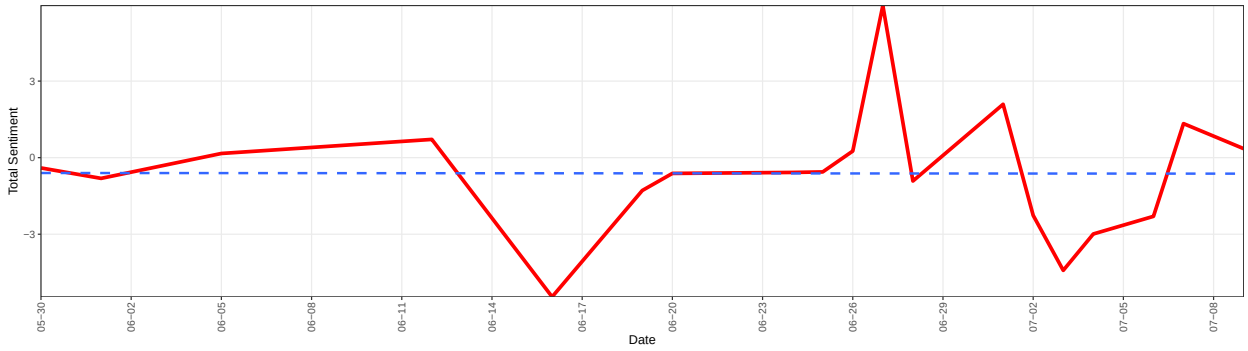


Figure 8.1: Sentiment Over Time by Comments

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

user	comment	comment_score
Alcohol Sales and Covid-19		
brighighs55	The primary aim of restrictions on alcohol sales is to prevent hospital beds being filled up with patients who have suffered injuries as a result of alcohol fueled fights etc. Weekends are peak times for trauma victims at hospitals. I read somewhere that over the last festive season the trauma ward at Baragwanath Hospital was empty over the New Year's weekend - because of lockdown restrictions on alcohol sales.	13
Covid-19 deaths in Gauteng much worse than reported		
gynnhays	And there are still people who say its overblown. Morons.	19
UpSiSunny	They are fucking morons. It is worse now than it has been. +- 26500 new cases reported on worldometer. We are fighting for each other's lives here and people are pissed off because they can't go to the pub.	19
EFF take to the streets to demand approval of Sputnik V and Coronavac Covid-19 vaccines - Independent Online		
IikGkyHisoryRSA	jislaaik, the eff can take the sputnik and coronavac vaccines and keep it for themselves.	15
Governments Vaccine rollout programme		
fixavs	I don't think the issue is the grouping around ages. It's the speed of the rollout. Phase 1 started in February, over 60s started in May, and now in July over 50s. Given this timeline, I'm only expecting to be able to get vaccinated next year. And this can be a problem if even newer variants show up. The rollout is too slow IMO. They're just putting vaccination efficacy at risk.	11
Lockdown Level 4 Megathread		
snai-iis	I love how he targeted whattsap groups like I know yall are spreading false information	53
Police and soldiers to arrest those not wearing Covid-19 masks in Gauteng Premier David Makhura		
AnomayNxus	&gtarrest And then what? Throw them in a tiny holding cell? ngl I had envisioned this army thing more as setting up field hospitals, maybe help transport oxygen...that sort of thing. Hopefully that's happening too & just not being mentioned	27
Proceed with caution: Employers have a right to insist that employees be vaccinated against Covid-19		
gynnhays	Good article. This one is going to cause lots of problems, but I do support it as whole. The employee is going to have to show that they have conducted a thorough risk analysis and put whatever measures they can in place, including measures to accommodate, where possible. I am wholly against religious exemptions for this, given that I cant think of any religious texts which expressly ban vaccination. Any one have any examples of such.	11
SA records 383 more COVID deaths & 19,506 new cases - Eyewitness News		
rashcanman2000	It's business as usual. Traffic is a little less congestive because Schools being closed but Businesses are still open. Our company has had 3 positive cases since last week Thursday with 14 additional Employees that are on sick leave but have not been tested. This is out of 62 people in our company. My Uncle has been on a ventilator and induced coma for two weeks and intensive care. They had to go through his tragea as the ventilator mask didn't ensure enough oxygen. They started him on dialysis on Sunday for his kidneys. Another Uncle of mine fell yesterday and we had to go to three Hospitals before one could assist us with a brain scan. Luckily the Neurologist said that his brain is fine but would have preferred to keep him there for his other injuries but they simply don't have any beds open so he is being treated at home. With all this I had to confront an old Lady at Spar an hour ago as she took her mask off to smell the cheese in the dairy isle. Firstly it's vacuum sealed so how can you smell it and secondly now is not the time for people to be negligent about following health protocol. Things aren't going well at this stage.	11
SA records more than 10,500 new Covid-19 cases in 24 hours - TimesLIVE		
Anon_Pannkok	This wave is much worse than the first two in Gauteng. Friends of mine that are doctors are literally crying as they cannot cope. Beds have run out, they are sending new patients to Bloemfontein, but they're dying on the way. They can't believe that the government is not imposing more restrictions with how severe it is.	14

9.2. Controversial Comments

Usually met with less comment score.

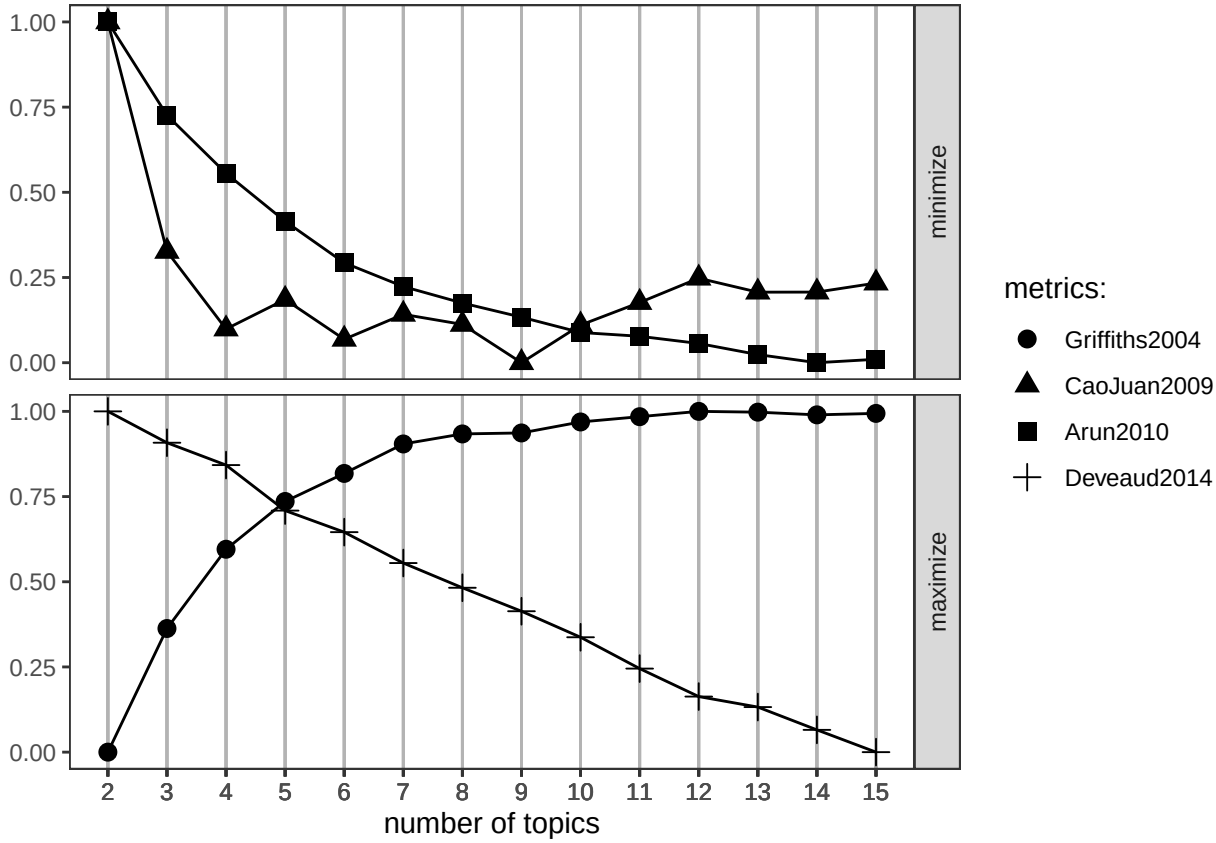
user	comment	comment_score	controversiality
Ramaphosa to address the nation on stricter Covid-19 lockdown measures at 19:00 - News24			
Pretty_Bison7682	i bet they ban booze sales for two months and limit outdoor gatherings but do nothing toward curfew when the only tool you understand is a hammer every problem looks like a nail	3	1
Over 5 600 new Covid-19 infections, 66 000 more people vaccinated - IOL			
ReganErasmus	you spelt lol wrong	2	1
Alcohol Sales and Covid-19			
charlakamavric	you see unfortunately no one really knows what the motives is of the sa government just like the smoking ban from last year only forced millions of cigarettes to be sold and used illegally	-1	1
EFF take to the streets to demand approval of Sputnik V and Coronavac Covid-19 vaccines - Independent Online			
Tokoloshe789	sputnik is apparently very good personally though i would like to load julius in a sputnik with his mate floyd and fire them off to the fuckin moon	0	1
Lockdown Level 4 Megathread			
teonicolaides	schools should stay open	2	1
Not-the-best-name	you goto understand that sa is a bit different now with covid even more so art is a luxury just like theoretical research that we cannot afford there is not many jobs in the arts going around right now also not in nuclear physics	-3	1
RoastyMcGiblets	tables can be socially distanced in a restaurant and people can safely eat especially a family that has been living together anyway restaurants serve society by employing many people and they cant keep everyone on payroll doing only takeout many businesses closed completely last time this is terrible for the economy and for a lot of the service workers funerals are full of people hugging i honestly dont get it	-1	1
slimm711	the problem is not wit the restaurants the problem is in the taverns they are hosting partys no social distancing mask what the f is that its sad after a yr n half ppl are taking covid lightly its sad when you question one to put his mask on his response is he dont ve covid but i like i always do educate the individual	0	1
cybercupcakes	is this honestly choosing the safety of our citizens or is this just short term thinking thinking long term the economic destruction of these lockdowns will lead to many families without work food medical and education and all these factors combined will cause suffering for the majority and eventually we will be just like zimbabwe dont get me wrong i will be for lockdown if we can afford it if we were in a first world country we could afford a lockdown but sadly we just cant afford it imo and i think the lockdowns will cause more long term damage than the short term damage of the virus	3	1
cybercupcakes	im sorry for your loss i am well aware people are dying no one is saying that is not happening all i am pointing out is that a failing economy can lead to a lot more sufferingdeath than the mortality rate of covid what is most important is that the government acts with logic science and data not with emotions running high because a loved one passed away nor with the selfish urge of wanting to go out to gatherings	4	1
shitcanfly	i get we need be cautious but my business is around weddings and commercial projects why should kzn suffer cause of jobrug rd fuckin time	4	1
humanfly___	klink uiters kak bly veilig mense lt van londen se wereld	-1	1
cybercupcakes	here i wanted to go watch black widow when it released	0	1
The_Angry_Economist	current data	-2	1
The_Angry_Economist	as in these people have been constantly behind the ball and why would i expect them now to get it right	1	1

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.

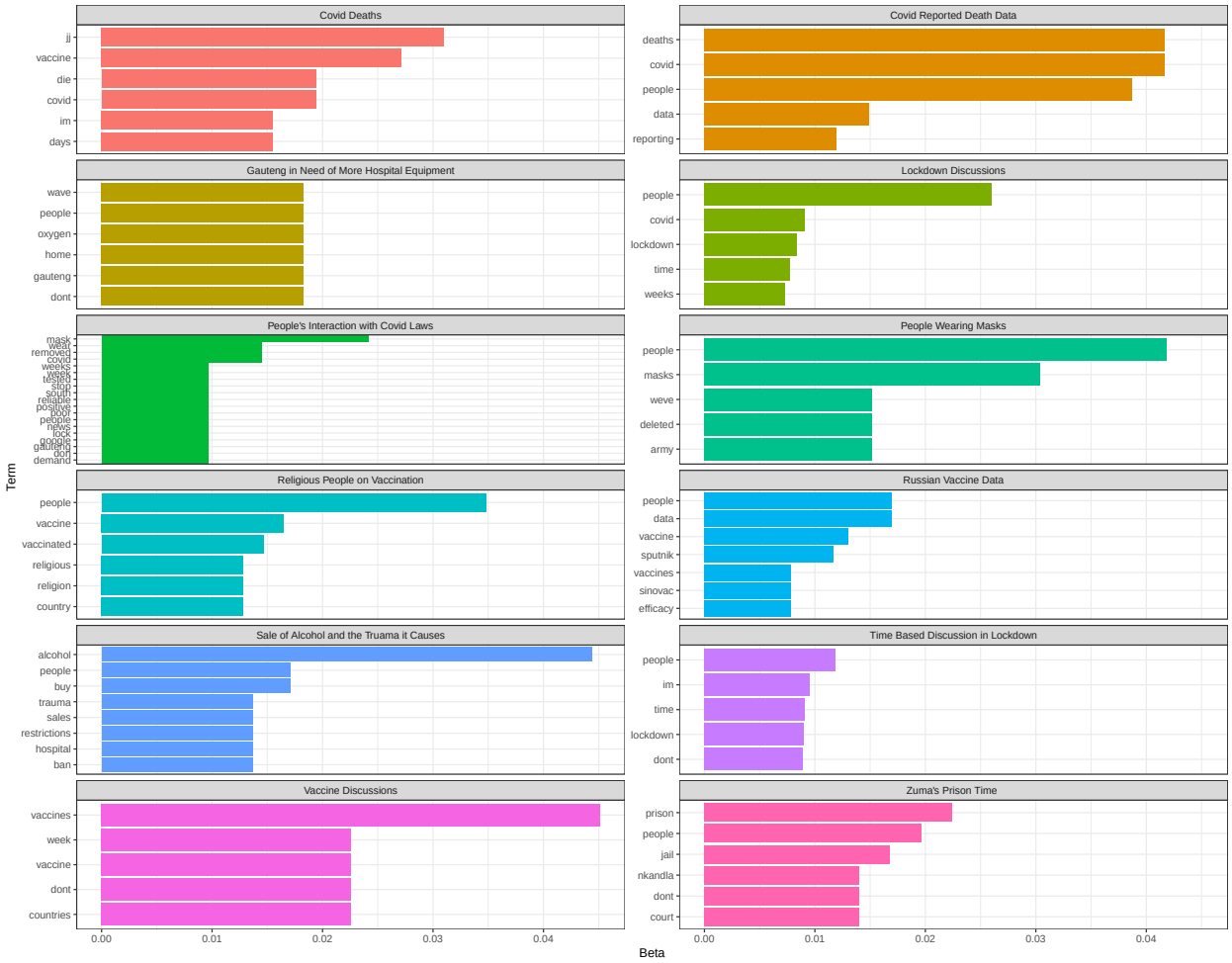


Figure 9.1: Modeled Topics (All Tweets)