Opleiding Informatica

The discovery of underlying topics within Coronavirus-related articles

Suzan Al-nassar

Supervisors:
Peter van der Putten (LIACS)
Jasper Schelling (ACED & Rotterdam University of Applied Sciences)

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

22/01/2022

Abstract

This thesis aims to discover underlying topics or themes within coronavirus-related articles. We have chosen five different Dutch media with different positioning towards the coronavirus debate: De Dagelijkse Standaard, Telegraaf, NRC, NU.nl and Volkskrant. We collected around 2000 corona-related articles from each newspaper between the first of January, 2020 until the end of April, 2021. With the help of text mining we used different analysis methods, mainly LDA and bi-gram networks, to discover the underlying themes or topics. We obtained results that can tell us more about the underlying topics, though results are qualitative and at times difficult to interpret. In future, a larger and more diverse data set could lead to more more differentiating and more clearer results. Therefore we can say that artificial intelligence can successfully contribute to discovering underlying topics or themes within an article.

Contents

1 Introduction

Nowadays, it is not hard to read, hear, or watch the latest news, especially since everyone has a smartphone now which makes news easily accessible. People have currently been informing themselves about the COVID-19 pandemic through all sorts of news media. Important to remark is that not every news medium shares information with only the intention to inform the reader, but some might also want to influence the reader to adopt alternative beliefs about the pandemic.

With our research, we aim to create awareness that media channels will have different positioning towards high profile societal topics such as the corona debate, and this will be reflected in their reporting. We aim to achieve this with the help of artificial intelligence [15]. In this thesis, the discovery of underlying topics or themes within corona-related articles is researched.

1.1 Research

Informing people about issues related to world health has always been a bit tricky, since there is more than one perspective to look at the certain case, e.g. the Mexican flu in 2009. [21] With the help of text mining and a data set, which exists of around ten thousand corona related articles from several different news media, we are able to extract underlying topics or themes that are hidden throughout the articles.

Furthermore, with the help of this topic extraction, we will analyse the differences between the different news papers. Therefore we decided to examine five Dutch newspapers that each differ in their styles of presenting news, ideological positioning and the audience they try to reach. In section 3.2 those newspapers are listed. Every newspaper has some sort of writing style and topics of interest, which define the type of newspaper. Within the corona pandemic, we are interested in how the different newspapers write their articles about corona and if they might include an underlying theme or topic within the article. This leads us to our research questions:

- With the help of artificial intelligence, is it possible to extract an underlying themes or topics of different Dutch newspapers with regards to the coronavirus articles?
- If there are underlying topics, what are these and how do these differ across media?

1.2 Topic extraction

Within this thesis, the term *topic extraction* will be used several times, therefore it is important to know what that means. When searching for the exact definition, the following definition is in our opinion the most complete as it describes the used methods in this thesis accurately:

Topic analysis (also called topic detection, topic modeling, or topic extraction) *is a machine learning technique that organizes and understands large collections of text data, by assigning "tags" or categories according to each individual text's topic or theme.* [10]

This means that when we have a big chunk of text, several topic extraction tools can be used to categorize this chunk of text. For example, when we take the books of the Harry Potter series, we are able to discover which spell belongs to which character with the help of a topic extraction tool.

The coronavirus has been discussed for over 2 years now, and on every medium possible. Because of this, we are able to collect a great amount of text articles which will make the topic extraction work better since it can be applied to a large database. Some tools of topic extraction that will be used in this thesis are: LDA, Dynamic Topic Modelling, and bi-grams combined with word association graphs. All of these tools extract topics from our database but there is a difference between each method. In section 3 the details and experiments with these methods can be found.

2 Background

In this section, we will discuss the background that has led to creating this thesis.

2.1 Evolution of the newspapers

Today, news is easily accessible, but this has not always been the case. Up until the rise of radio, television, and internet, newspapers were the main source of incoming news. The first published newspaper in the Netherlands is called *Courante uyt Italien, Duytslandt, etc.* and it was first published in 1618. [1] The newspaper has been around for more than four centuries now. Over the past centuries a lot has happened that changed and transformed the newspapers.

From the year 1880 until 1960, the Netherlands was divided into four social groups that were based on religion. We refer to this time as the pillarization ('verzuiling') of the Netherlands. During the pillarization, each social group had their own educational institutes, their own shops, and likewise their own newspaper. It was obvious that each newspaper had a specific type of audience and focus, for example, a roman Catholic was forbidden to read the news paper published by the protestant Christians because it would include a lot of elements of the other religion.

It was not until the 1960s that the pillarization came to an end. There is a research done about the pillarization in Flanders which aims to discover if there has been an effect of the pillarization on the newspapers of nowadays. The newspapers of nowadays discuss ideology less than expected as a consequence of the ending of pillarization. Another effect shown in the aforementioned research is that nowadays, the different newspapers - despite of their historical differences - support the government's decisions in general. Back in the era of pillarization, this was not accepted because as a member of the Church, one could not agree with the government without the support of the Church. [22]

Although the pillarization has ended a long time ago, the news media nowadays still carry the same concept of writing their articles for a certain public with a certain focus and hold different political beliefs and ideology. However, the news reporting of nowadays *should* be objective according to the code of ethics of journalists. [11] When one reads a news article, the person should come up with an opinion themselves instead of copying an opinion from the newspaper. With this in mind, we want to research whether different newspapers still have an underlying topic or theme.

2.2 Text mining and Corona

From the first of January, 2020 (around the beginning of corona) until the end of April 2021 (the end of collecting data for this thesis), there has been a lot of news about this virus. In the very beginning, the news seemed far away from a Dutch perspective, since the virus started in China,

Wuhan but in March 2020, the coronavirus started to spread in the Netherlands and every news medium was talking about this hot topic. In the time we live in currently, fake news is widespread. [6] There have been quite a few studies on misinformation and fake news with regards to the COVID-19 virus. The results showed that there is proof that people believe in the fake news about the coronavirus. [17] This is a done with the help of text mining. Text mining is useful in researching answers for this project because we have to sort our text from our articles into topics and text mining automates this. Regarding the pandemic, some research has been conducted with the help of text mining.

2.2.1 Mining corona related discussions in social media

With such an enormous and worldwide pandemic, there is a lot of information to be found about this topic. Because of this, many research is done about (mis)information analysis within corona related discussions, in social media.

The first analysis we studied, is of Englmeier [7] who uses text mining with the purpose of detecting fake news about the coronavirus. Their research is done with the help of the *Contexter system*. This is a *"prototypical system that operates on Named Entity Recognition and uses theme-specific Bag of Words to identify semantic markers in text that point to the specific meaning of the text fragment."* The contexter system is still in progress but it has great potential to eventually detect fake news and misinformation within corona related articles. For our research, this paper has insightful information about what is possible with text mining.

The next paper by Kaila et al. applies topic modelling on Tweets that include the hashtag *#coronavirus* [12]. They found out that the most relevant sentiments had negative aspects like 'fear' as well as positive aspects like 'trust'. Topic modelling is a great way to extract several topics from a large corpus of text. This is a form of unsupervised text learning and it can tell us a lot with the right techniques applied.

A next paper written by Pratama et al. [19] also uses text mining to analyse information about corona but based in Indonesia. This analysis is also done on tweets from twitter. Here the methods used are similar to what we want to do with our research. They start with pre-processing the text of the tweets to get clean and workable data, then they will do analysis based on Term Frequency to get the most relevant information out of these tweets. Eventually they concluded that it is possible to get information out of Twitter with the help of text mining. We aim to get such a result ourselves but then within newspapers.

The following paper, by Hossain et al. [9], is also about detecting misinformation within tweets from twitter. However, they do this research using an interesting approach, namely by assigning a misconception to a tweet and then their program must say whether the misconception is true, false or another option is to take no stance on the misconception. They achieved this by using Natural Language Processing models which is a text mining library. With their research, they aim to eventually combat fake news about COVID-19.

2.2.2 Mining corona related discussions in news

In this subsection, we look at papers that research corona-related articles in news. We start with a paper done by Krawczyk et al. [16] who analyzed 26 million online news articles from the front pages of 172 major online news sources in 11 countries. This research aimed at discovering the

overall sentiment and the most relevant themes within those articles. This is achieved with the help of sentiment analysis by a tool called Vader, and the themes are subtracted with the help of topic extraction. They concluded that the main purpose of the articles is to inform the people about the COVID-19 situation and they found that the sentiment was not negative, however within wide heterogeneous reporting of the pandemic, the sentiment of around 16 % of this coverage is highly negative. This research is similar to our analysis and therefore can be used as a good source of inspiration.

A next paper by John Kastner [13] and others is about the development of an application that is *"usable on desktop and mobile devices that allows users to explore the geographic spread in discussion about the virus through analysis of keyword prevalence in geotagged news article"*. This is done with the help of key term extraction with TF-IDF applied. This is an interesting approach on portraying the results from this text analysis since it engages the users by using their application. It is also a good way to sort a large collection of articles and key terms.

2.2.3 Mining COVID-19 medical literature

An important reason for using text mining with regards to the coronavirus is to contribute to medical research.

For a research that is corona-related, there is a collection of COVID-19 related scientific papers and research called CORD-19 [29]. This is an open research data set which keeps growing with new scientific papers and research every week. This data set can be used to gain in depth insights about corona-related analysis since there are many tools used and many graphs are created which show the effects of those tools. The purpose of CORD-19 is to connect the research results from machine learning with the biomedical research in order to gain information about effective treatments and management policies for COVID-19.

The paper by Reddy et al. [24] is a great example of how CORD-19 can be used and focuses on clustering insightful information that could contribute to corona related medical literature with the help of text mining. Reddy's research aimed at filtering medical terms out of the data set CORD-19 with the help of Natural Language Processing and they used spectral clustering to get their results. They found many terms from the clusters that could be helpful for the medical world in their research.

These papers are a good inspirational examples of what is possible with text mining. We aim to use text mining in order to answer our research questions.

3 Experiments

In this section, we will explain our approach at a high level, then we show the detailed experiments and finally we will show our obtained results. We have done several experiments that contribute to answering our research questions. We used the same overall corpus, and where indicated, subsets of this corpus based on time periods, or various newspapers in scope.

3.1 Research objectives and approach

The purpose of this research, is to create awareness for readers of Dutch news media. We want to do this by extracting the underlying topics of news articles with the help of text classification.

Dutch newspaper	Number of collected articles
De Dagelijkse Standaard	1067
Het NRC	1154
NU.nl	1182
De Volkskrant	2504
De Telegraaf	8337
Total: 14244	

Table 1: Collected newspapers

We start with scraping articles about the coronavirus from several Dutch newspapers. Once we obtained the required data, three main experiments have been carried out to analyze topics and themes across various news sources.

Since we are dealing with unsupervised classification, the first experiment is meant to get a more detailed insight of our data. We want to research what the main differences are between the writing style of the newspapers. Once we get terms that define a certain newspaper, we can move on to more in depth experiments.

So, for this first experiment we used Latent Dirichlet Allocation (LDA) which is a type of topic extraction based on the frequency of a word. [2] It is a method where each article in a corpus is represented as a distribution over those topics. Section 4.2.1 gives a more thorough explanation of LDA.

The second experiment has the purpose of extracting and following topics over time. The coronavirus pandemic period can be divided into different phases over the first 14 months, by researching the topics with time as a variable, we might discover other topics. For this experiment, we use Dynamic Topic Modelling which is a method that analyses the change in specific topics over time. [18]

The third experiment focuses on bi-grams in combination with word network graphs. Bi-grams are a pair of words that appear commonly in an article, an example would be "Artificial Intelligence". This experiment gives an insight on how the newspaper uses their words and how they are associated in a word graph.

3.2 Data and preprocessing

The data was obtained from the newspapers as listed in Table 1. We filtered out the articles that contained the key term *corona* in the URL link of an article which was usually the title of that article. We gathered the data of 16 months, between January 1st, 2020 until April 30, 2021. An important remark is that we let the algorithm run until it was finished for all the newspapers except NU.nl. This is because NU.nl had a different URL structure than the other newspapers. The URLs of NU.nl were based on IDs while the URLs of the others newspapers were based on the published date of the articles. We believe that the reason that we gathered relatively more articles from the Telegraaf than the other newspapers, is because the Telegraaf's URLs contained the term corona more often where the other newspapers used another term for corona in their URLs such as COVID-19.

The reason we choose these newspapers is because each of them has another type of characterization so if we would to compare them, we would expect differences between the obtained results.

De Dagelijkse Standaard is seen as corona critical, yet democratic, and placed on the very end of the conservative-right end of the spectrum. De Volkskrant is more progressive-left leaning targeting an audience with a higher education. De Telegraaf aims to target a broad and mainstream audience by providing accessible content that is easy to read. It is leaning towards conservative-right and the way they write their articles is in a more sensational way than the other newspapers, however it is far from a tabloid. NU.nl leans towards progressive-left and more liberal, their articles are easily accessible and it is a good medium to get a quick update about the news headlines. The NRC is seen as the most neutral newspaper and is positioned as a newspaper-of-record targeting high quality journalism for a higher educated audience. These assumptions are based on our own opinions on how people view these newspapers. Those assumptions are made by talking with other people about those newspapers and by reading the articles from each newspaper.

In research from December 2017 [3], three of our selected newspapers were featured; NU.nl, de Volkskrant, and de Telegraaf. Of those three newspapers, it was found that NU.nl is used the most, followed by de Telegraaf. As we assumed, of the 32% that read de Telegraaf, 13% is right-leaning and 7% is left-leaning. Within the Volkskrant and NU.nl, the majority is left-leaning. That research gives more insight about trust within the newspapers and it appears that those with populist views trust the news less than those without populist views. The trust in newspapers did not have a certain division within the left/right spectrum.

The URLs of the articles were scraped with Python's library *BeautifulSoup*. This was done by looking on the website for the page that contained an archive of corona articles. Once we knew where in the website those URLs could be found, we then scraped them one by one with only changing the date or ID number in the URL. An example of "De Volkskrant" looked like this: `https://www.volkskrant.nl/archief/' + str(year) + '/' + str(month) + '/' + str(day)`. Here we made sure that every date since the beginning of the covid situation had been scraped. We did this by writing a simple algorithm that looped through the year by changing the `day`, then the `month` and then the `year`. If a date did not exist, for example 30 February, then nothing happened and the algorithm would just continue.

Now that the URLs had been scraped, we had to get the content of those articles. So the title, main text and date, were scraped with Python's library *Trafilatura*. This library made it possible to automatically scrape the specified field of the HTML page of an URL. We specified four fields:

- Title: this is the title of the article;
- Text: this is the main text of the article;
- Date: this is the date on which the article was published;
- Host name: this is the newspaper source of the article (e.g.: NRC).

Once we had all this data, it was of the form: `Source|Date|Title|Text`

The text part of an article sometimes contains noise that we do not want to have during our experiments. So we had to clean up the collected text data. We start by replacing every enter with a space, this way the text body stays in the same row within the csv file where all the data is stored. We then lower cased each word so that it becomes clear what the frequency for each word is. Then, we removed the following symbols `[, \. ! ? ]` and thereafter, we removed the stop words which was a standard stop word list from python's library *Natural Language ToolKit*

Algorithm 1 Web scraper

```
open file # this contains the urls

for i in file do
    fetch = trafilatura.fetch_url(i)
    data = bare_extraction(fetch, include_comments = False)
    if data is not None then
        data['hostname'] = data['hostname'].replace('\n', ' ')
        data['date'] = data['date'].replace('\n', ' ')
        data['title'] = data['title'].replace('\n', ' ')
        data['text'] = data['text'].replace('\n', ' ')
        data['text'] = data['text'].replace(''Het beste van De Telegraaf', ' ')
        tuple = ( data['hostname'], data['date'], data['title'], data['text'] )
        write(tuple) to newfile.csv
```

(NLTK). We did add some words to this list, namely the newspaper names. Some newspapers, had standard sentences, copyright marks or ads in their text body. For example, Telegraaf contained: *"Het beste van De Telegraaf"*. These had to be filtered manually. Algorithm 1 shows the pseudocode of this process.

3.3 Experiment 1: LDA

To understand our first experiment, we will now give a brief explanation about LDA. LDA is short for **L**atent **D**irichlet **A**llocation which is a type of topic extraction based on the frequency of a word. LDA consists of two main principles: "Every document is a mixture of topics and every topic is a mixture of words". This means that every document contains at least one topic but usually consists of multiple (sub)topics. The second part of the sentence means that a cluster of words form a topic. LDA works by estimating those two parts at the same time. [26]

The input needed for LDA is a corpus and a dictionary. A corpus is essentially a collection of text documents and a dictionary contains all the unique words within the corpus. For our experiment, we created a corpus for each newspaper individually. LDA is usually used with a corpus in bag-of-words format which converts a text into vectors based on term frequency. We also used this representation for our experiment.

For the experiment, we filtered out the terms that are less relevant on their own when we look at their meaning, for example a verb does not give us much meaning since it does not say something about another term. We applied *Part-of-speech tagging* from Python's library *spaCy*. This works by breaking a sentence down in words and predicting the type of context of each word. For our experiment the following text tags were removed: {`"VERB", "ADP", "PUNCT", "NUM", "SYM", "AUX", "ADV", "CONJ", "DET", "PART", "PRON", "SCONJ", "X"`}. An overview of all the text tags and their meaning can be found in figure 1 [8]. Now, the remaining text mostly contains nouns which will give us the most meaningful results. With these nouns, we created a corpus that contains the nouns of each article of one of the Dutch newspapers. We also created a dictionary of all these words so the LDA tool can compare which nouns are more important.

Then we prepared the input variables for the LDA tool so we tokenized the sentences within our

data and we converted this data set into a bag-of-words format. Our corpus is now complete. For the dictionary, we used the same data set and applied a function that converted this data set into a dictionary. We used all the unique words, for each newspaper we gathered this many unique words:

- De Volkskrant: 48095
- De Telegraaf: 30724
- Het NRC: 23593
- Dagelijkse Standaard: 13352
- NU.nl: 10908

Other parameters that were needed for the LDA function, are the amount of topics of which we will give a thorough explanation in section 3.3.2 and the amount of passes which is how many times we run the LDA function in order to get more accurate results. Based on trial and error, we kept the number of passes on the amount of three. The function then looks like this: `gensim.models.LdaMulticore(corpus=corpus, id2word=dictionary, num_topics= 5, passes = 3)`. Now that everything is ready, we are able to run the LDA tool, the following subsection shows the results.

3.3.1 Salient terms

We applied the LDA function on each of the newspapers independently. Below the results can be found for each of the newspapers. Initially we choose to distribute the set of articles over 10 topics where each topic contains 30 of the most important terms. The LDA shows then 10 topics where each topic contains 30 salient terms and a general top 30 salient terms which are the salient terms over all 10 topics. Table 2 shows the general top 30 salient terms for each newspaper. The salient terms are the terms that are the most informative terms for the determination of the generated topic. [5] The saliency is computed by calculating the likelihood that a term w is generated by topic t and by comparing this word to the likelihood that another, randomly chosen, term w' is generated by the same topic t. [5] The following formula shows how this calculation is done.

$$\text{saliency}(w) = \text{frequency}(w) \cdot (\sum p\,(t|w) \cdot \frac{\log\,(p(t|w))}{p(t)})$$

This does not mean that the salient terms are the terms that are characteristic the writing style of the newspapers but rather that those are the most relevant terms that distinguish the different topics. So with the salient terms, we are able to interpret the topics for each newspaper independently. Therefore we can extract from the salient terms whether the topics of the different newspapers are distinguishable from each other.

In order to make the difference between the salient terms of each newspaper more obvious, we will remove the terms that are present in all the newspapers, besides this, we will also remove the terms that give an indication of time. We end up with Table 3.

To try to answer our research question of whether there are underlying topics, we will have a closer look at the cleaned up table, Table 3. In this table, the terms kept their original place in the top 30 salient terms. The placing tells us which term is more important within a newspaper where

POS	DESCRIPTION	EXAMPLES
ADJ	adjective	*big, old, green, incomprehensible, first*
ADP	adposition	*in, to, during*
ADV	adverb	*very, tomorrow, down, where, there*
AUX	auxiliary	*is, has (done), will (do), should (do)*
CONJ	conjunction	*and, or, but*
CCONJ	coordinating conjunction	*and, or, but*
DET	determiner	*a, an, the*
INTJ	interjection	*psst, ouch, bravo, hello*
NOUN	noun	*girl, cat, tree, air, beauty*
NUM	numeral	*1, 2017, one, seventy-seven, IV, MMXIV*
PART	particle	*’s, not,*
PRON	pronoun	*I, you, he, she, myself, themselves, somebody*
PROPN	proper noun	*Mary, John, London, NATO, HBO*
PUNCT	punctuation	*., (,), ?*
SCONJ	subordinating conjunction	*if, while, that*
SYM	symbol	*$, %, §, ©, +, −, ×, ÷, =, :), 😝*
VERB	verb	*run, runs, running, eat, ate, eating*
X	other	*sfpksdpsxmsa*
SPACE	space	

Figure 1: An overview of all the spaCy text tags

Newspaper	Top 30 salient terms
Dagelijkse Standaard	1. Mensen; 2. Nederland; 3. Coronavirus; 4. Nieuwe; 5. Maatregelen; 6. Kabinet; 7. Virus; 8. Goed; 9. Natuurlijk; 10. Overheid; 11. Heel; 12. Aantal; 13. Corona; 14. Jaar; 15. Jonge; 16. Landen; 17. Tijd; 18. Echt; 19. Grote; 20. Land; 21. Snel; 22. Rivm; 23. Lockdown; 24. Dag; 25. Nederlandse; 26. Coronacrisis; 27. Keer; 28. Week; 29. Beleid; 30. Cijfers.
NRC	1. Mensen; 2. Aantal; 3. Jaar; 4. Week; 5. Procent; 6. Vaccin; 7. Nieuwe; 8. Coronavirus; 9. Nederland; 10. Miljoen; 11. Kabinet; 12. Dag; 13. Positieve; 14. Dinsdag; 15. Rivm; 16. Uur; 17. Patienten; 18. Land; 19. Virus; 20. Avondklok; 21. Maandag; 22. Minister; 23. Ziekenhuizen; 24. Europese; 25. Landen; 26. Woensdag; 27: Deel; 28. Tweede; 29. Weken; 30. Politie.
Nu.nl	1. Mensen; 2. Vaccin; 3. Aantal; 4. Jaar; 5. Week; 6. Nederland; 7. Procent; 8. Miljoen; 9. Coronavirus; 10. Dagen; 11. Kabinet; 12. Rivm; 13. Avondklok; 14. Britse; 15. Januari; 16. Dinsdag; 17. Maandag; 18. Variant; 19. Dag; 20. Nieuwe; 21. Woensdag; 22. Uur; 23. Tweede; 24. Virus; 25. Positieve; 26. Ruim; 27. Weken; 28. Lockdown; 29. Politie; 30. Doses.
Volkskrant	1. Mensen; 2. Jaar; 3. Corona; 4. Virus; 5. Procent; 6. Nederland; 7. Nieuwe; 8. Aantal; 9. Goed; 10. Week; 11. Coronavirus; 12. Grote; 13. Tijd; 14. Heel; 15. Dag; 16. Tweede; 17. Coronacrisis; 18. Maatregelen; 19. Kabinet; 20. Land; 21. Euro; 22. Miljoen; 23. Weken; 24. Lockdown; 25. Landen; 26. Patienten; 27. Vaccin; 28. Nederlandse; 29. Snel; 30. Onderzoek.
Telegraaf	1. Mensen; 2. Aantal; 3. Coronavirus; 4. Corona; 5. Nieuwe; 6. Jaar; 7. Week; 8. Patienten; 9. Proces; 10.Coronacrisis; 11. Nederland; 12. Virus; 13. Dagen; 14. Coronapatienten; 15. Miljoen; 16. Dag; 17. Positief; 18. Ziekenhuizen; 19. Tweede; 20. Positieve; 21. Rivm; 22. Land; 23. Uur; 24. Ziekenhuis; 25. Amsterdam; 26. Vaccin; 27. Maatregelen; 28. Ruim; 29. Weken; 30. Euro.

Table 2: Top 30 salient terms per Dutch newspaper

1 is the most important term and 30 is the less important term. One way of a newspaper to have underlying topics within an article is by expressing their opinion. A manner to express an opinion is by using adjectives because they modify the meaning of a word [30], therefore we will count the amount of adjectives within the cleaned up top 30 salient terms. We got the following results:

- Dagelijkse Standaard: 7 times
- Volkskrant: 6 times
- Nu.nl: 4 times
- Telegraaf: 4 times
- NRC: 3 times

Based on these results, we can conclude that De Dagelijkse Standaard uses adjectives the most, followed closely by De Volkskrant. This may mean that those newspapers are more likely to express their opinion than the other newspapers. This could lead to creating underlying topics that may influence the reader to adopt the same opinion as the underlying opinion of the newspaper. We are not able to conclude whether a more extensive use of adjectives leads to creating more underlying topics because the usage of adjectives do not directly cause the creation of subtopics.

When we apply the LDA function, we also have to specify the term relevance λ which essentially implies the relevance of that term. This is a metric that is adjustable between 0 and 1 λ and

Newspaper	Top 30 salient terms
Dagelijkse Standaard	5. Maatregelen; 6. Kabinet; 8. Goed; 9. Natuurlijk; 10. Overheid; 11. Heel; 13. Corona; 15. Jonge; 16. Landen; 17. Tijd; 18. Echt; 19. Grote; 20. Land; 21. Snel; 22. Rivm; 23. Lockdown; 25. Nederlandse; 26. Coronacrisis; 27. Keer; 29. Beleid; 30. Cijfers.
NRC	5. Procent; 6. Vaccin; 10. Miljoen; 11. Kabinet; 13. Positieve; 15. Rivm; 17. Patienten; 18. Land; 20. Avondklok; 22. Minister; 23. Ziekenhuizen; 24. Europese; 25. Landen; 26. Woensdag; 27: Deel; 28. Tweede; 30. Politie.
Nu.nl	2. Vaccin; 7. Procent; 8. Miljoen; 11. Kabinet; 12. Rivm; 13. Avondklok; 14. Britse; 15. Januari; 18. Variant; 21. Woensdag; 23. Tweede; 25. Positieve; 26. Ruim; 28. Lockdown; 29. Politie; 30. Doses.
Volkskrant	3. Corona; 5. Procent; 9. Goed; 12. Grote; 13. Tijd; 14. Heel; 16. Tweede; 17. Coronacrisis; 18. Maatregelen; 19. Kabinet; 20. Land; 21. Euro; 22. Miljoen; 24. Lockdown; 25. Landen; 26. Patienten; 27. Vaccin; 28. Nederlandse; 29. Snel; 30. Onderzoek.
Telegraaf	4. Corona; 8. Patienten; 9. Proces; 10.Coronacrisis; 14. Coronapatienten; 15. Miljoen; 16. Dag; 17. Positief; 18. Ziekenhuizen; 19. Tweede; 20. Positieve; 21. Rivm; 22. Land; 24. Ziekenhuis; 25. Amsterdam; 26. Vaccin; 27. Maatregelen; 28. Ruim; 30. Euro.

Table 3: Cleaned top salient terms per Dutch newspaper

according to the original paper, the optimal relevance is set at 0,6. [2] Therefore, we applied this relevance score to our analysis.

3.3.2 Number of topics

Initially we choose to experiment with 10 topics. The reason for this is that with such a large data set, the expected result is that there will be a lot of different topics. An amount of 10 topics would definitely cover the most important ones. However, when we looked at the results, there was a lot of overlap between the topics. This overlap caused clusters to form, thus we believe that it is better to experiment with less amount of topics so we can get a better and clearer interpretation of the topics.

We experimented with the following amounts:

- 10 topics: here we saw that this amount caused lots of overlapping between the topics which caused clustering;
- 7 topics: still some overlap within most newspapers;
- 5 topics: less overlap and the topics are more defined;
- 3 topics: no overlap and the topics are less detailed.

Figure 2 below shows examples of this experiment with different amount of topics. They all belong to the same newspaper, "De Telegraaf".

The articles are collected over a time span of 14 months and they all have one topic in common, the coronavirus. Because of this, we cannot expect that there will be a lot of different topics. When we looked at the results with 7 and 10 topics, we noticed that it was quite difficult to distinguish different topics and because of the overlap, the topics had many terms in common which made it more difficult to see a difference. Therefore, we will only discuss the results with topic amounts 5. However, even with this amount of topics, the differences were not distinguishable unfortunately.

Because of this, we decided to apply TF-IDF into our LDA analysis, this is a classifier that is not just based on term frequency but rather on term relevance. The following subsection provides more insight onto this experiment.

3.3.3 TF-IDF

To get a better understanding of the topics and the term relevance, we analyzed our corpus based on TF-IDF computations. We computed this score for every document within a newspaper, then we summed up all the scores for each term and calculated the mean score value. This computation makes it possible to view the relevant terms within the corpus instead of the most frequent terms. The computations are done on each newspaper independently from the other newspapers.

The understand the difference between regular *bag-of-words* and *TF-IDF* weighted text representations better, we will explain how TF-IDF computations are done. TF-IDF stands for **T**erm **F**requency **I**nversed **D**ocument **F**requency and it measures the term *relevance* instead of the term *frequency* which Bag-of-words does. The formula below shows how this is done: [28]

$$tf(t) * idf$$

where $tf =$

$$\frac{\text{term count of t}}{\text{total amount of terms}}$$

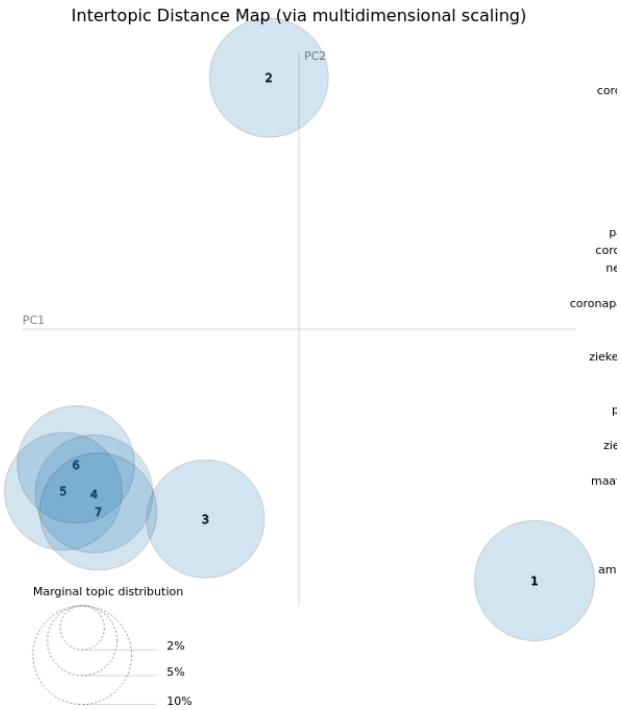
and where $idf(t) =$

$$\log(\frac{\text{total number of documents}}{\text{number of documents where term t appears}})$$

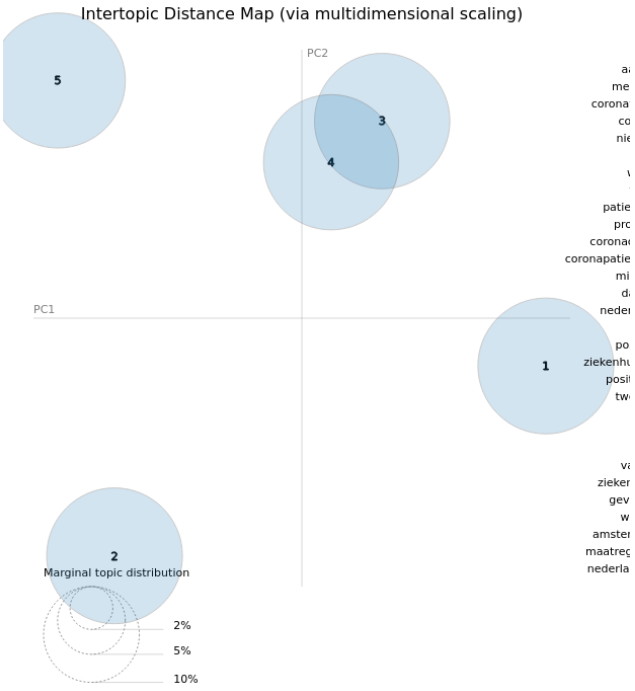
As becomes visible, the main difference between TF-IDF and bag-of-words is that TF-IDF searches all the documents where a certain term appears which contributes to the relevance of a term. For example, the word "the" is used a lot in the English language, however for text data analysis, "the" does not have a relevant meaning. These type of words are then filtered out since they occur a lot in every document. By doing this, we get the most insightful results.

For our LDA analysis, we did not compute this ourselves but used a python in-build function called TfidfVectorizer from the library *sklearn*. We first created a corpus which contained all the documents for a particular newspaper, just like with the bag-of-words LDA analysis. Then we applied the TfidfVectorizer which tokenizes the words in our corpus and then applies the TF-IDF analysis on those words. We went a step ahead and calculated the average value for each term and then we took the top 40 terms that scored the highest average (so also a considerably smaller set

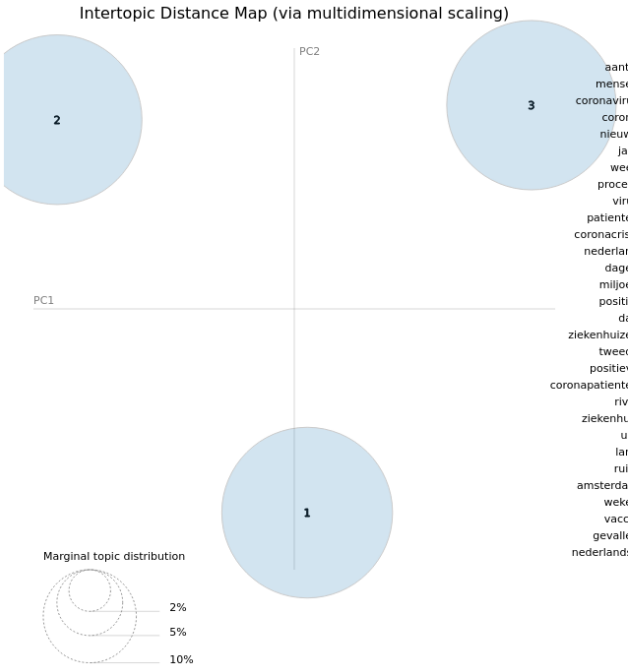
(a) 10 Topics



(b) 7 Topics



(c) 5 Topics



(d) 3 Topics

Figure 2: The different amount of topics within LDA analysis of the Telegraaf

of keywords). Now we have a new corpus to work with. We applied a TfidfModel on this corpus which enables us to fit this model in our LDA analysis. [25]

Figures 4 till 7 below show the results of this experiment. Unfortunately the computations of the Telegraaf did not succeed and therefore we cannot show it's result.

Newspaper	**NRC**
Topic 1	1. Afgelopen; 2. Week; 3. Ziekenhuizen; 4. Coronavirus; 5. IC; 6. Etmaal; 7. Zondag; 8. Ziekenhuis; 9. Europese; 10. Meldingen; 11. Doses; 12. Tests; 13. Meldt; 14. Gemeld; 15. Opgenomen; 16. Ontbreken; 17. Pfizer; 18. Mogelijk; 19. Premier; 20. Daling; 21. Reguliere; 22. Zorg; 23. Zuid; 24. Februari; 25. Getest; 26. Lockdown; 27. Versoepelingen; 28. Hoogste; 29. Bekend; 30. Gaan.
Topic 2	1. Aantal; 2. Jaar; 3. Procent; 4. Patienten; 5. Kabinet; 6. Dag; 7. LCPS; 8. Cijfers; 9. Politie; 10. Intensive; 11. Euro; 12. Zegt; 13. EU; 14. Sterfgevallen; 15. Sinds; 16. Wel; 17. Elementen; 18. Onderzoek; 19. Overleden; 20. President; 21. Janssen; 22. Tweede; 23. Maart; 24. Vertraging; 25. Blijkt; 26. Toegenomen; 27. Jongeren; 28. Zei; 29. Middel; 30. Curacao.
Topic 3	1. Vaccin; 2. Rivm; 3. Positieve; 4. Covid; 5. Vaccins; 6. Nederland; 7. Minder; 8. Vrijdag; 9. Liggen; 10. Weer; 11. Astrazeneca; 12. Maandag; 13. Jonge; 14. Besmettingen; 15. Britse; 16. Care; 17. April; 18. Ema; 19. Printversie; 20. Afdelingen; 21. Corona; 22. Coronacrisis; 23. Mogen; 24. Gemeente; 25. Positief; 26. Amsterdam; 27. Duitse; 28. Bezoekers; 29. Geregistreerd; 30. Kamer.
Topic 4	1. Nieuwe; 2. Miljoen; 3. Eerste; 4. Eerder; 5. Variant; 6. Uur; 7. Woensdag; 8. Coronatests; 9. Virus; 10. Land; 11. Per; 12. Twee; 13. Vanaf; 14. Werden; 15. Minister; 16. Artikel; 17. Miljard; 18. Duitsland; 19. Ruim; 20. Tussen; 21. Testen; 22. Totaal; 23. Momenteel; 24. Prik; 25. Liveblog; 26. Dagen; 27. Open; 28. Besmet; 29. Ministerie; 30. Vaak.
Topic 5	1. Mensen; 2. Volgens; 3. Coronapatienten; 4. Dinsdag; 5. Zaterdag; 6. Donderdag; 7. Landelijk; 8. Spreiding; 9. GGD; 10. Rutte; 11. Landen; 12. Moeten; 13. Nederlandse; 14. Avondklok; 15. NB; 16. Januari; 17. Demissionair; 18. Vorig; 19. Doorgegeven; 20. WHO; 21. OMT; 22. Opzichte; 23. Tijdens; 24. Scholen; 25. Alle; 26. Bedrijven; 27. Weken; 28. Versie; 29. Eiland; 30 Rijksinstituut.

Table 4: Results NRC

Newspaper	**NU.nl**
Topic 1	1. Vaccins; 2. Dagen; 3. Avondklok; 4. Eerste; 5. Zeven; 6. Jonge; 7. Pfizer; 8. Coronapatienten; 9. Lockdown; 10. Weer; 11. Dinsdag; 12. Sterfgevallen; 13. Vaccineren; 14. Zaterdag; 15. Weken; 16. Abonneren; 17. Zorgmedewerkers; 18. Mailtje; 19. Ingeent; 20. Krijgen; 21. Kamer; 22. Extra; 23. Personen; 24. Effect; 25. Kwartaal; 26. Goed; 27. Quarantaine; 28. Groep; 29. Tussen; 30 Acute.
Topic 2	1. GGD; 2. Nieuwe; 3. Doses; 4. Europese; 5. December; 6. Woensdag; 7. EMA; 8. Gemeente; 9. Vaccinatie; 10. Moeten; 11. Wel; 12. Moderna; 13. Scholen; 14. Beluisteren; 15. Ministerie; 16. Podcasts; 17. Podcastapp; 18. Bedrijf; 19. Premier; 20. Etmaal; 21. Mutatie; 22. Cijfers; 23. Vorig; 24. Onze; 25. OMT; 26. Verpleeghuizen; 27. Daling; 28. Klachten; 29. Prik; 30. Maart.

Topic 3	1. Aantal; 2. Mensen; 3. Britse; 4. Ziekenhuizen; 5. Minder; 6. Januari; 7. Patienten; 8. Virus; 9. Kinderen; 10. Zondag; 11. Uur; 12. Maandag; 13. IC; 14. Podcast; 15. Voorgaande; 16. Coronavirus; 17. Gevaccineerd; 18. Rutte; 19. Zuid; 20. Stuur; 21. Positief; 22. Tijdens; 23. Februari; 24. Land; 25. Ziekenhuis; 26. Inwoners; 27. Gemeld; 28. Koninkrijk; 29. Eerder; 30. Meldingen.
Topic 4	1. RIVM; 2. Miljoen; 3. Nederland; 4. Kabinet; 5. Gemiddelde; 6. Zorg; 7. Tests 8. Testen; 9. Nieuws; 10. Woordvoerder; 11. Onderzoek; 12. Miljard; 13. Overleden; 14. Covid; 15. Abonneer; 16. Per; 17. Zegt; 18. Donderdag; 19. Euro; 20. Werden; 21. Getest; 22. Landen; 23. Maatregelen; 24. Open; 25. Advies; 26. Regio; 27. Meldt; 28. Besmet; 29. Minister; 30. Apparaat.
Topic 5	1. Vaccin; 2. Jaar; 3. Positieve; 4. Procent; 5. Afgelopen; 6. Week; 7. Politie; 8. Astrazeneca; 9. Variant; 10. Besmettingen; 11. Dag; 12. EU; 13. Spotify; 14. Medewerkers; 15. Opgenomen; 16. Feedback; 17. Gedownloade; 18. Middag; 19. Maken; 20. Prikken; 21. Verenigd; 22. Ochtendpodcast; 23. Via; 24. Dagcijfers; 25. Nederlanders; 26. Verspreiding; 27. Corona; 28. Intensive; 29. Hoger; 30 Dosis.

Table 5: Results NU.nl

Newspaper	**Dagelijkse Standaard**
Topic 1	1. Aantal; 2. Besmettingen; 3. Procent; 4. Rutte; 5. Gewoon; 6. Premier; 7. Onderzoek; 8. EU; 9. Europese; 10. Oh; 11. Aanpak; 12. Hugo; 13. Leerlingen; 14. Britse; 15. Wappies; 16. Afgelopen; 17. Keer; 18. Crisis; 19. Geld; 20. Zorgen; 21. Italie; 22. Alternatief; 23. Snel; 24. Volkomen; 25. Variant; 26. Economische; 27. Inconsistent; 28. Zingen; 29. Aantreden; 30 Namelijk.
Topic 2	1. Kinderen; 2. Coronabeleid; 3. Onze; 4. Maatregelen; 5. Overheid; 6. Nieuwe; 7. Weken; 8. Steeds; 9. Heel; 10. Avondklok; 11. Willen; 12. Amsterdam; 13. Spelen; 14. Kamer; 15. Vanaf; 16. Willem; 17. Term; 18. New; 19. Hadden; 20. Positief; 21. Middelbare; 22. Thierry; 23. Eigen; 24. Deal; 25. Grote; 26. Werk; 27. Ultieme; 28. Logischerwijs; 29. Grootse; 30. Chinese.
Topic 3	1. Jonge; 2. Kabinet; 3. RIVM; 4. Weer; 5. Partij; 6. Kritiek; 7. Lockdown; 8. Kans; 9. Gaan; 10. Cijfers; 11. Virus; 12. Lijkt; 13. Baudet; 14. Terwijl; 15. School; 16. Landen; 17. Week; 18. Vaccin; 19. GGD; 20. Nooit; 21. Besmet; 22. Nederlanders; 23. China; 24. Dupe; 25. Gisteren; 26. Sigrid; 27. Horeca; 28. Nieuws; 29. Coronafonds; 30. Toekomstige.
Topic 4	1. Beleid; 2. Wel; 3. Coronavirus; 4. Jaar; 5. Jongeren; 6. Testen; 7. Open 8. Gaat; 9. Corona; 10. Land; 11. Getest; 12. Patienten; 13. Ziekenhuis; 14. Per; 15. Duurzame; 16. App; 17. Klaver; 18. Bijna; 19. Blijft; 20. Duitsland; 21. Gouden; 22. Wappies; 23. Ziekenhuizen; 24. CDA; 25. Houden; 26. Zelfs; 27. Interviewer; 28. IC; 29. Refereert; 30. Referendum.
Topic 5	1. Mensen; 2. Nederland; 3. Moeten; 4. Mondkapjes; 5. Totaal; 6. Ondernemers; 7. FVD; 8. Ondertussen; 9. Euro; 10. Politie; 11. Dag; 12. Neemt; 13. Doden; 14. Minister; 15. Kaag; 16. Echt; 17. Risico; 18. Idee; 19. Vandaag; 20. Gestoorde; 21. Koning; 22. Mogen; 23. Oprichter; 24. Partijcongres; 25. Zinken; 26. Green; 27. Filosofie; 28. Gretig; 29. Wet; 30 Scholen.

Table 6: Results Dagelijkse Standaard

Newspaper	**Volkskrant**
Topic 1	1. Jaar; 2. Virus; 3. Weer; 4. Maatregelen; 5. Ziekenhuis; 6. Jonge; 7. President; 8. Week; 9. Dag; 10. Test; 11. IC; 12. Nederlanders; 13. Geld; 14. Vaccinatie; 15. App; 16. Ouderen; 17. Den; 18. Werknemers; 19. Gemeente; 20. Zorgen; 21. Artsen; 22. Mondkapjes; 23. Verpleeghuizen; 24. Zei; 25. Reizigers; 26. Steeds; 27. Ondernemers; 28. Hoger; 29. Tests; 30 Sociale.
Topic 2	1. Besmettingen; 2. Euro; 3. Vaccin; 4. Coronacrisis; 5. Wel; 6. Minder; 7. Avondklok; 8. EU; 9. Bedrijf; 10. Eerste; 11. Minister; 12. Premier; 13. Onze; 14. Klachten; 15. Overheid; 16. Getest; 17. Economie; 18. Regio; 19. Coronavaccin; 20. Ziekenhuizen; 21. Regels; 22. VS; 23. Wereld; 24. Johnson; 25. Besmet; 26. Tijdens; 27. Man; 28. Correspondent; 29. Advies; 30. Lidstaten.
Topic 3	1. Zegt; 2. Landen; 3. Miljoen; 4. Lockdown; 5. GGD; 6. Nieuwe; 7. Per; 8. Laatste; 9. Live; 10. Jongeren; 11. Nieuws; 12. Volg; 13. Britse; 14. Inwoners; 15. Kwartaal; 16. Vaccineren; 17. Groep; 18. Extra; 19. Duitsland; 20. Open; 21. Ouders; 22. Volgens; 23. Nederlandse; 24. OMT; 25. Laten; 26. Ruim; 27. Economische; 28. Stad; 29. Sterfte; 30. Horeca.
Topic 4	1. Procent; 2. Aantal; 3. Coronavirus; 4. Covid; 5. Rutte; 6. China; 7. Onderzoek 8. Regering; 9. Europese; 10. Amerikaanse; 11. Golf; 12. Uitbraak; 13. Chinese; 14. Wij; 15. Variant; 16. Moeten; 17. Goed; 18. Ministerie; 19. CBS; 20. Coronavaccins; 21. Media; 22. Gemeenten; 23. Dollar; 24. Italie; 25. Commissie; 26. Coronabeleid; 27. Astrazeneca; 28. Vrouw; 29. Twee; 30. Rotterdam.
Topic 5	1. Mensen; 2. Kabinet; 3. Nederland; 4. Corona; 5. Patienten; 6. Vaccins; 7. Testen; 8. RIVM; 9. Cijfers; 10. Duizend; 11. Miljard; 12. Kinderen; 13. Tweede; 14. Land; 15. Trump; 16. Bedrijven; 17. Zorg; 18. Kamer; 19. Europa; 20. Liveblog; 21. Elkaar; 22. Crisis; 23. Scholen; 24. Heel; 25. Onderwijs; 26. Pfizer; 27. Biden; 28. Positief; 29. Gevaccineerd; 30 Personen.

Table 7: Results Volkskrant

From the results, it becomes clear that there still is not a lot of differentiation between the different newspapers and topics. However, we are able to conclude some statements. The Dagelijkse Standaard mentions "vaccin" only once while the other newspapers mention it more than once, another conclusion is that within the NRC topic 5 stands out from the other topics as it contains many institutions that expert in certain topics of corona. For a more thorough explanation of the results, please refer to the Discussion in section 4.2.1.

3.4 Experiment 2: Term changes over time

The governmental measures to combat the coronavirus varied over time. We would like to get more nuanced results to answer our research questions so the following experiment analyses the difference

in the newspapers throughout time.

3.4.1 Relative term frequency

We selected some keywords of which we will think that will develop differently over time. This is calculated as follows:

$$\frac{\text{term frequency per month}}{\text{total term frequency}} * 100\%$$

Figures 3 and 5 show the percentage of the frequency of the selected keywords within the *Dagelijkse Standaard* and Figures 4 and 13 show the percentage of the frequency of the selected keywords within the *Volkskrant*. In figure 3 we can clearly see a connection between "Lockdown " and "Baudet". We took his name in our selected keywords since he influences people to get into the beliefs of his party, he was also one of the political parties whom was very active and vocal about the corona measures. The Dagelijkse Standaard is also oriented more towards conservative readers so it was expected that the name "Baudet" would appear in this newspaper. When we compare this to figure 4, we see that there does not seem to be a connection between "Lockdown" and "Baudet". Moreover it seems that "Baudet" is mostly used in the beginning of the pandemic and in the beginning of 2021 whereas "Lockdown" is used over the whole period of time.

We also see that within both the newspapers, the terms "China" and "Wuhan" score high in the beginning of the pandemic and lower towards the end. This is because the coronavirus started there and this was in the beginning of the year. The corona virus did not reach the Netherlands until around march and it was then that the Netherlands had the first lockdown, the terms "Wuhan" and "China" were largely used then.

Another observation point is that the term "Lockdown" is not used much during the summer months in figure 3. This is because there were less people who got the virus during the summer and as a result, there was no lockdown necessary. Whereas in the winter months, people naturally get more sick with corona and as a consequence we got in lockdown.

In figures 5 and 13 five other key terms are shown of which we think will develop over time. There is a rise in the terms "Wappie", "Vrijheid" and "Avondklok" at around the same time frame within both the figures. This is because of the evening curfew that was around this time. "Wappie" is a term that describes the people in the Netherlands that are against the corona measures and do not agree with it, as a consequence, when the evening curfew was announced, a lot of those people started to protest and then people called them the "Corona Wappies".

There is also a increase visible towards the end of the graph in figure 5 of the term "Complot" (conspiracy). This is because a lot of fake news and fake theories started to spread when the vaccine became available. Within figure 13 we notice that the term "Wappie" is also used around the beginning of the pandemic while that term was not widely used then, we also notice that the term "Complot" appears throughout the whole period of time with peaks at the beginning of the pandemic and during the end of 2020. A reason for this might be that with this virus there was a lot of uncertainty in the beginning and therefore much speculation, and a reason for the peak to be around the end of 2020 is because after a year people were starting to be fed up with this virus and create conspiracy theories.

Now we have an idea of what terms we can expect to change we will do a dynamic topic modelling experiment

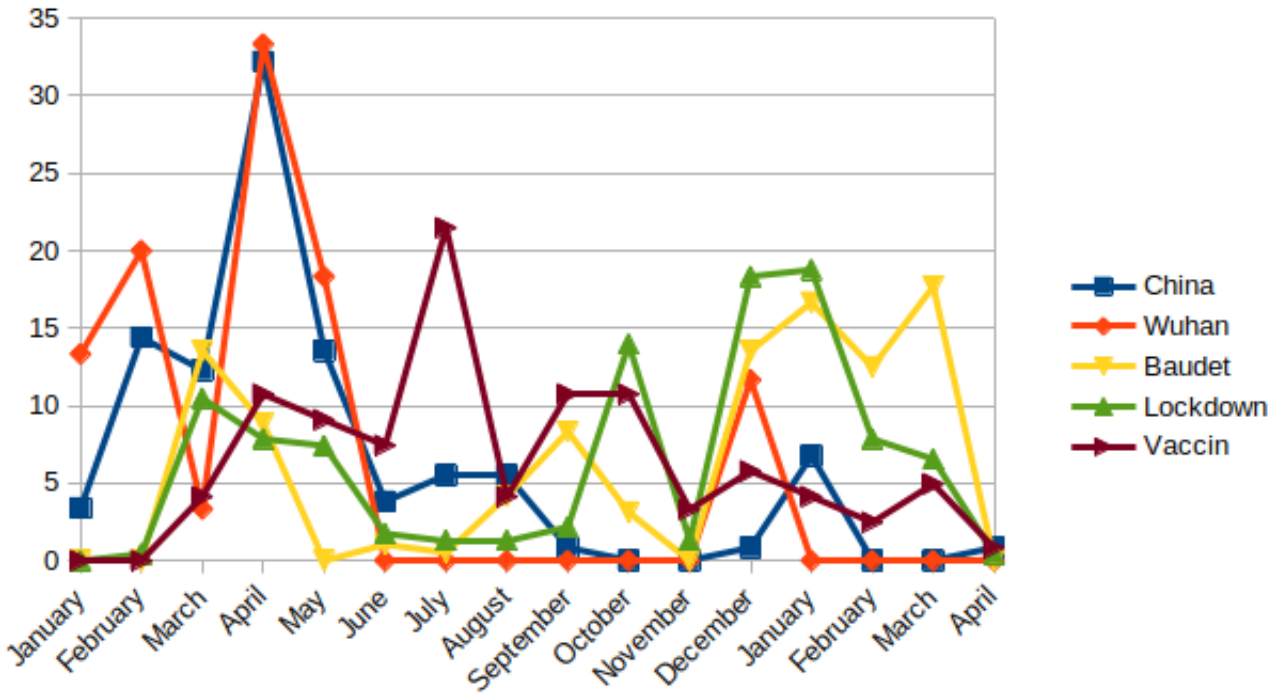


Figure 3: Term frequency from Jan 2020 until April 2021, Dagelijkse Standaard

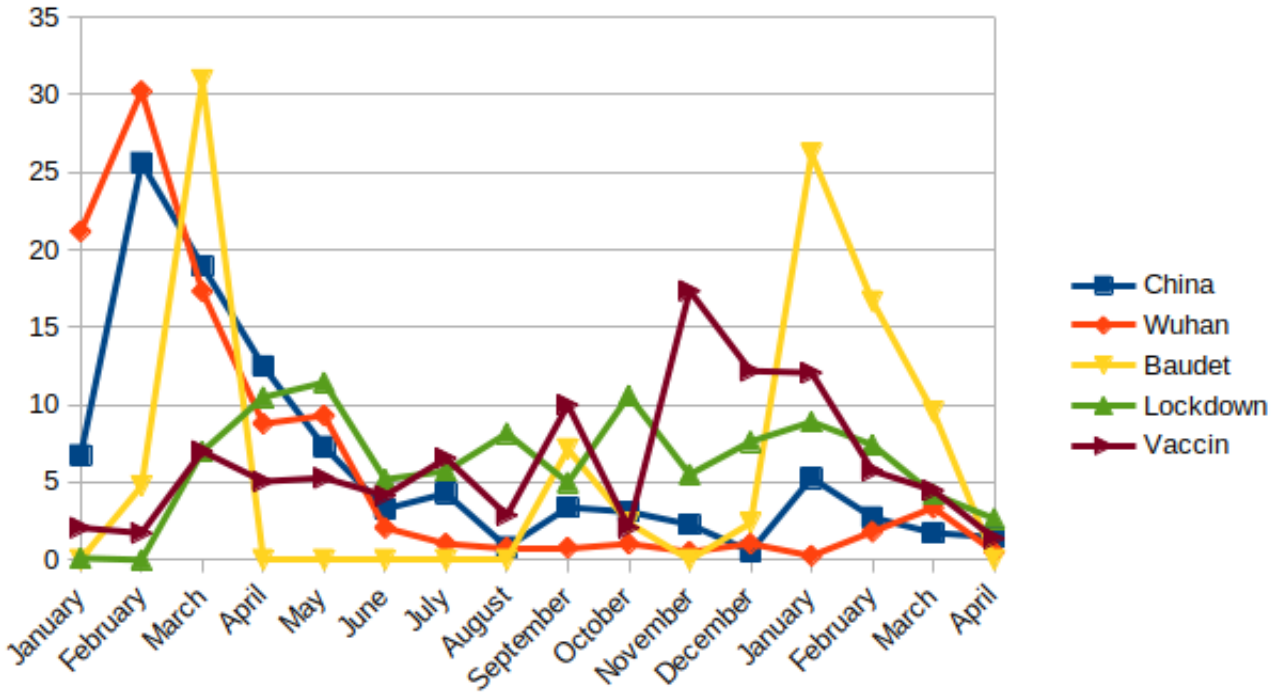


Figure 4: Term frequency from Jan 2020 until April 2021, Volkskrant

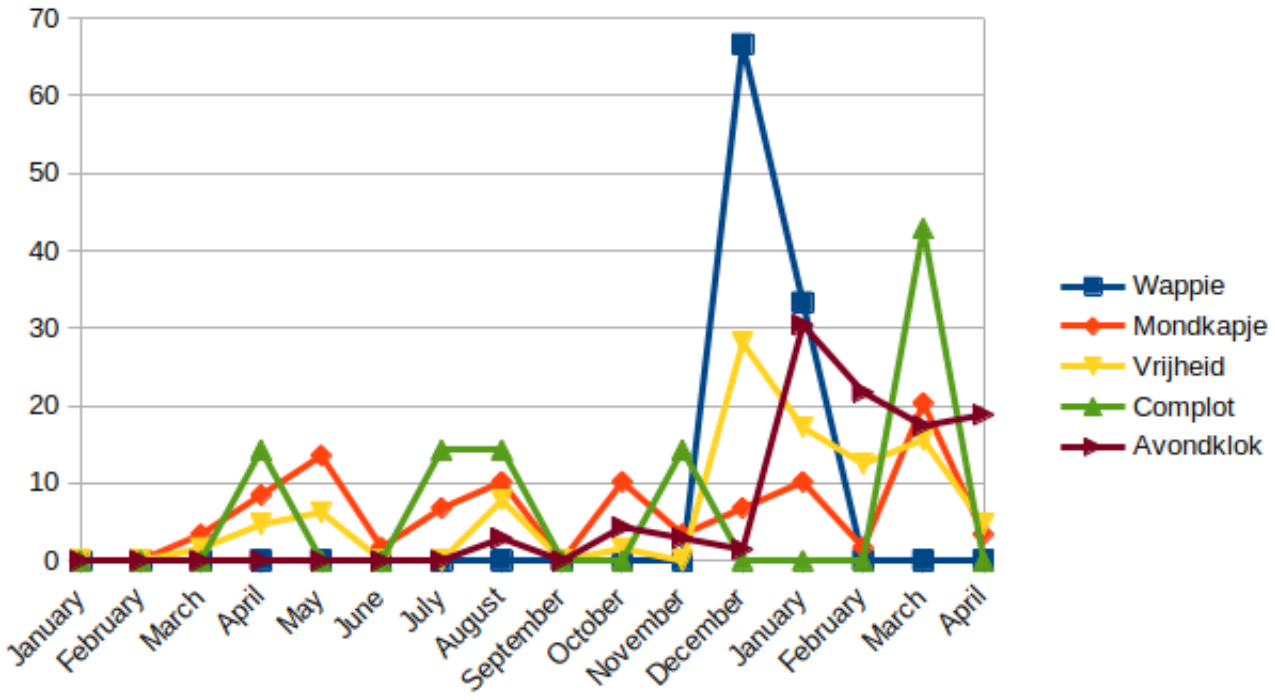


Figure 5: Term frequency from Jan 2020 until April 2021, Dagelijkse Standaard

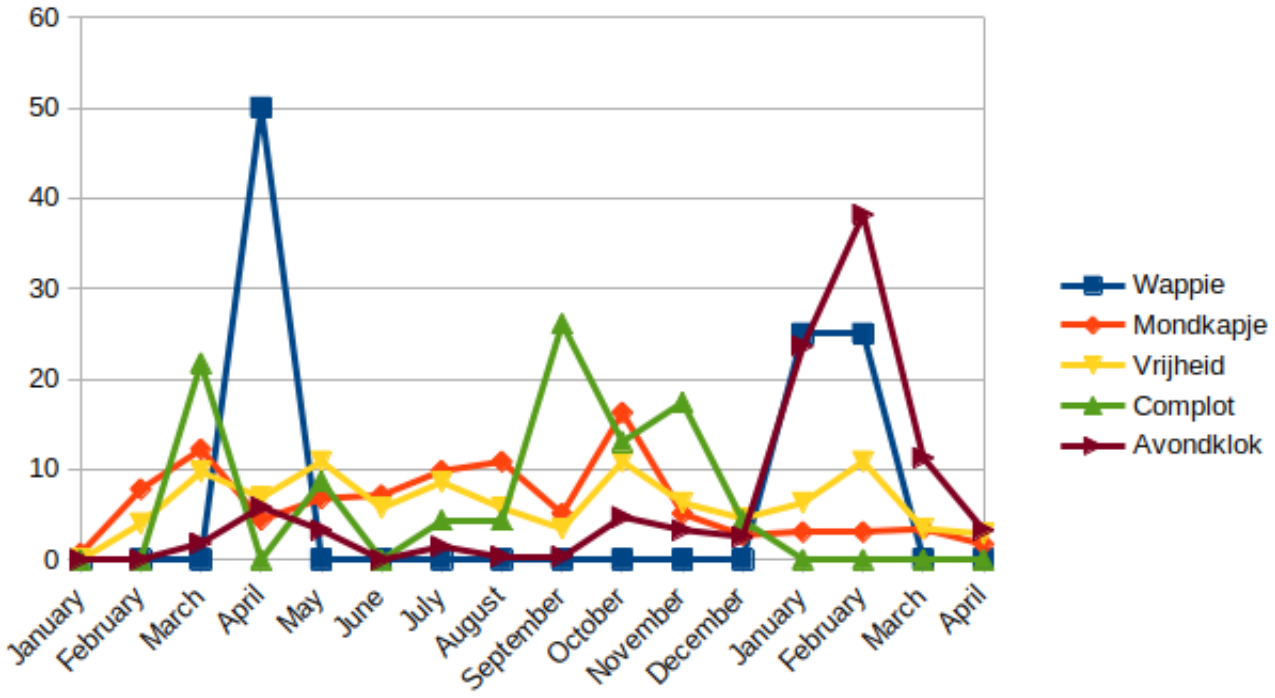


Figure 6: Term frequency from Jan 2020 until April 2021, Volkskrant

3.4.2 Dynamic topic modelling

Dynamic topic modelling is a model which evaluates topics within a data set over time by keeping the same context over different time frames but the key terms could change. [23] Originally the dynamic topic modelling is done in C++ so we used a gensim wrapper so we can use this function in Python. We want to see topic development over the months by dividing the data set over the time. We divided the articles over 16 months where each month contained the corresponding articles. Unfortunately this function did not work well with our data set, maybe because the data set was not big enough. Figure 7 shows how each term within that topic developed over time, as becomes visible, those terms did not change a bit. Unfortunately, the other 4 topics had the same result so we will not show those. Figure 8 shows how those topics as whole changed over time.

To get something out of the results, we show the salient terms for the topics:

- Topic 0: "aantal", "coronavirus", "kabinet", "lockdown", "maatregelen", "mensen", "nederland", "nieuwe", "tijd", and "virus".
- Topic 1: "wethouders", "experts", "covid-19", "kort", "macrocijfers", and "slogans".
- Topic 2: "belangen", "bescherming", "dekkingsgrens", "eind", "fors", "knot", "optimale", "pensioenverhoging", "schulden", and "solidariteit".
- Topic 3: "goed", "jonge", "kabinet", "mensen", "minister", "natuurlijk", "nederland", "overheid", and "wet".
- Topic 4: "discipline", "draghi", "economen", "engelse", "hoogleraren", "kapitaal", "koppeling", "percentage", "saldo", and "verslaafd".

We believe that Topic 0 has a higher distribution than the other topics, because it contains key words that were frequently used throughout the whole pandemic, especially the terms "coronavirus" and "lockdown". Both topic 2 and 4 seem to describe economics which explains why they are relatively close in the graph. We also see that topic 3 rises in the sixth period which is in July 2020. As becomes visible, it is difficult to interpret these results.

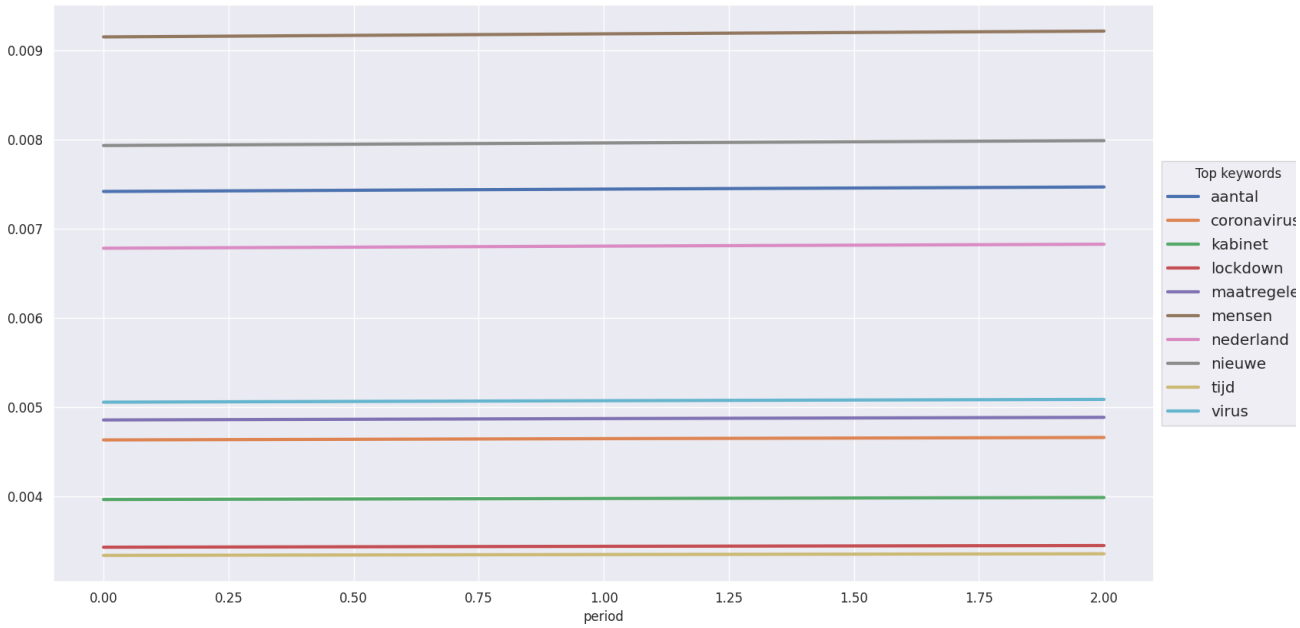


Figure 7: First topic

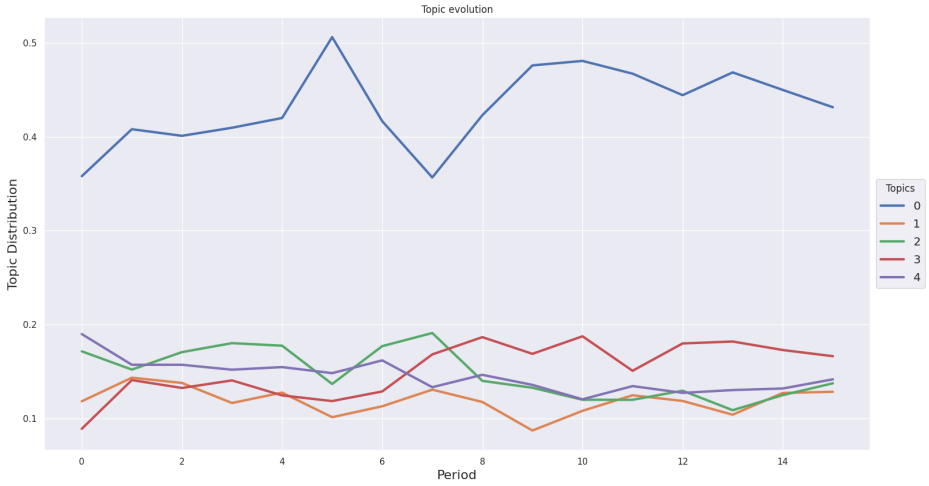


Figure 8: Topics plotted over time

3.5 Experiment 3: Bi-gram network graph

Based on the LDA experiment, we noticed that there were many individual terms within the topics and salient terms that did not have a specific meaning. For example the term "Tweede" (second) and the term "Kamer" (room) do not give us much insight when we analyse the results, however

when we take those two terms and read them as one we get "Tweede Kamer" (the government of the Netherlands) suddenly the whole meaning changes and becomes more valuable.

A pair like this is called a bi-gram and it is another way to look at the differences and underlying subtopics between the newspapers. Bi-grams are pairs of words that appear together often. When we look for bi-grams in a text analysis, it will help with understanding the meaning of the text. For example, the words *artificial* and *intelligence* have different meanings individually, however, when we see it as one term *artificial intelligence*, it gets a different meaning. So, in able to get a better meaning of our text analysis, we will use bi-grams. [20]

To get a bi-gram network graph, we first have to gather the bi-grams in our data set. We do this by using a function called `ngrams()` from python's library *nltk* [27]. First we tokenized every word in the articles for each newspaper. So per newspaper independently, we got a list of all the unique words, then we created all the bi-grams possible by linking one word to another and finally we counted for each of those bi-grams how often they occurred in the articles per newspaper. For the graph, we took the 100 most frequent bi-grams.

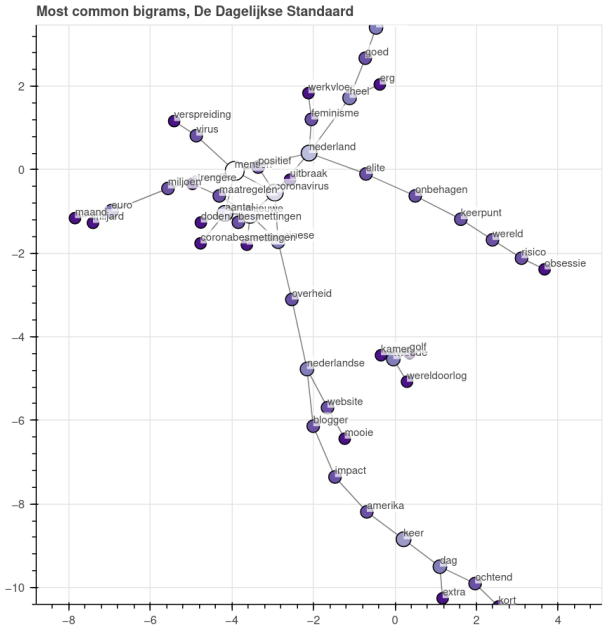
Now that we have our bi-grams, we create a graph using python's library *Bokeh*. The visualisation methods work by plotting the data into a Cartesian 2D field, the x- and y-axes do not contain any meaning in this case, they are only used to create a space in which the graph can be projected. In our generated graph, the nodes contain one part of a bi-gram and each vertex leads to the second part of the bi-gram. So when a node has three vertices, it means that there are three different bi-grams which all have this one node in common. Figures 9 and 10 on the next page show these bi-gram graphs for all the newspapers. A side note is that these graphs are all zoomed in, to view the full bi-gram graphs, have a look at the appendix.

Each cluster forms a topic, in figure 9d there are four clusters visible where it appears that the most left cluster is a topic that contains advertorial bi-grams that do not have any relation to corona.

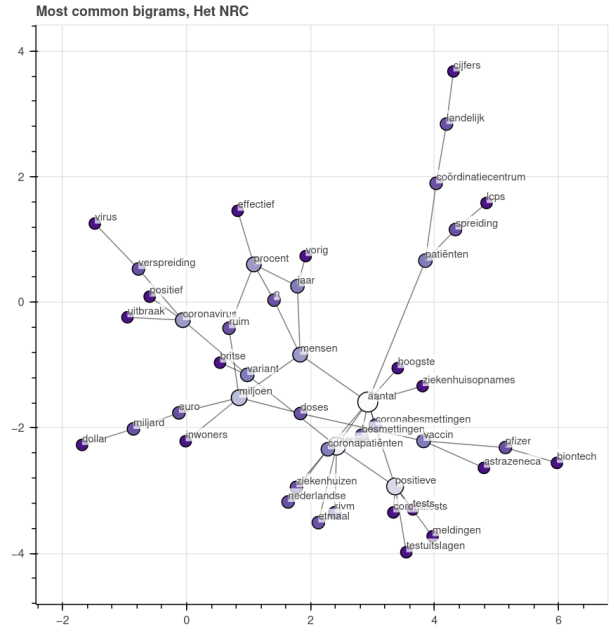
In the figures, we see that each graph is distributed differently, this distribution tells us something about the fragmentation of the topics. Figure 9a shows that the Dagelijkse Standaard has three trails which are about corona unrelated topics. The trail towards the bottom shows a subtopic about America, probably because of the elections, and the right-most trail seems to form a cluster of opinions about the coronavirus.

In Figure 9b we see three main branches and they all have corona as main topic. It seems that most-right topic is about the vaccinations, the most-left branch seems to be about the seriousness of the coronavirus, and the top branch seems to be about factual numbers of the coronavirus. In Figure 9c we see mostly one whole topic, a branch towards the bottom is leading to a subtopic about China and Wuhan, the city where the corona virus began. We also see a small cluster in the bottom which is about the minister of the public health of the Netherlands. In Figure 9d we see a couple of clusters, the cluster on the top left contains advertorial terms that are used within the articles which we apparently did not filter out well. Then, the cluster in the middle seems to contain general coronavirus terms. It differs from the cluster on the right because that cluster seems to contain more detailed coronavirus terms like different covid-19 variants and mutations, but also about the vaccines. Then, finally in figure 10 we see three clusters of which two are about advertorial terms which were not filtered out properly. The right-bottom cluster is about the coronavirus.

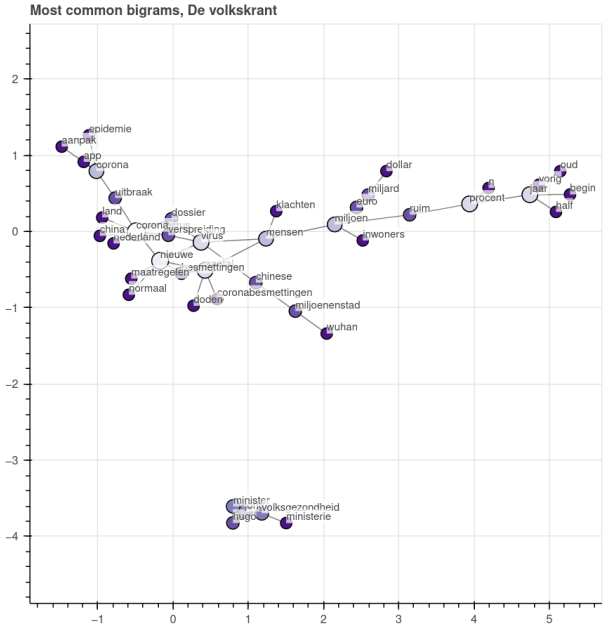
To see the complete graphs, please have a look at the Appendix in section 6.



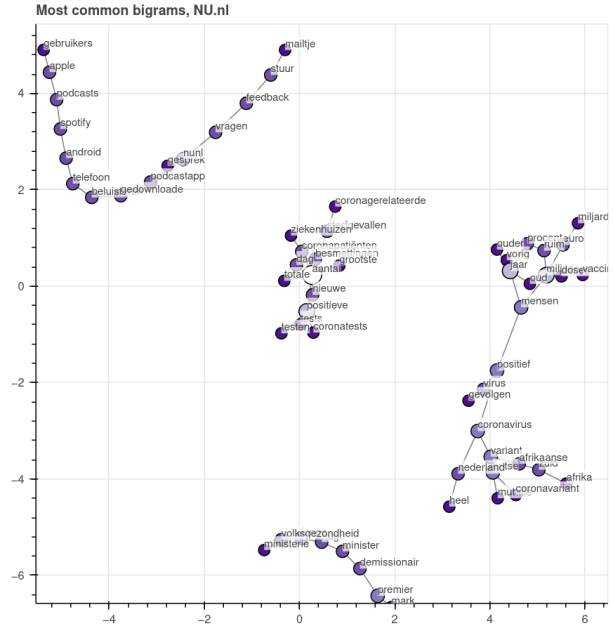
(a) De Dagelijkse Standaard



(b) Het NRC



(c) De Volkskrant



(d) NU.nl

Figure 9:
Top 100 most common Bi-gram graphs for each newspaper

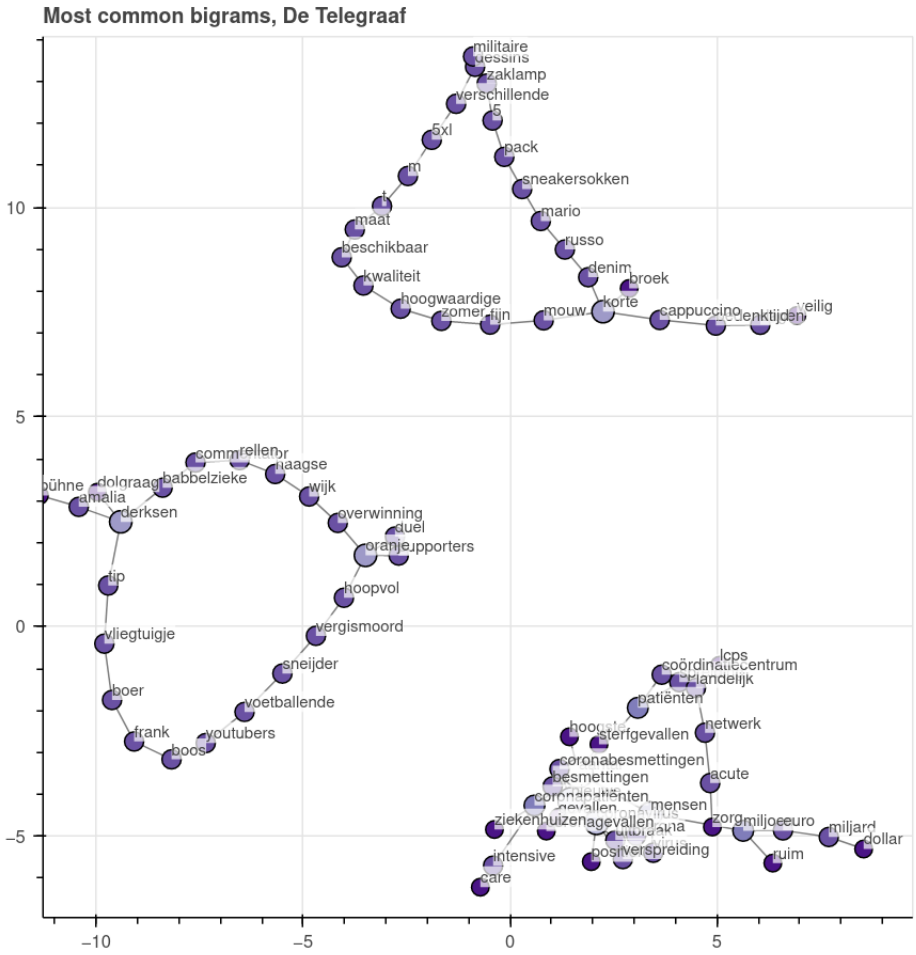


Figure 10: Top 100 most common bi-grams, De Telegraaf

4 Discussion

In this section, we will discuss our research findings.

4.1 General limitations

In this subsection, we will discuss the general limitiations we encountered when we were doing our analysis.

4.1.1 Data set size

Our data set is a bit limited in the way that it is not very large, most newspapers have at most 2000 articles and even some of these articles are short. Such data analysis works a lot better when the data set is larger. Another limitation which contributed to our small data set, is that we only searched for the word "Corona" within the titles of the articles. This term is the most general term for the coronavirus so it would deliver most of the articles. However, we could have used more key terms to get more subtopics, for example the term "Vaccin" would deliver more

articles about vaccination which is related to corona but it is a subtopic. More terms that we could have used are “Covid”, “Mondkapje” (mouth mask), “Maatregelen” (measures), “Persconferentie” (press conference), and “Variant”. We believe that these keyterms would have delivered more differentiation in the articles which would have led to a higher probability of discovering subtopics

4.1.2 Time limitation

Another limitation is the time frame of this research. The corona virus did not exist before the year 2020 in the Netherlands but it is still present in 2021 and will be in 2022. Because of the moment we started our data acquisition, we could not get all the necessary information to create this research. So this research could be viewed as the analysis of the beginning of the coronavirus.

4.1.3 Bias

To do a research like this, where the purpose is to find underlying meanings behind the articles, it is important to stay objective and neutral. However, we know all of these newspapers and their stigma/political views so when we did some of the research, we tried to stay neutral but also tried to conform our biases about the selected newspapers. It is a bit difficult to stay neutral when we know what to expect.

4.2 Interpretations

In this subsection, we will discuss our interpretations of the obtained results from our experiments.

4.2.1 Experiment 1

From the salient terms, we could not get great differentiation out of the terms between the newspapers, we believe that this is because the most frequent terms over all the newspapers will be overall the same since all the articles have the same topic in common. Though, we did discover that the salient terms of some newspapers used more adverbs than others which indicates a more subjective write style. Normally a couple more adverbs would not be linked to a more subjective write style but now we are dealing with the top 30 so each adverb means that that specific adverb was used so much that it is important for the write style. From the LDA experiment itself, we could again, not differentiate the terms, we believe this is because of the same reason as with the salient terms; the words are too commonly used within all the newspapers.

We can conclude that LDA with Bag-Of-Words unfortunately did not work well with our data set. The main topic within the newspapers is the same and this causes the experiment to get less differentiation than when we would do this experiment with totally different topics.

It would work better if the data set would be larger but also if we would filter out the commonly used words. We could have done this by removing each word that passes a threshold of frequency within each news medium. For example, the word “mensen” which means “people” is used frequently within each news medium, it is likely that each news medium uses this word at least 2000 times. We would filter out words like these which would have left the data set with more meaningful words that define the write type of the newspaper better.

This solution is actually what TF-IDF does so we experimented with this as well. Within a newspaper, the topics are more defined but still closely related. We will now give a detailed interpretation about the obtained TF-IDF tables 4 till 7. Since the topics are not greatly distinguishable, it becomes a bit difficult to interpret these results. We have to use our own interpretation which could possibly deliver biased conclusions, therefore, we will try to keep it objective. In order to answer our research question, we have to look at words that may imply an underlying message or a double meaning.

A good way to interpret these obtained results, is to do a sentiment analysis. We interpret the results of the LDA tables by looking at the frequency of negative and positive sentiment. We used a lexicon for this which we derived from Kaggle [4]. We got the following results:

- Volkskrant got 7 negative words which are: 'crisis', 'zorg', 'zorgen', 'uitbraak', 'klachten', 'virus', and 'besmet'.
  And 9 positive words which are 'wel', 'goed', 'eerste', 'open', 'ruim', 'premier', 'positief', and 'week', 'advies';

- Dagelijkse Standaard got 15 negative words which are: 'gewoon', 'zorgen', 'dupe', 'virus', 'besmet', 'volkomen', 'inconsistent', 'blijft', 'risico', 'zinken', 'weken', 'crisis', 'kritiek', 'doden', and 'partij'.
  And 11 positive words which are: 'echt', 'positief', 'week', 'gretig', 'werk', 'willen', 'snel', 'duurzame', 'kans', 'wel', and 'premier';

- NU.nl got 12 negative words which are: 'prikken', 'maken', 'zeven', 'daling', 'zorg', 'weken', 'virus', 'overleden', 'besmet', 'klachten', 'woordvoerder', and 'prik'.
  And 10 positive words which are 'eerste', 'dagen', 'goed', 'wel', 'advies', 'positief', 'premier', 'open', 'krijgen', and 'week';

- NRC got 9 negative words which are: 'overleden', 'ontbreken', 'virus', 'daling', 'weken', 'vertraging', 'zorg', 'liggen', and 'prik'.
  And 8 positive words which are: 'eerste', 'premier', 'week', 'dagen', 'ruim', 'hoogste', 'wel', and 'open'.

Since this sentiment analysis is only done on the key terms, we cannot make statements or conclusions about the whole articles, therefore when we make statements or conclusions in the next paragraph, we mean to speak about the key terms.

From this sentiment analysis, it becomes apparent that the Dagelijkse Standaard has used the most negative words within their articles assuming we only consider the salient terms. The Volkskrant uses the least amount of negative words, only 7 of which most of them are words that are inevitable when we talk about the coronavirus while the Dagelijkse Standaard also uses negative words apart from the coronavirus. The Dagelijkse Standaard is the newspaper that uses the mosed positive words, although it does not differ much from NU.nl which used 1 word less. We notice that the NRC and the Volkskrant still use few sentiment words within the positive sentiment analysis, from this we can conclude that the NRC and the Volkskrant write their articles in a neutral sense while NU.nl and the Dagelijkse Standaard are more strongly opinionated since they use more sentiment in their articles.

4.2.2 Limitations experiment 1

We can not make hard conclusions based of these analyses because this sentiment analysis is done over the topics which are derived from LDA and we can never be certain about the sentiment lexicon since it is quite difficult to attach a certain sentiment to a word. For example, the word "Positief" which means positive is marked as a positive sentiment word but it could have a negative meaning within an article; "Positief getest" which means a positive covid test. That would not be as positive.

4.2.3 Experiment 2

The first exploratory experiment that we did, was done manually to provide an insight of what we can expect when we do the dynamic topic modelling experiment. We can deduce from figures 3, 4, 5 and 13 that there are certain terms we expect to change over time. An important remark within these graphs is that the terms are relative and are not related to the other terms. So the term "Wappie" is out of proportion since this term was found only four times in the whole data set within the Volkskrant. We did not search terms that pass a certain threshold but rather picked out some terms of which we expect them to change over time.

The figures 7 and figure 8 show the results of the dynamic topic modelling. Unfortunately, those results are not usable for our analysis. When we look at figure 7, which displays the terms within topic 1 over time, we see no dynamic change while there are terms that would change over time, for example the term "Lockdown" would definitely be more or less frequent over time. Then, in figure 8 we see some dynamic change. We see that the first topic, topic 0, changes a lot but there are several reasons why we cannot interpret these results:

- First, when we derived this topic, it did not became clear what this topic is exactly about. The terms in figure 7 do not form a word network which would point at a clear topic. So when this whole topic changes over time, we do not know what exactly is changing;
- Another reason why we cannot use these results is that the dynamic change is not big enough for us to interpret this change. It seems as if there is barely any change over the time for the topics;
- Lastly, the topics all sort of cluster together, this might has to do with the fact that the topics are not very distinguishable. Another feasible reason for this clustering is that the analysis do not work like we expect it to work.

We conclude that these results are not very interpretable and less useful for our research.

4.2.4 Limitations experiment 2

We believe that the greatest limitation is that we simply did not have a large time span. We only had data of 16 months where each month did not contain many articles. We think that our data was too small for this function to work.

4.2.5 Experiment 3

The word network graphs are a good way to discover if a newspaper writes their corona related articles with an underlying message behind them. This is because we can see more clearly how the clusters are forming since there is a connection between the words.

The word network graph of the Dagelijkse Standaard, Figure 9a contains some terms that are unusual to find in a corona related article. Terms like feminism, World War, and obsession. When we compare these results with the results of the first experiment, we notice that table 6 does not contain these unusual words but it does contain more adverbs which are a bit unusual to use in an objective written article like "Gestoorde" in topic 5 which means troubled/disturbed. We also notice that there are a bit more political terms in the LDA table compared to the word network graph.

When we look at the word network graph of the NRC in figure 9b we notice that unlike the other word network graphs, the NRC seems to focus on the vaccines since we see the different vaccine brands "Pfizer"/"Biontech" and "Astrazeneca". We notice this as well when we analyse the results of the first experiment in table 4. There we also notice the other vaccine brand "Janssen". Furthermore, we notice that the NRC word network graph does not contain much adverbs, the terms seem to be fairly neutral. We noticed this as well when we did the positive/negative sentiment analysis in the discussion of the first experiment, there we discovered that the NRC, together with the Volkskrant, contained the least amount of positive/negative words.

The other three newspapers, the Volkskrant, NU.nl, and the Telegraaf, seem to contain the same type of terms. Naturally they are all corona related but they do not stand out. When we compare these findings to the sentiment analysis in the discussion of the first experiment, we see that NU.nl scores a bit higher than the other newspapers. We notice that the clusters of the word network graph 9d of NU.nl are more but also smaller than those of the the other newspapers. This could mean that NU.nl discusses some aspects of the coronavirus more detailed than the Volkskrant and the Telegraaf, and when we look at the terms within the clusters, this speculation may be correct. NU.nl seems to discuss the corona variants more detailed than the other newspapers.

4.2.6 Limitations experiment 3

For our word network graphs, we simply took the 100 most common bi-grams. We did not experiment a lot with this number so we could have get better results with another amount. We did experiment with the 50 most common bi-grams and compared to the amount of 100, we can say that the current amount tells us more details which are needed for interpreting our results. A higher number would be more effective or maybe not, we did not test this and that could be a limitation.

Another remark that is not necessarily a limitation, is that there are many ways to do an analysis like in experiment 3. A paper by Kats et al [14] about distinguishing commercial content from editorial content in news, also implement a type of bi-gram graph. Rather than pairing each word with another and counting the occurrence of each bi-gram, they calculated the number of times where a word W1 appeared in the same sentence with another word W2. If the frequency passed a certain threshold, they created a link between the nodes and displayed it in a graph. This is an interesting approach and if we would have done this, we could have compared it to our bi-gram graph and analyse the differences. This also shows us that there are many ways to do an analysis based on the coherence of two words that are used frequently in a sentence.

4.3 Future work

Since the coronavirus is an ongoing development, there could be done further experiments and research. If we would to continue on this research, there are several aspects that we would add or change.

4.3.1 Data set

In order to do text analysis, the data set to work with has to be of high value. For further work we would do the following things to achieve this:

- Data set: we would gather more corona related news articles to increase our data set and we would search for more key terms like we listed in the limitations. By doing this, the research would automatically be improved already. We would also search the key terms in both the titles as well as the text body of the articles. Another improvement within the data set is to gather the articles over a longer period of time;
- Clean data: we already noticed that we did not properly clean up our data as expected when we saw the sponsored texts in experiment 3 within the results of De Telegraaf. We would be more precise by cleaning this up by filtering this out manually. Another improvement on cleaning up our data, is to add a stemmer because in experiment 1, we see the same term appear often but in other forms. By adding a stemmer we would end up with the same term only once which would create more space for other, more meaningful, words to appear in our analysis;
- Word type: In experiment 1, we filtered out most of the word types which left us with nouns and adverbs. For our future work we would experiment more with this to get the best possible results.

4.3.2 TF-IDF

In the first experiment, we already saw that applying TF-IDF had a big improvement on analysing the topics within and between the newspapers. We would probably get more out of this if we would have applied TF-IDF on our third experiment as well. TF-IDF tries to filter the meaningless terms that are already often used in each newspaper so by applying that, we would definitely get better results. We would also try to create the TF-IDF function ourselves in experiment 1 instead of using the out of the box function, so that we can tweak it, for example by comparing time periods. By creating the function our own, we could get a better understanding of how it is applied on our data set and we would get a better insight on how we could improve this function.

4.3.3 Compare, then and now

Since the coronavirus is an ongoing development, we could do a similar research that focuses on the recent developments and compare the results with our results now. We believe that the second experiment, about the dynamic topic modelling might also get better results if we would do the research with all the data up until the most recent time possible. We would also try to get more subtopics by including articles that are related to corona indirectly.

For future work, we would try to make an analysis based on time as we tried in experiment 2. We would compare the results from January 2020 until April 2021 with the results from May 2021 until January 2022. The pandemic has had many developments over these time frames so we would for example compare the words which are relatively the most frequent and see how this frequency would change over time. It is obvious for example that the term "Wuhan" used to be relatively very frequent in the beginning of the pandemic but soon this frequency would decrease. This was also visible in graph 3 in experiment 2. We believe that by doing such experiments based on time, we would achieve insightful information on how a topic changes and therefore exposing underlying subtopics.

4.3.4 Visualisation

With text analysis, it is of great importance that we turn our results into some sort of visualisation so we can interpret the results better. With experiment 1 it was a bit difficult to interpret the results since there were a lot of terms that all overlapped a bit within the newspapers. We would do a type of visualisation where at least two of the newspapers would show on so we can really compare them.

5 Conclusions

The purpose of this thesis was to discover whether artificial intelligence could help us in discovering an underlying theme or topic within our selected newspapers and to interpret these underlying topics if they could be found. To answer those questions, we did three experiments, all of them were a form of topic extraction. The first experiment, for which we used LDA, resulted in 5 topic clusters for each newspaper. We applied TF-IDF on the corpus to get more meaningful results. From that experiment, we discovered that the Dagelijkse Standaard and NU.nl scored a higher in sentiment analysis than the other newspapers.

The second experiment was dynamic topic modelling in which we tried to discover change in terms and topics over time but unfortunately this experiment did not result in interpretive results. The third experiment was the creation of word network graphs with bi-grams. These graphs showed us fairly detailed information which helped us to interpret the results better. Here we discovered that the Dagelijkse Standaard used some unusual terms and that NU.nl discussed their topics more in detail.

By combining our results, we believe that artificial intelligence can be used to discover an underlying topic or theme within text analysis. It does require knowledge of term sentiment but it is doable. We can conclude that the NRC stays the most neutral when writing their articles. We can not conclude whether the other newspapers have an underlying topic because there where no indicating terms found.

We believe however, that the underlying theme of the Dagelijkse Standaard, is that it has a negative attitude towards the coronavirus. A reason for this is that it scored the most amount of negative terms in the sentiment analysis and another reason is that within experiment 1, the political party Forum voor Democratie (FVD) and its leader Thierry Baudet, are mentioned often and it is known that Baudet is against the corona measures.

We believe that with some additions and improvements in future work, we could get better and more specific results. This is easily achievable by getting a larger and more diverse data set and

cleaning it up more properly, but also by doing more visualisation methods and by applying more data preparation such as TF-IDF to our data set. Nevertheless, we truly believe that artificial intelligence can be used to discover underlying topics as this was clearly visible with the bi-gram experiments, where each cluster formed a subtopic. Therefore we see this research as a success.

References

[1] Christoph van den Belt. "De krant, een cultuurgeschiedenis". In: *Historiek* (2019).

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation". In: *Machine Learning Research* 3 (2003).

[3] Pew Research Center. *News Media and Political Attitudes in the Netherlands*. 2017. URL: https://www.pewresearch.org/global/fact-sheet/news-media-and-political-attitudes-in-netherlands/.

[4] Yanqing Chen and Steven Skiena. "Building Sentiment Lexicons for All Major Languages". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Baltimore, Maryland: Association for Computational Linguistics, June 2014, pp. 383–389. DOI: 10.3115/v1/P14-2063. URL: https://aclanthology.org/P14-2063.

[5] Jason Chuang, Christopher D. Manning, and Jeffrey Heer. "Termite: Visualization Techniques for Assessing Textual Topic Models". In: *Proceedings of the International Working Conference on Advanced Visual Interfaces*. AVI '12. Capri Island, Italy: Association for Computing Machinery, 2012, pp. 74–77. ISBN: 9781450312875. DOI: 10.1145/2254556.2254572. URL: https://doi.org/10.1145/2254556.2254572.

[6] De Europese Commissie. *Aanpak van desinformatie over het coronavirus*. 2020. URL: https://ec.europa.eu/info/live-work-travel-eu/coronavirus-response/fighting-disinformation/tackling-coronavirus-disinformation_nl.

[7] Kurt Englmeier. "The Role of Text Mining in Mitigating the Threats from Fake News and Misinformation in Times of Corona". In: *Procedia Computer Science* 181 (2021). CENTERIS 2020 - International Conference on ENTERprise Information Systems / ProjMAN 2020 - International Conference on Project MANagement / HCist 2020 - International Conference on Health and Social Care Information Systems and Technologies 2020, CENTERIS/ProjMAN/HCist 2020, pp. 149–156. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2021.01.115. URL: https://www.sciencedirect.com/science/article/pii/S1877050921001538.

[8] Afham Fardeen. *Tutorial on Spacy Part of Speech (POS) Tagging*. 2021. URL: https://machinelearningknowledge.ai/tutorial-on-spacy-part-of-speech-pos-tagging/.

[9] Tamanna Hossain et al. "COVIDLies: Detecting COVID-19 Misinformation on Social Media". In: *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. Online: Association for Computational Linguistics, Dec. 2020. DOI: 10.18653/v1/2020.nlpcovid19-2.11. URL: https://aclanthology.org/2020.nlpcovid19-2.11.

[10] MonkeyLearn Inc. *Topic Analysis: The ultimate guide*. URL: https://monkeylearn.com/topic-analysis/.

[11] Nederlandse Vereniging van Journalisten. *Code voor de journalistiek, door het Nederlands Genootschap van Hoofdredacteuren*. 2008. URL: https://www.nvj.nl/themas/ethiek/ethiek/code-journalistiek-nederlands-genootschap-hoofdredacteuren-2008.

[12] Rajesh Prabakhar Kaila and Krishna Prasad. "Informational Flow on Twitter – Corona Virus Outbreak – Topic Modelling Approach". In: International Journal of Advanced Research in Engineering and Technology (IJARET), 11 (3), 2020, pp 128-134. Available at SSRN: https://ssrn.com/abstract=3565169 (2020).

[13] John Kastner, Hanan Samet, and Hong Wei. *NewsStand CoronaViz: A Map Query Interface for Spatio-Temporal and Spatio-Textual Monitoring of Disease Spread*. 2020. arXiv: 2003.00107 [cs.IR].

[14] Timo Kats, Peter van der Putten, and Jasper Schelling. "Distinguishing Commercial from Editorial Content in News." In: Preproceedings BNAIC/Benelearn 2021 (Luxembourg). 2021.

[15] Joost N. Kok et al. "Definition, Trends, Techniques. In Knowledge for Sustainable Development: an Insight into the Encyclopedia of Life Support Systems". In: *UNESCO Publishing-EOLSS Publishers, Oxford, UK* 1 (2002).

[16] Konrad Krawczyk et al. "Quantifying Online News Media Coverage of the COVID-19 Pandemic: Text Mining Study and Resource". In: *J Med Internet Res* 23.6 (June 2021), e28253. ISSN: 1438-8871. DOI: 10.2196/28253. URL: https://doi.org/10.2196/28253.

[17] Marco Meyer, Mark Alfano, and Boudewijn de Bruin. "Epistemic Vice Predicts Acceptance of COVID-19 Misinformation". In: Available at SSRN: https://ssrn.com/abstract=3644356 or http://dx.doi.org/10.2139/ssrn.3644356 (2020).

[18] Loic Muhirwa. *Topic Modelling and Dynamic Topic Modelling : A technical review*. 2021. URL: https://www.statcan.gc.ca/en/data-science/network/topic-modelling.

[19] Enda Esyudha Pratama and Rizqia Lestika Atmi. "A Text Mining Implementation Based on Twitter Data to Analyse Information Regarding Corona Virus in Indonesia". In: *Journal of Computers for Society* (2020).

[20] *Python text processing*. 2021. URL: https://www.tutorialspoint.com/python_text_processing/python_bigrams.htm.

[21] *Raad van Europa: autoriteiten zaaiden paniek over Mexicaanse griep*. 2010. URL: https://www.europa-nu.nl/id/vidwefmc9ezd/nieuws/raad_van_europa_.

[22] D. Raeijmaekers. "Little debate". In: *Ideological media pluralism and the transition from a pillarized to a commercialized newspaper landscape (Flanders, 1960-2014)* (2018).

[23] Mohit Rathore. *Dynamic Topic Models Tutorial*. URL: https://markroxor.github.io/gensim/static/notebooks/ldaseqmodel.html.

[24] Sandeep Reddy et al. "Use and validation of text mining and cluster algorithms to derive insights from Corona Virus Disease-2019 (COVID-19) medical literature". In: *Computer Methods and Programs in Biomedicine Update* 1 (2021), p. 100010. ISSN: 2666-9900. DOI: https://doi.org/10.1016/j.cmpbup.2021.100010. URL: https://www.sciencedirect.com/science/article/pii/S2666990021000094.

[25] Radim Rehurek. *TF-IDF model*. 2021. URL: https://radimrehurek.com/gensim/models/tfidfmodel.html.

[26] Julia Silge and David Robinson. *Text mining with R*. O'Reilly Media Inc., 2017.

[27] *Source code for nltk.model.ngram*. 2015. URL: https://www.nltk.org/_modules/nltk/model/ngram.html.

[28] Bruno Stecanella. *Understanding TF-ID: A Simple Introduction*. 2019. URL: https://monkeylearn.com/blog/what-is-tf-idf/.

[29] Lucy Lu Wang et al. "CORD-19: The Covid-19 Open Research Dataset". In: *CoRR* abs/2004.10706 (2020). arXiv: 2004.10706. URL: https://arxiv.org/abs/2004.10706.

[30] C. S. Wyatt. *Guide for Writers: adjectives and adverbs*. 2021. URL: https://www.tameri.com/edit/adjadv.html.

6 Appendix

Here there are figures which are too large to put between the main text. To look more closely at the bi-gram word graphs, please look at this link: Bi-gram wordgraphs and click on *Download* to see the html contents.

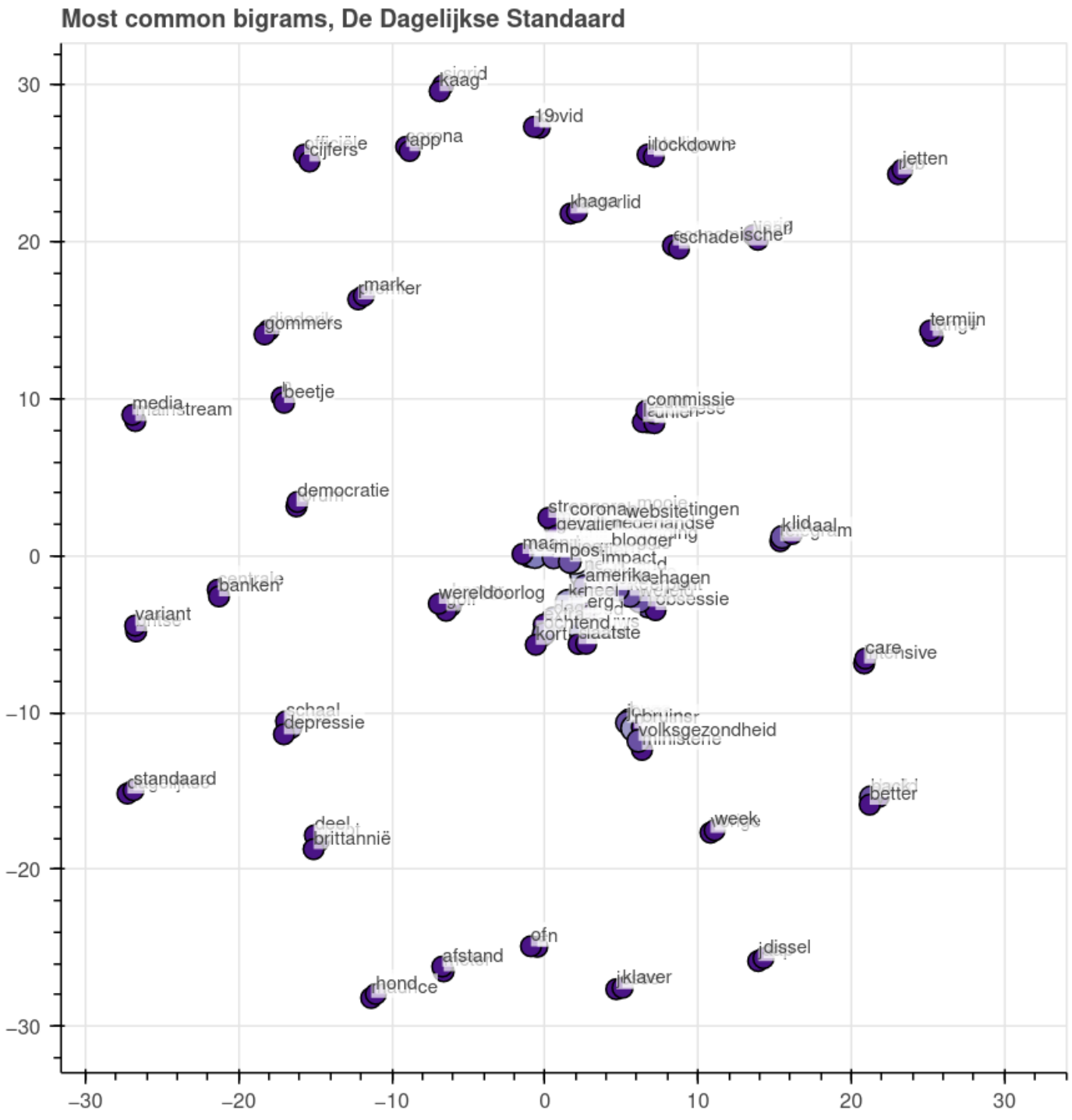


Figure 11: Word graph of the Dagelijkse Standaard

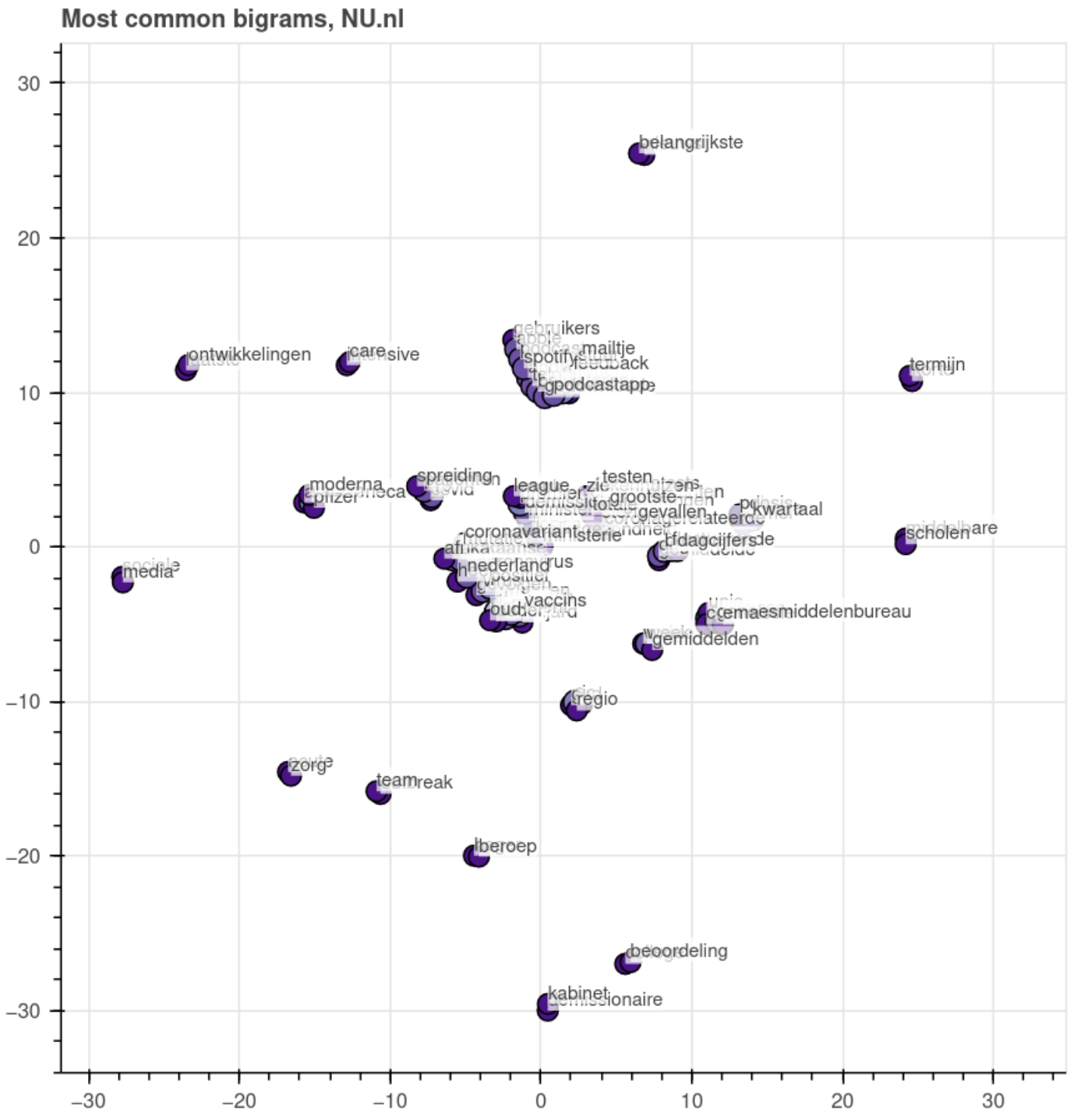


Figure 12: Word graph of NU.nl

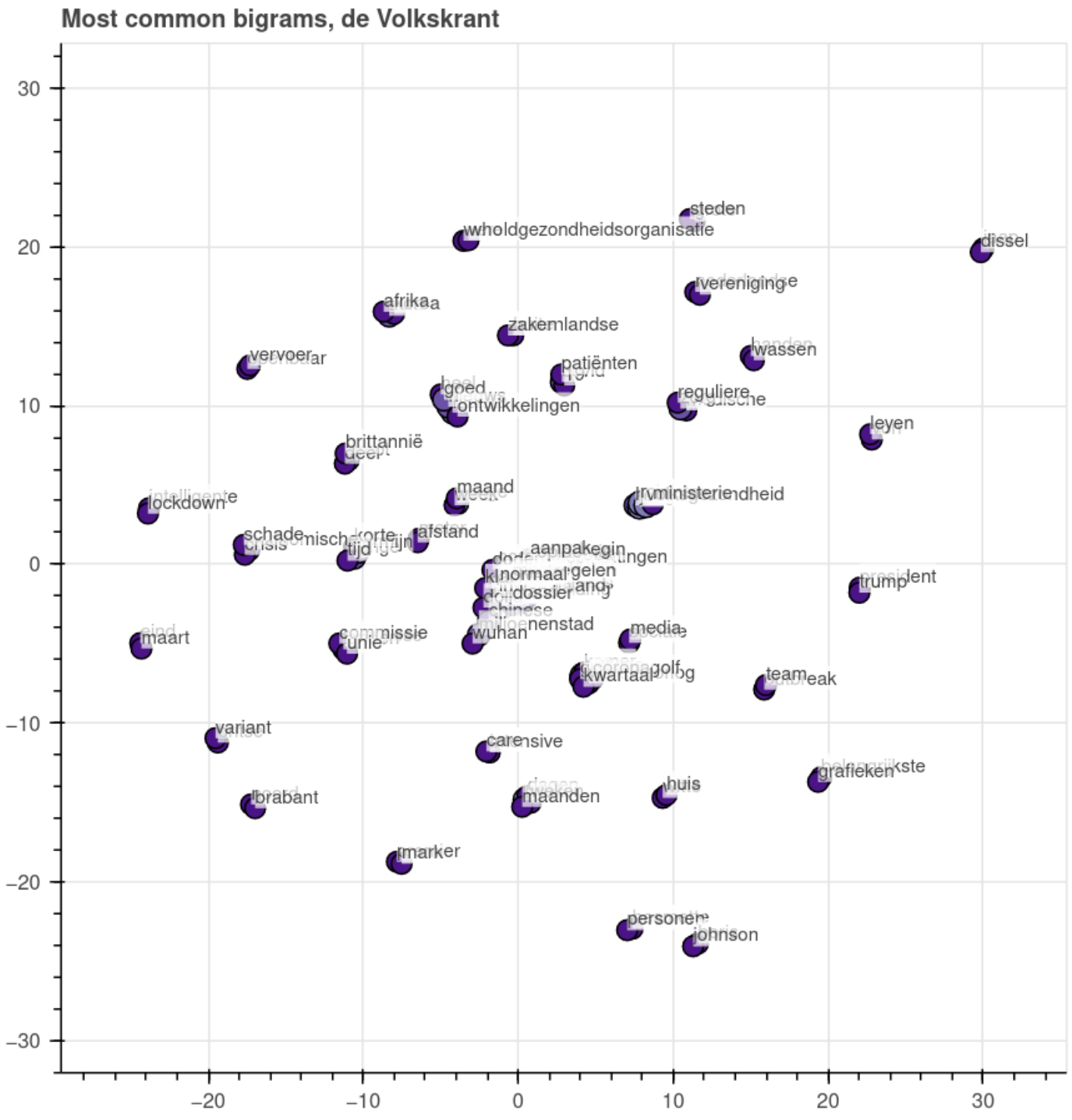


Figure 13: Word graph of the Volkskrant

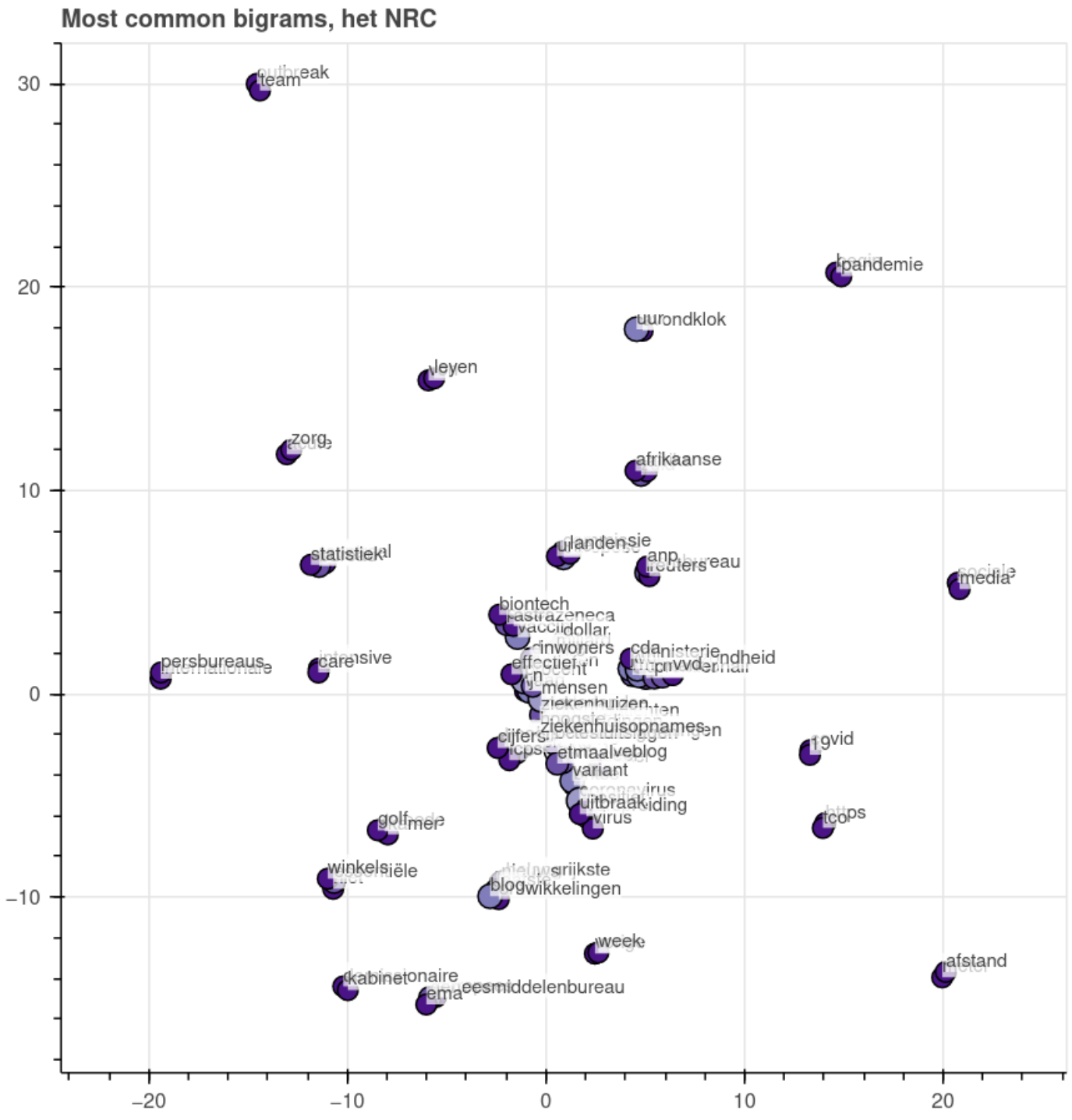


Figure 14: Word graph of the NRC

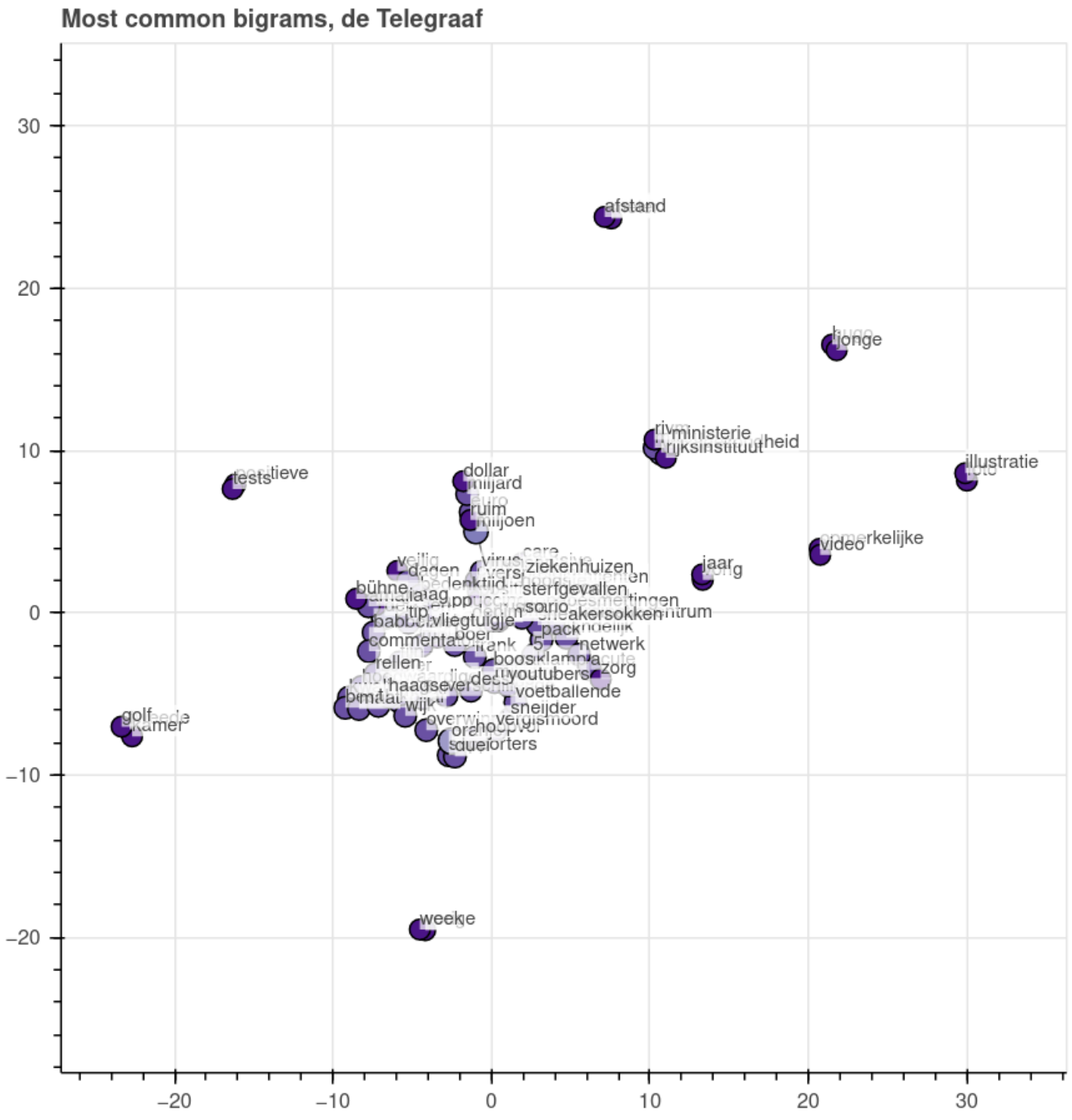


Figure 15: Word graph of the Telegraaf