University of Leipzig, Germany
Faculty of Mathematics and Computer Science
Department of Computational Humanities at the institute of Computer Science

Project Report

Stories of the Pandemic

Methods and Applications in Digital Humanities

April 1st, 2022

Student Names: Niklas Ehrlich, Oscar Kirchner, Irene Martin, Maria Schweren

Email: {ne70puzy, ok94cevi, im50diju, ue45ojap}@studserv.uni-leipzig.de

Program of Study: Digital Humanities / Data Science (Master of Science)

Supervisor: Dr. Andreas Niekler

Abstract

The occurrence of the COVID-19 pandemic at the beginning of 2020 caused great shock worldwide and has since been one of the most important topics in society. Consequently, no other topic is as present in the media as this one, since on the one hand there are many different opinions, but on the other hand the pandemic also covers very many different topics, such as health, economy, sports, politics and culture. In view of this fact, it is of great interest, especially in retrospect, which topics were in focus of the pandemic's coverage and how they changed over time. Consequently, the aim of this paper is to conduct a text analysis of the news coverage of the COVID 19 pandemic over the last two years. The focus is on different text analysis methods such as word frequencies, the TF/IDF measure, co-occurrence analyses and especially topic modeling. Furthermore, it will be investigated whether it is possible to visualize and narrow down the stories of the pandemic into a map of knowledge and what are the limitations of such a representation. The results of our investigations show that it is possible to uncover the relevant topics of the COVID-19 pandemic using different methods. Counting word frequencies and TF/IDF calculations still represent very simple procedures, which mainly refer to single words only without relationship to other words or semantic connections. However, the creation of co-occurrence networks provides more information, especially regarding the co-occurrence of words in articles. Another approach is topic modeling, which creates a semantic context around different terms, so that topic groups with words from similar meaning contexts are formed. Using this method of text analysis, we also created a Map of Knowledge, which contains a high level of information regarding the topics of the COVID-19 pandemic, but also has its limitations especially in terms of temporal information. In summary, our work provides a comprehensive overview of the main topics in the news coverage of the pandemic and provides a good basis for further work. In future work, especially the combination of co-occurrence analysis with topic modeling would be of great interest.

Contents

1. Introduction

Since the beginning of the year 2020 there is one major topic existing in all kinds of media: the COVID-19 pandemic. This topic is also the subject of many scientific works, statistics and general publications. These provide insights into current developments as well as retrospective ones. From these publications and also from public political decisions, the media publish articles online and in print. Society opinions, debates and developments were also recorded in those articles. What was relevant at what time and how terms have developed can be determined from these texts. In view of the fact that the published articles consider different topics and opinions regarding the pandemic, it is interesting to examine them in more detail. Therefore, this research will explore the use of topic modelling and other text analysis methods to visualize or summarize the different stories of the pandemic in a certain text-corpus. The main data source will be consisting of a text corpus of various Guardian news articles on the COVID-19 pandemic over the last two years, provided in the so called "Guardian-API"[1]. This research study focuses specifically on three research questions: First, What topics regarding the COVID-19 pandemic appear in the Guardian news articles and how have they changed over a certain period of time? Second, Is it possible to visualize and narrow down the stories of the pandemic into a map of knowledge? And Third, What can such a representation do? What are its limitations and how can this representation contribute to enlightenment?. These questions are intended to provide more information in the field of research into the course of the pandemic regarding the change of topics.

[1] A more detailed description can be read in chapter 3 Methodology.

2. Related Work

As already described, some articles refer to scientific publications. We also take this as the basis of our research. However, we refer to works that are topic model based. Research about topic modeling on the topic of COVID-19 can be seen for example in Jing Xuan Koh [2022] of analyzing the topic loneliness in times of the pandemic. Therefor they analyzed about 4.500 Twitter feeds from between May to July 2020 and started the topic modeling with the data. In result, prevalence of the themes were examined over the time and across the number of followers from the twitter accounts. They had to struggle with a big amount of data, which were scrapped and organized for hierarchical Modeling and identifying overarching themes. In contrast to our work, there was also a focus on global differences. Since we only use articles from the British The Guardian in this work, we do not observe any location-dependent differences in development.

In addition, Niaz Mahmud Zafri [2021] conducted a content analysis of newspaper coverage of the COVID-19 pandemic in Bangladesh. Their objective was to identify the key topics related to the pandemic and their timeline of discussion in order to develop a framework for pandemic management based on their results. For this purpose, they calculated a Latent Dirichlet Allocation topic model with twelve topics and presented their results in wordclouds on the one hand, and on the other hand they examined the occurrence of the twelve topics over time. However, our goal in this work, is to create a map of knowledge from the results of the topic modeling, which both represents the different topics and the proportions of the topics. Furthermore, we apply other text analysis methods like co-occurrence analysis in the context of our investigations. In order to show the impact on scientific papers, Ahamed Sabber [2020] also pursue a strategy of text mining in health-magazine articles that shows the necessity of using different methods when exploring big corpus objects, such as topic modelling in the classic way (using TF/IDF and LDA) and the use of Natural Language Processing. Their results are depicted in multiple graphs and intend to open up new approaches for research-questions regarding COVID-19. As this paper was one of the first studies in this order of magnitude on this subject, we were able to use some of the ideas like the consideration of the relevant keywords.

3. Methodology

The following chapter outlines the used methodology for this project and describes the information retrieval, corpus preprocessing and usage of the different methods of text analysis for achieving a resulting map of knowledge. Our Programming tasks for the experiment are based on the Programming language R[2]. It is best suitable for this project because it is a language for statistical computing and graphics. All of the achieved R-Scripts and source code, depicting these methods can be found in our Git-repository[3].

3.1. Map of knowledge

The overall goal of this work is to generate knowledge out of big data collected from the Guardian newspaper API. The retrieval of this amount of data in form of text-data and the knowledge generation trough data-retrieval and text-processing methods will be described in the following chapters of this work. When generating knowledge, an important last step is to analyze and visualize this gained knowledge in an interpretable and comprehensible form through methods like plotted graphs or network structures Alberto J Canas [2005]. A so called map of knowledge can be a simple and highly graspable, visual solution to representing the gained knowledge of the "stories of the pandemic" and of this work. It's a visualization technique for collected and correlating data and the generated knowledge from the data and in structure comparable to a so called mind map. This maps are very popular in all branches of knowledge generation and collection Yu-Cheng Lin [2006], for example in economic studies or tourism, but also in computational studies and humanities and there as also in the digital humanities Dalia E. Varanka [2018]. Software like mind manager is used as a common solution to organizing knowledge in big knowledge maps which also resemble network-structures between mapped topics Corporation [2022]. Different kinds of knowledge can be mapped,

[2] Further details can be found on this website: https://www.r-project.org/about.html

[3] Git-repository: https://github.com/1r3n3/StoriesOfThePandemicProject.git

in this case the map of knowledge will consist of the topic knowledge gained by using topic modelling (chapter 3.7). It should show which topics where the most popular regarding the COVID-19 pandemic in the guardian corpus of the years 2019 to 2022 and their correlations to each other. The visualization of a topic network with under-topics and common topics that seem to occur together anyway, but are especially mentioned together with pandemic-relevant topics is possible. So if COVID-19 topics occur together, do they also show correlation of their most frequent terms? Can there be a visualization of statistical significance regarding word frequencies, TF/IDF and topic models?

3.2. Data retrieval

As the base for this experiment and for collecting data, the Guardian API for scraping Guardian news articles was used.[4] Although Sars-Cov-Virus is not a new phenomenon at all, this work wants to focus on the current situation and not on situations like the influenza (H1N1) pandemic of 1919 which often is taken as a similar pandemic event and compared to Sars-Cov-2, though it's another strain of virus and belongs to another family of viruses Neumann [2019]. For this purpose, a period of time was defined, which we take as the basis for our investigation Ahamed Sabber [2020]. The pandemic started in December of 2019, and we decided to start with the articles from this date, to get a whole view of the international spreading of pandemic. We planed scraping the articles till the end of February 2022 but then decided to postpone the deadline to the 15. of March in 2022, because we waited for the release of the inactivated vaccine Valneva [2022], that was supposed to be released in February but took longer than expected. As part of the data retrieval process, all articles from the defined time period were scraped that were related to the following terms: Covid, COVID-19, Corona, Coronavirus, and SARS-CoV-2. The result of scraping were about 147.300 articles that are related to the COVID-19-Pandemic and therefore relevant for the corpus we wanted to create. Since a

[4] Further details can be found on this website: https://open-platform.theguardian.com/explore/

large amount of data was available due to the very large time period chosen, we wanted to remove some articles to save computational time on the one hand and and on the other hand to identify possibly less relevant articles on the COVID-19 pandemic. For this reason, we examined the title and text of the articles and counted the frequency of occurrence of the strings "Corona," "Covid," or "SARS-Cov-2.

The result is shown in Figure 1, where the number of articles is plotted against the corresponding frequency of the three strings. Furthermore, there were, for example, also some articles that did not even contain text but comics and were therefore not useful for our text corpus. From these findings, we finally decided on the following requirements that we had on the resulting corpus: Articles in which either the strings "Corona", "Covid" or "SARS-Cov-2" did not occur once, or in which there was no text in general, have been excluded from the corpus.

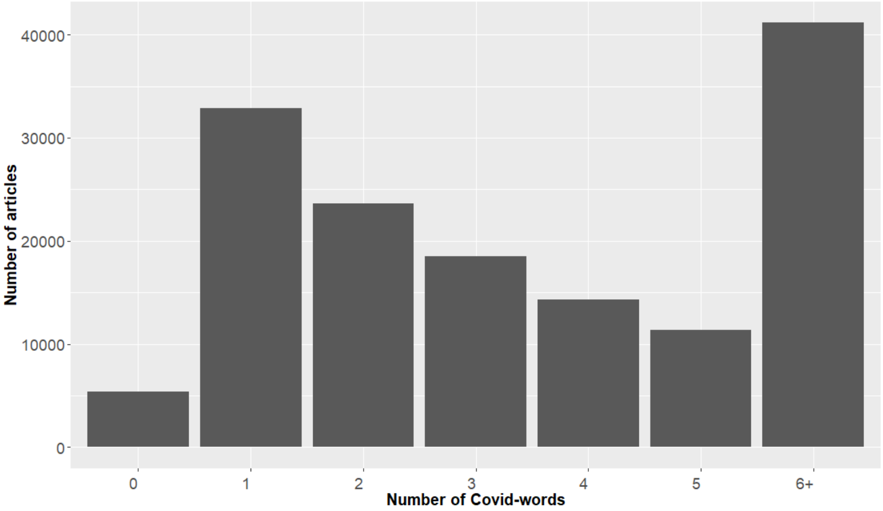

Figure 1: Frequency of Covid-words in Articles

After removing the irrelevant articles, the corpus contained a total of 141.753 articles that where used for our experiments. They consist of nearly 10.000.000 sentences, which gave us an average of 67 sentences per article. The whole corpus of articles contains 225.661.115 words, so that a sentence consists of an average of 24 words and each article consists round about 1592 words on average. All statistics about our data are written in Table 1.

Data	**Statistics**
Period of time	1. December 2019 - 15. March 2022
Number of all articles	147.296
Number of sorted out articles	5.543
Number of used articles	141.753
Number of sentences	9.538.568
Average sentences per article	67,29
Number of words	225.661.115
Average words per article	1591,93
Average words per sentence	23,66

Table 1: Data and Statistics

Figure 2 also provides a timeline showing the number of articles over the course of the pandemic. As can be seen, reporting on the Covid 19 pandemic did not really begin until February 2020 and then reached its maximum in the following three months. Since May 2020, the number of articles has been decreasing nearly continuously, although a lot is still being written about the pandemic to this day.

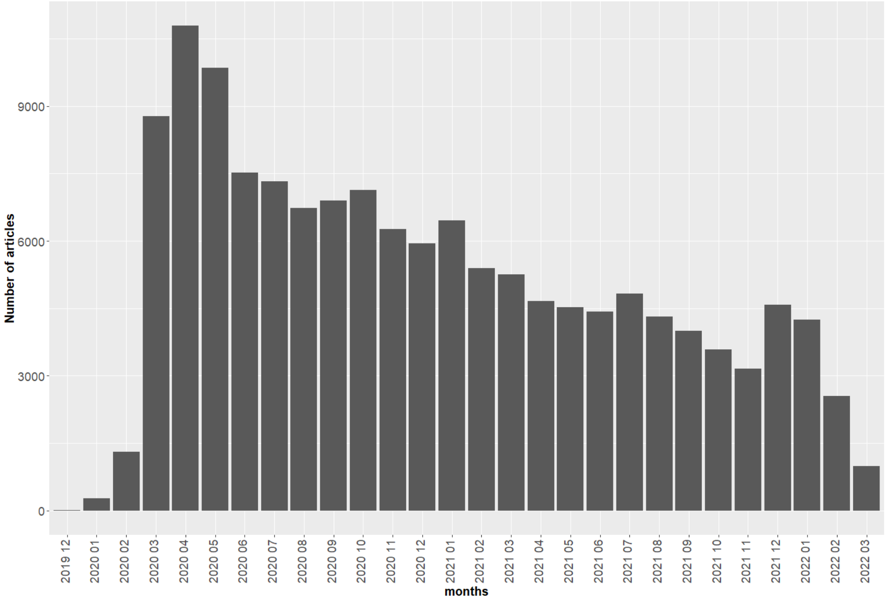

Figure 2: Number of articles over time

3.3. Corpus preprocessing

The data from all the scrapped articles has been saved as a csv (comma separated value) file. The file has the format: "doc_id"; "id"; "url"; "date"; "title"; "text". After loading the whole corpus into RStudio, the first thing to do was to reduce the database. As mentioned above, this involved identifying articles that are not written primarily for COVID-19 or do not contain any useful information. For the subsequent preprocessing steps, the data were first transfered into a corpus object of the `quanteda`-package. The preprocessing of the text data required several steps: tokenization, lowercase transformation, lemmatization, removal of non-word symbols (including punctuation) and English stop-words Schweinberger [2022c]. Based on the preprocessed corpus, a document term matrix (DTM) is created, but terms which occur in less than 1% of all articles are removed. Finally, the DTM includes 141751 rows and 5025 columns, which corresponds to the number of articles in the corpus and the number of different word forms of the vocabulary.

3.4. Absolute Word frequencies

Counting word frequencies is one of the simplest methods in text analysis Schweinberger [2022b]. It provides information about which words are used most frequently in a corpus and, consequently, may be of great importance for the content discussed in it. In our case, the frequency information of words in the corpus can be determined with the help of the DTM, since it corresponds in principle to a word frequency list. The DTM stores how often each word of the vocabulary occurs in the articles. Stop-words and other irrelevant tokens are not considered - as already described above. By summing up all of the column vectors of the DTM, we finally obtain the absolute frequency of a word in the entire corpus. A suitable representation for visualizing the most frequently occurring words in the corpus is a word cloud, which displays the words larger or smaller depending on their frequency. In the context of our research, we also want to measure the frequencies of selected terms over time. This allows us to determine from which point in time certain topics were discussed more frequently or less frequently, which in turn could indicate a certain event of the COVID-19 pandemic. For example, a more frequent use of the word "vaccination" over time could indicate the beginning of COVID-19 vaccinations. To obtain the time course of frequencies of selected words, we first reduced our DTM to those words that should be considered, and then aggregated their frequencies in single articles per month to get a meaningful representation of term frequencies over time.

3.5. TF/IDF

In the field of text analysis, especially when it comes to information retrieval from a big text corpus, one of the basic tasks can be found in the determination of not only the absolute importance of lemmas, but also on a way more relative frequency measure. This measure is known as TF/IDF (short for: term frequency/ inverse document frequency) and is used for spotting lemmas and showing the correlation between a lemma and its importance in the text object. When using the absolute wordcount, one can only measure the importance based on the word frequency of a lemma and not how the lemma can be ranked when it comes to contextual importance Taylor Arnold [2018]. Although in the context of this work the absolute word frequency is a good measure in showing certain distributions of words and their correlations in the guardian corpus, as shown in Figure 1, the use of TF/IDF can come in very handy when trying to determine a contextual importance between different articles over the years for example. The idea behind the TF/IDF measure is to weight terms according to their semantic contribution to a document. On the one hand, this means that a term is considered more informative the more frequently it occurs in a document (cf. term frequency). But also that with increasing occurrence of a term in a document, its information content for this document decreases - this is also called inverse document frequency. Finally, the TF/IDF weight for a term is the product of these two frequency measures. When working with TF/IDF on a corpus object, the preprocessing steps are common to other text retrieval methods, so indexing and detecting common stop words are the first steps in cleaning and reducing the corpus. When working with the programming language R-script, the packages `quanteda`, `dplyr`, `tidyverse` are of the most use for these tasks. The steps of preprocessing the corpus object are thoroughly shown in chapter 3.3. While the term frequencies can be calculated quite easily from the column sums of the DTM, the calculation of the inverse document frequency is more complicated. For each term, the total number of documents is divided by the number of documents in which the respective term occurs, and the logarithm is calculated from this. Finally, as mentioned before, the TF/IDF weights of the terms are the product of the two measures

3.6. Co-occurrence analysis

When analyzing textual data one must not only concentrate on absolute word frequencies, but also on the occurrence of words in certain contexts. But whilst the occurrence of a certain word alone might be relevant for a text or even a text corpus, for example the word thermae in an article about roman baths, the appearance of this certain word would relate to other certain words from the same contextual theme. In other words, if a word tends to occur in a certain context, the possibility is high, certain other words are co-occurring in the same context, having a similar meaning Scott Deerwester [1990]. This correlation is known as co-occurrence and the method using these co-occurrences, is named co-occurrence analysis. The relations between those words can be syntagmatic, meaning the significant words occur in a closed compound, so they are directly related to one anotherCharles Bally [2011]. The relation between co-occurring words could also be paradigmatic, determining words that occur together in a context, but not as close group, also having similar contextual meaning and grammatic Scott Deerwester [1990]. The computational methods used in this thesis are mostly relying on paradigmatic relations between the most relevant words describing the COVID-19 pandemic in the guardian text corpus. Our goal when using this method is to produce a network structure that shows the occurrences between the most significant words and directly relating words to them, as seen in Figure 8. The object of context in co-occurrence analysis is given as a sentence. The underlying theory is that sentences build a unit usable for semantic text analysis in a document Schweinberger [2022a]. Therefore a new corpus object can be build by splitting the old corpus up into divided sentences, which is crucial for the co-occurrence analysis, using the `quanteda` package for this purpose. Segmenting the sentences has to be one of the first steps even before the other preprocessing steps of the corpus so the word forms can remain intact and will not be reduced or split apart by filtering out certain stop words for example. Thereafter, the preprocessing of the corpus takes place so that the co-occurrence analysis is performed on a cleaned and reduced corpus object. The only difference to other used methods is the representation of the corpus object as a binary document term matrix, not a DTM. We chose to use these methods on the whole corpus object as well as on split up parts of said corpus, to show a change of the co-occurrence network over time. For the new corpus objects, following time periods were used: articles only from 12.2019 – 12.2020 or from 01.2021 – 03.2022. Accordingly, 3 different corpus objects were used for the co-occurrence analysis, one for the entire time period and one for each of the two sub-periods considered. When analyzing co-occurrence between words, a binary matrix is constructed out of these terms, named `binDTM` (binary Document Term Matrix) following the instructions by Schweinberger [2022a]. The important part therefore is only to find the binary match for the occurrence of a term in a document. When a term is matched within a sentence or a document a 1 will be written into the matrix, if not, a 0 will be written into itPerfetti [1998]. The `binDTM` builds the base on which the calculation of the co-occurrences is taking place. For plotting the resulting network graphs, the `igraph` package was used. The results are showing co-occurrences in relation to the term COVID-19 e.g. Figure 16 and the term vaccines Figure 18.

3.7. Topic modelling

While the other Methods that have already been referred to where rather frequentistic models or models that break a corpus down into smaller contextual pieces, topic modelling is a statistical representation model for topics found in a whole text corpusTaylor Arnold [2018]. The fundamental idea of topic models is to find hidden semantic structures in text data, in our case to discover specific topics in a selection of articles. Each topic is composed of different terms that occur together in documents and can therefore be assigned to the same semantic topic. For example, if you have a document that deals with a certain topic, you can assume that words related to this topic occur more often in this document than in documents with other main topics. With the help of text models it is possible to understand large collections of text data. For topic modeling in this thesis, a Latent Dirichlet Allocation Model (short: LDA Model) using Gibbs Sampling algorithm is computed based on the DTM Schweinberger [2022c]. The LDA approach is based on the assumption that a given document is composed of several topics. It allows, on the one hand, learning the topic distribution in each document and, on the other hand, characterizing the topics by assigning words to each topic. The most important parameter in computing an LDA model is the number of topics `K`. The goal is to find an optimal number of topics so that the resulting semantic contexts are neither too general (small K), nor too detailed and too difficult to interpret (big K). The best number of topics can be found in R using the package `ldatuning` and the function `FindTopicsNumber`. Several LDA models are trained with different values for K and their performance is evaluated using four metrics (Griffiths2004, CaoJuan2009, Arun2010, and Deveaud2014). The best number of topics shows low values for CaoJuan2009 and Arun2010 and high values for Griffith2004 and Deveaud2014. As shown in Figure 3 the optimal number of topics for this corpus resulted in being 21 topics. At first a bigger range of topics was chosen to work with, but while using the LDA-tuning a decrease of the optimum curve after 21 topics could be seen, with a optimum of 21 topics. With this number of topics, a LDA topic model was finally calculated, which tries to uncover hidden semantic structures in the corpus text data as best as possible. The results of topic modeling were then used to generate a map of knowledge and to plot topic distributions over time (see chapter 4.3).

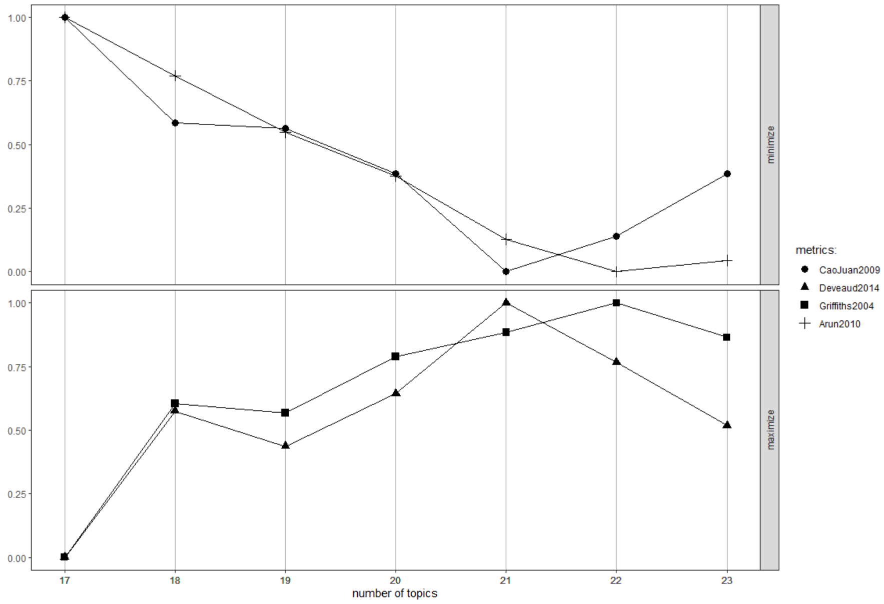

Figure 3: LDA-Tuning to find the best number of topics

4. Results

After discussing the methodology of this thesis in Chapter 3, this chapter presents the results of our projects. First, the results of a corpus analysis regarding certain characteristics are explained, this includes analyses of word frequencies and also their time courses. Furthermore, created co-occurrence networks are described. Finally, we will present and discuss the results of our research with topic modeling.

4.1. Corpus analysis:

The corpus was examined from a basic overview to specific details. So first we looked at the titles of the articles. We have filtered out the most frequent words and this resulted in three different word clouds.

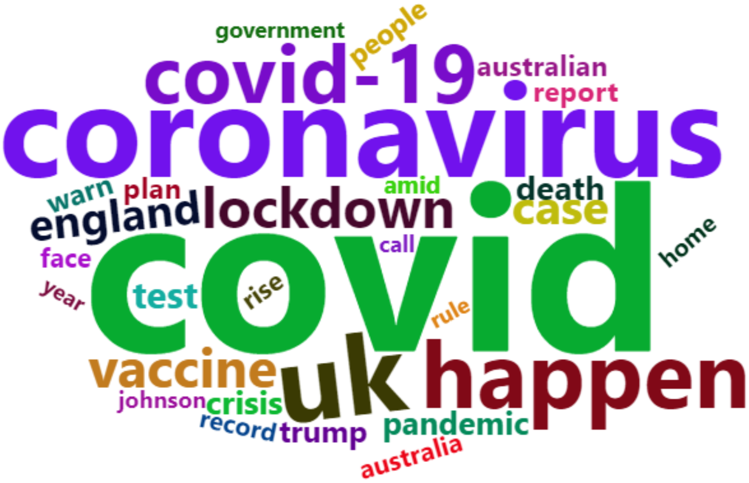

Figure 4: Most frequent words in title

We were able to determine that "covid", "coronavirus" and "covid-19" appeared the most in Figure 4. The words "lockdown", "vaccine", "uk" and "happen", were used in the titles as well. As well as less common words like "goverment", "people", "australian", "report", "warn", "plan", "england", "amid", "call", "death", "case", "face", "year", "test", "rule", "death", "case", "home", "johnson", "crisis", "record", "trump", "australia" and "pandemic". How does this relate to the text in the articles?

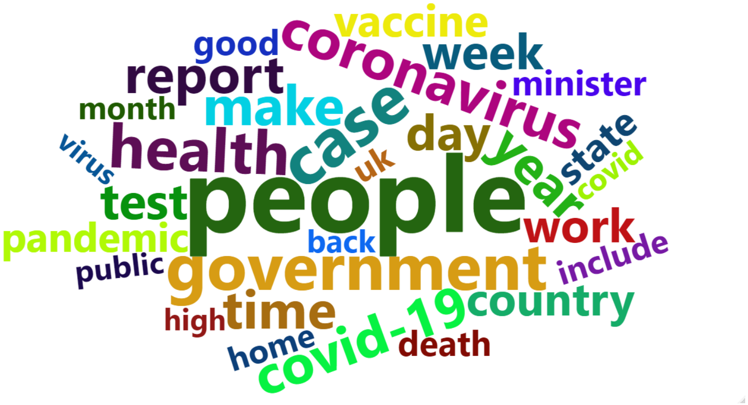

Figure 5: Most frequent words in text

"People", "health" and "government" were used mostly in Figure 5. Then "coronavirus", "case", "make", "vaccine", "week", "report", "day", "year", "country" and "covid-19" as words that appear not that often. Less common words were "good", "minister", "month", "virus", "pandemic", "state", "covid", "uk", "day", "year", "pandemic", "public", "back", "work", "include", "high", "home" and "death".

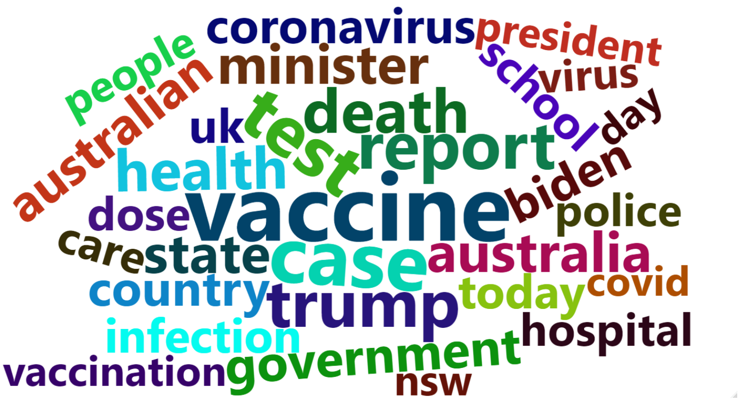

Figure 6: Most significant words based on TF/IDF measure

For the third word cloud TF/IDF measures were used as the basis for calculating the frequencies. As result in Figure 6 the most significant words were "vaccine", "case" and "trump" as the most common. Then words that almost all appear evenly such as, "coronavirus", "people", "minister", "president", "school", "virus", "australian", "uk", "test", "death", "report", "day", "health", "dose", "care", "state", "biden", "police", "country", "australia", "covid", "today", "infection", "vaccination", "government", "hospital" and "nsw". We also could determine the proportions of covid terms in the corpus Figure 12. The statistically most appeared word was "covid". Followed by the word "virus" and "coronavirus".

Stories of the pandemic: Another experiment related to word frequencies was to map them by monthly frequency. As a result we created three graphics, each with different focused characteristics.

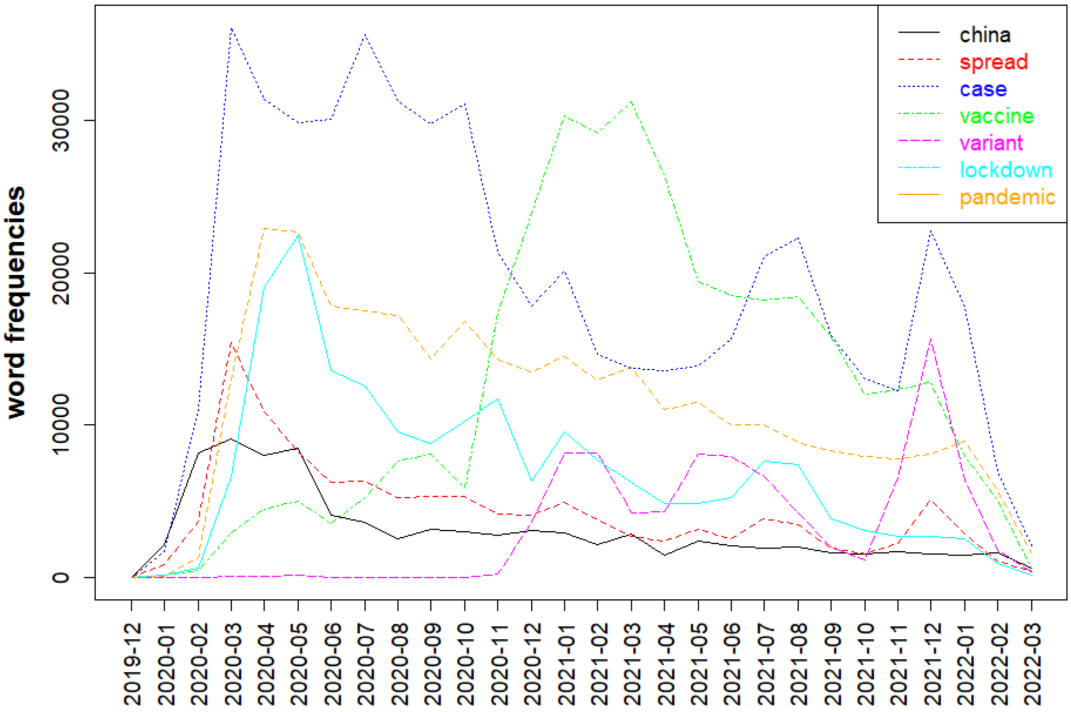

Figure 7: Word frequencies over time (milestones)

First, we focused at the general monthly word frequencies. It can be seen in Figure 7 that the word "case" is the most common from January 2020 and will then be overtaken by the word "vaccine" in November 2020. From June 2021, "case" will be listed as the most common word again. "Case" is reaching its peak in February 2020 and "vaccine" in March 2021. The word "pandemic" is rising in April 2020. The word "lockdown" is following and reaches its peak in May 2020. The words "china" and "spread" are also rising and reaching their peak March 2020. One word stands out in particular: "variant". As its starting to rise in December 2020 and reaching its peak in December 2021. With those results in mind we wanted to take a closer look in two specific time frames. First we looked at the results regarding the measures in Figure 13. The words "test" and "vaccinations" stood out in the entire time span. Surprisingly, the words "lockdown" and "home" had a similar trajectory. We were also able to find something similar with the words "mask" and "distance". Then we looked at the issue of vaccination in Figure 14 as this stood out in Figure 13 among the measures. Firstly astrazeneca appears on the map slowly rising. Then pfizer had its peak at January 2021. Followed by the peak of astrazeneca in March 2021. Towards the end of our considered period of time, the word booster appears. Which, however, cannot be assigned to an exact vaccine manufacturer. In contrast to the vaccinations already mentioned, moderna does not have enormous maximum values. The course after the vaccination appeared was almost linear with small elevations. It is also noteworthy that the vaccinations of sputik, janssen have not received much attention during the period of time we have considered.

4.2. Co-occurrence analysis:

After the results already described, we made further experiments with the co-occurrence analysis. This resulted into seven different Co-occurrence network graphics. Firstly we had a look over the entire period focused on COVID-19.

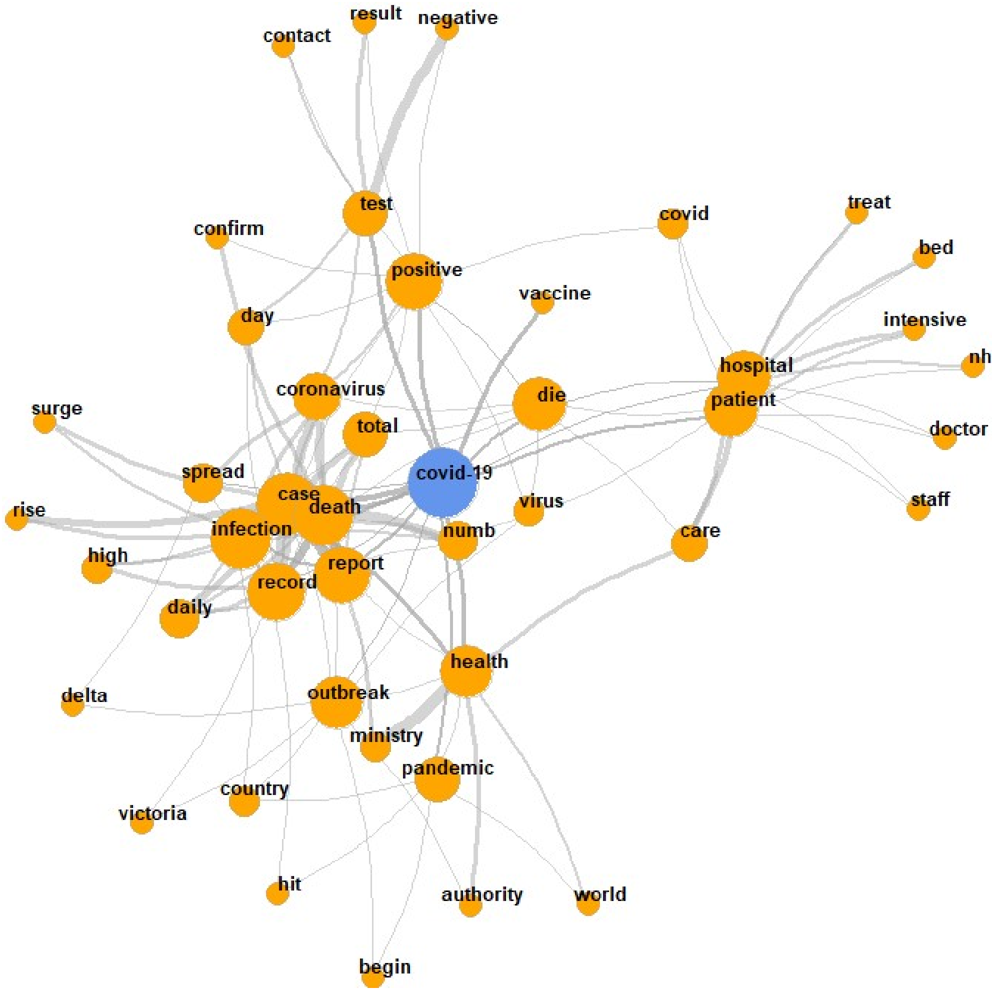

Figure 8: Co-occurrence network over the entire period

In figure Figure 8 different clusters with certain words can be seen. Nodes that have less than two connections have been removed from this network. The most closely clustered words are "hospital" and "patient". While in relation to COVID-19 it is also associated with the word "die". Then the words "case", "death", "infection", "record" and "report" are forming a larger cluster. The words "spread", "daily" "coronavirus" and "total" are closely related to this. Another cluster is dominated by the word "health", which is strongly associated with "ministry". The latter is also clustered with the words "outbreak" and "pandemic". The word "positive" is associated with the word "test" in the network pictured. "Test" also correlates with the word "daily". Which in turn is related to "case", "death" and "infection". In contrast to this network, in Figure 15 we have mapped one with the full number of nodes. Which clusters the individual words quite similarly, but into a larger and more specific network. In relation to both networks, we have created two more specific ones. These each have a period of at least one year. The second has a period of one year and two months. Which covers our entire investigative time span. In figure Figure 16 a time period from December 2019 to December 2020 was selected. Knots that had less than two connections were removed again. The result shows a similarity of the clustered words of figure Figure 8. However, pandemic and outbreak are clustered more closely. Only health close to both clustered and ministry is related to other words. The words case, death, infection, and report are clustered together again. The word confirm is added. The word "record", on the other hand, only seems to play a secondary role here. "Patient" and "hospital" are also clustered together again. Basically, only small things change in this network in contrast to the first one described. In Figure 17, on the other hand, there are some changed clusters. For example, "patient" is clustered without "hospitalized". In addition, there is a clustering with vaccine and dose. Then "health" is clustered with "reuters" and "test" with "positive". Finally the words case, death, infection, record and report continue to be clustered together.

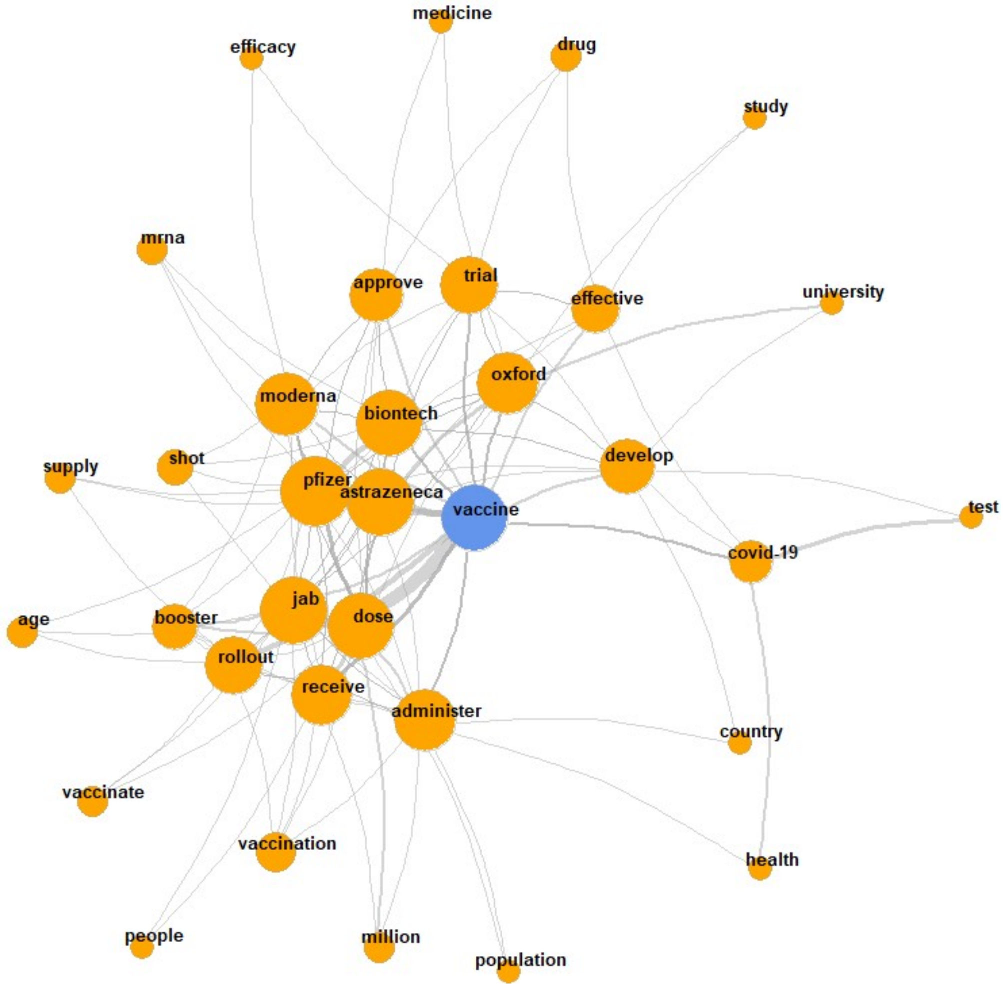

Figure 9: Co-occurrence network over the entire period

As a further experiment, we used the procedure from Figure 8 and Figure 15 again, but this time with the focus on the word vaccine. In Figure 9 we have applied the procedure of removing nodes that have less than two connections. As a result, we see that "moderna", "biontech", "pfizer" and "astrazeneca" are clustered in the network. As well as "Jab", "dose roll out", "receive", "booster" and administer. The words "approve", "trial", "Oxford", and "effective" are also related, but not very closely. On the other hand, in Figure 18 we can see differently weighted clustering because of the lack of removing nodes. A large clustering is in focus here: "vaccinate", "million", "roll out", "vaccination", "receive", "age", "supply", "booster", "jab", "dose", "administer", "shot", "moderna", "pfizer", "astrazeneca", "mRNA", "oxford" and "biontech" are in relation to "vaccine". There are also individual clusters such as "approve", "develop", "covid-19", "effective" and "trial". Which are associated with less weighted words.

4.3. Topic Modelling:

As a final method, we applied topic modeling to our corpus. We created a map of knowledge and three other graphs showing the topic proportions over time.

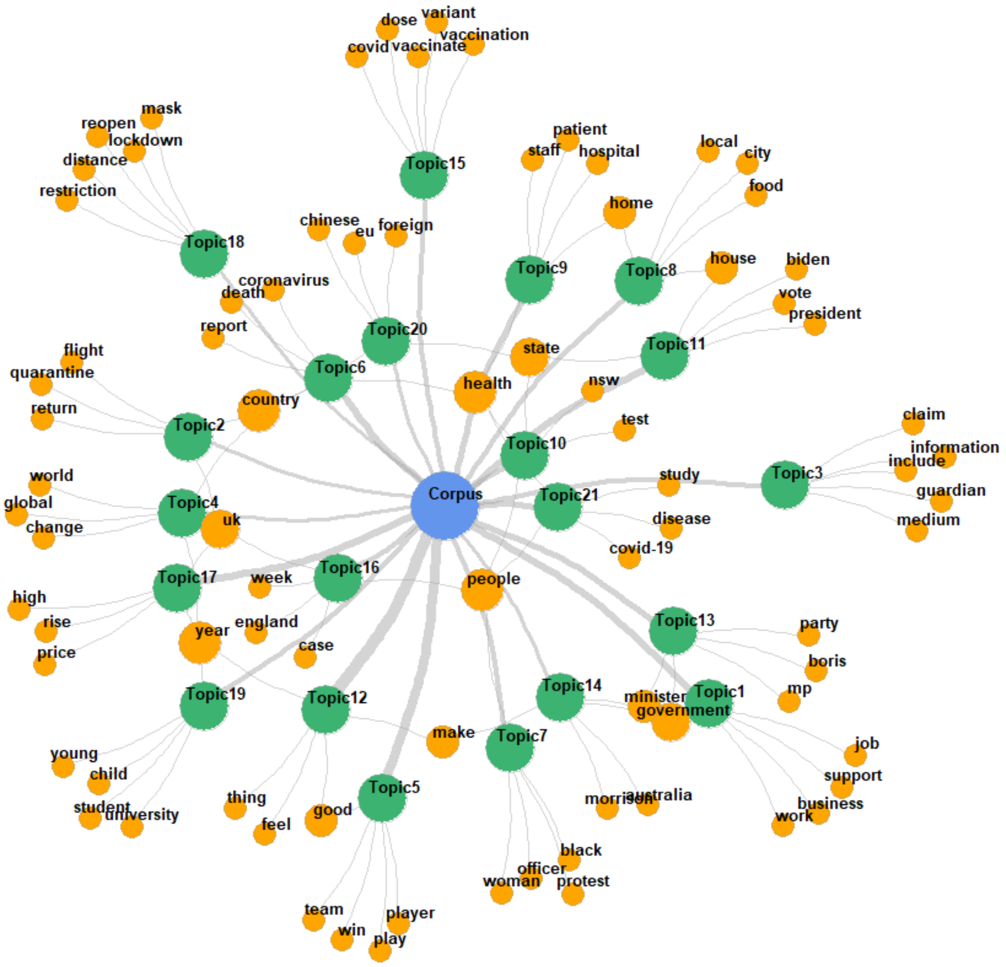

Figure 10: Map of Knowledge (Topic Modelling)

As the first experiment, we had twenty-one topics created automatically in Figure 10. This map of knowledge shows the connection between the corpus and the topics. The thickness of the respective knots reflects the presence of the topic throughout the corpus. The five most common words within the topic are derived from the individual topics. Since words within this graphic are only sited once, words can be shared with other topics. Like the word "country" in the map. This word is shared by Topic2, Topic6 and Topic4. Topic16 contains the following five most common words: "people", "test", "england", "government" and "uk". Topic15 contains vaccine, "vaccination", "covid", "dose" and "vaccinate". Topic10 includes "case", "people", "health", "nsw", and "test". Topic1 includes "pay", "job", "work", "government" and "business". The temporal distribution of each topic will be investigated in more detail in the next experiments.

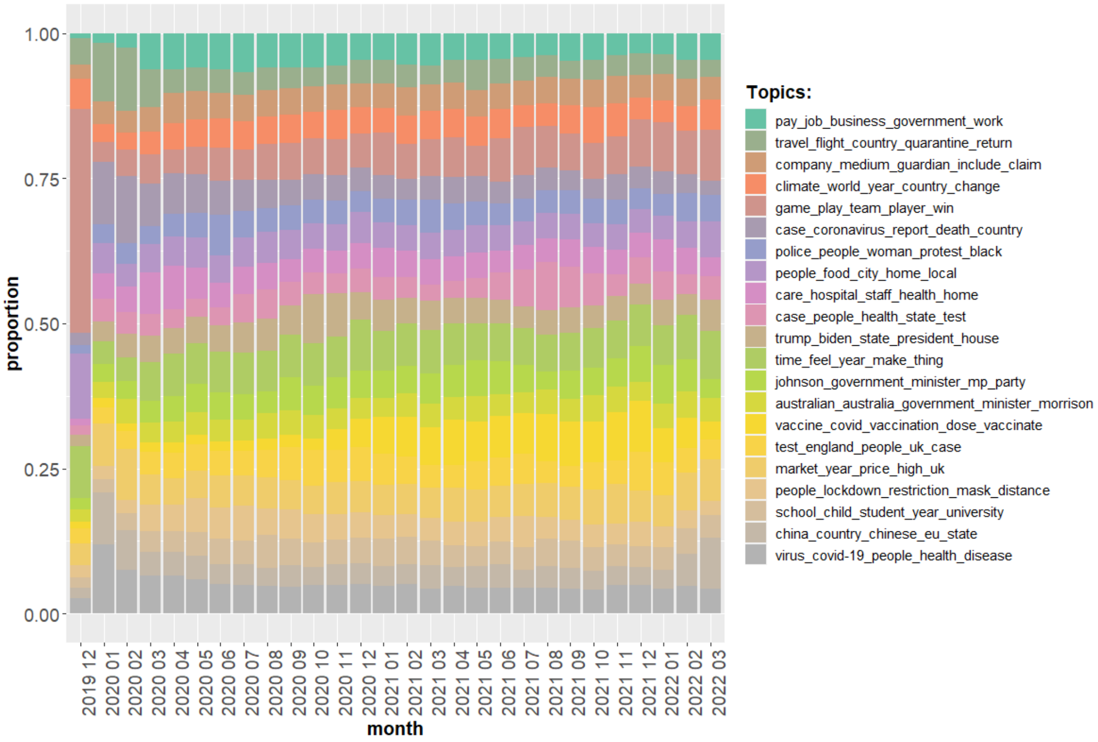

Figure 11: Topic proportions over time

In Figure 11 the topics are listed and the proportions over time can be derived. In order to be able to look at this in more detail, the entire time span had to be halved in each case. So we created two more graphics. In Figure 20 the period was halved. The two periods considered are from December 2019 to December 2020 and January 2021 to March 2022. However, the individual trends cannot be clearly identified here, so the period was halved again. In Figure 21 we have four time periods. The first period runs from December 2019 to June 2020, the second from July 2020 to December 2020, the third from January 2021 to June 2021 and the fourth from July 2021 to March 2022. Here we can see that Topic16 (people_test_england_goverment_uk) increased in the second period. However, in the third and fourth periods it has slightly lost its relevance. At Topic15 (vaccine_vaccination_covid_dose_vaccinate) we can see that the incidence increased in the second period and increased sharply in the third and fourth. This can be explained by the relevance of the booster vaccination that took place at that time. In Topic10 (case_people_health_nsw_test) we can see that the relevance increased in the second period, but decreased again in the third. In the fourth, however, she gained relevance again. For Topic1 (pay_job_work_goverment_business) we can see a steady decrease over the four periods. Thus, at the beginning of the pandemic, the topic was of increased relevance. It should be noted here that to create the topic proportions in Figure 20 and in Figure 21, an LDA model with a lower alpha parameter value $\alpha = 0.2$ was calculated to make the topic distributions better visible.

5. Discussion

The results were surprising in some aspects. Of course, words like "Covid" and "Coronavirus" stood out clearly in the headlines of the articles in our corpus. However, the text itself was more about "people" or "government". In the last two years, there have been many important findings and information published by the Guardian regarding the global pandemic. The articles often refer to government assessments or actions. Since there were often many approaches and opinions, the keywords "people" or "government" were used frequently in the text, while the headlines did not repeat words and tended to be short and concise. It was assumed that words like "covid" and "coronavirus" would appear in the articles with some regularity over time. We expected topics around the lockdowns to occur in phases. Vaccination on the other hand was not expected to become increasingly relevant until the spring of 2021, when the first vaccines were approved. Figure 5 in section 4.1 shows that we were partially correct in our assumptions. It is interesting to note, however, that in the case of vaccination, for example, a decline in reporting is again evident from summer 2021 onward. It is also good to see that virus variants did not play a role at the beginning and only came into the discussion towards the end of 2020. The various lockdowns can be guessed at in the data, but we had expected a clearer result. This is probably due to the British corona policy, which deviated from these measures relatively early, after very strict and hard lockdowns at the beginning, and had relatively few overall social restrictions that count as lockdowns. With the Co-occurrence analysis we figured out some correlations between certain cue words. It could be shown that around the topic "covid-19" actually similar much of the death, as of the infection itself was reported. There was also a close connection with the topic of vaccination. Here, the various well-known vaccine manufacturers (Pfizer, Moderna, Astrazeneca and Biontech) played a central role. Another cluster was formed by important terms for vaccination processes such as the dose, the prick or the booster.

Let us now take a look at the results of the topic modeling. For our entire corpus, the optimal number of topics was 21. If we look at these more closely, they can be grouped into sub-topics again. As was to be expected, some topics related to the corona pandemic can be easily assigned: there is a field around the vaccinations, one around the measures, one around the reporting, and so on (cf. figure 10, chapter 4.3). In the resulting diagram, it is also easy to see the connections between individual topics. For example, young people, students and educational institutions often seem to have been discussed in the same context as economic and financial issues. The common term that is meant here is "the year". However, this may well be due to the fact that in relation to young people, "the school year" or the "years of youth" are often reported, while economic topics are about "the economic year" or the "financial situation after a year. If we look at the development of the topics over time (see Figure 11), two things in particular stand out. First, there is a big difference between December 2019 and the rest of the time until today. This can have several reasons. First, we do not have the same amount of data for December as for the other months, since COVID-19 only played a more international role from mid-December on. In addition, most of the reports at that time only referred to the spread in china and people still believed that they could prevent a worldwide pandemic by isolating themselves. We also see that the third topic in the graphic takes up a lot of space. However, this topic cannot be classified anywhere well and seems to be a statistical error due to frequent words used in the reporting. Otherwise, it is surprising that the remaining topics remain relatively similar over the two years. There are no discernible fluctuations, for example in connection with the lockdowns or the like. Only in the area of vaccination can we see an increase over time. Of course, this can be explained by the fact that the realistic prospects of vaccination have increased over the course of the pandemic and that it took a while for vaccination to become available to society as a pandemic fighter. In the last two months, however, this has decreased again, presumably because a large proportion of the British population has now been vaccinated.

6. Conclusion

During our work with the data from the Guardian, we had to overcome some difficulties. We used a variety of approaches to analyze the data, each of which produced a different presentation of the results. We started with frequency, simply counting the number of words in relation to the time of appearance of the articles. This approach is relatively simple to implement, but can be improved by using document frequency. To do this, we used the TF/IDF approach, which allowed us to represent the uniqueness of the words. However, it was important to always look at the results in comparison to the results of other words. Another approach we followed was topic modeling. The advantage here is that, when used correctly, it creates a semantic context around different terms and these can be easily represented. One difficulty here, however, is to bring data into a temporal relationship. If, for example, the data set is divided into the years 2020 and 2021, different topics can emerge and make a direct comparison of the data nearly impossible. This is where our computing power reached its limits, and one of our computers sometimes had to work through the entire weekend. The most time-consuming part was the co-occurrence analysis. We discovered that it makes sense to directly delete nodes with fewer than two connections. In the end, however, this made it very easy to see which terms appeared frequently in connection to others and which appeared less frequently. All in all, we have tried various ways to bring all our acquired knowledge together in a map of knowledge. The main difficulty was the time component, as we also wanted to show changes in reporting over the past 28 months. We didn't manage to look at all the approaches together, but we did find some ways to present and compare our results. In principle, it should also be possible, with considerable additional effort and time, to combine at least the co-occurrence analysis and the topic modeling, since the entire period was examined here in particular.

A. Appendix

A.1. Covid Terms

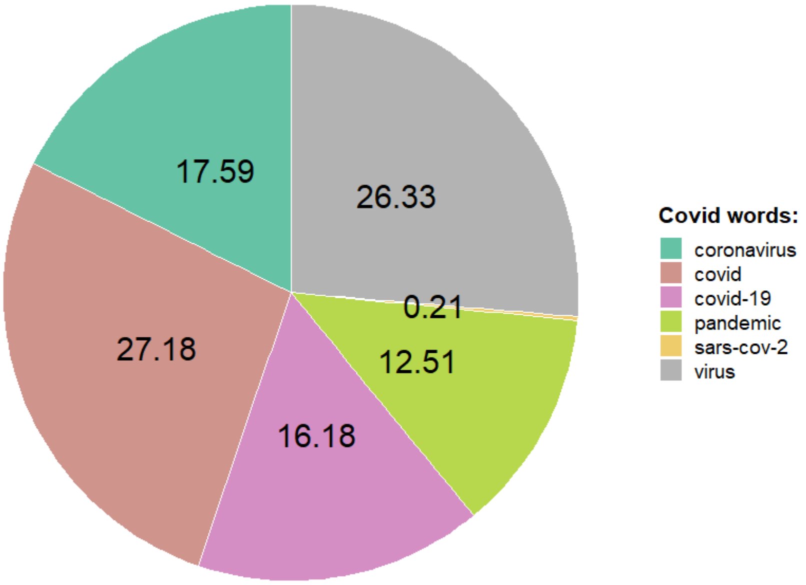

Figure 12: Proportions of covid terms in corpus

A.2. Word frequencies Measures

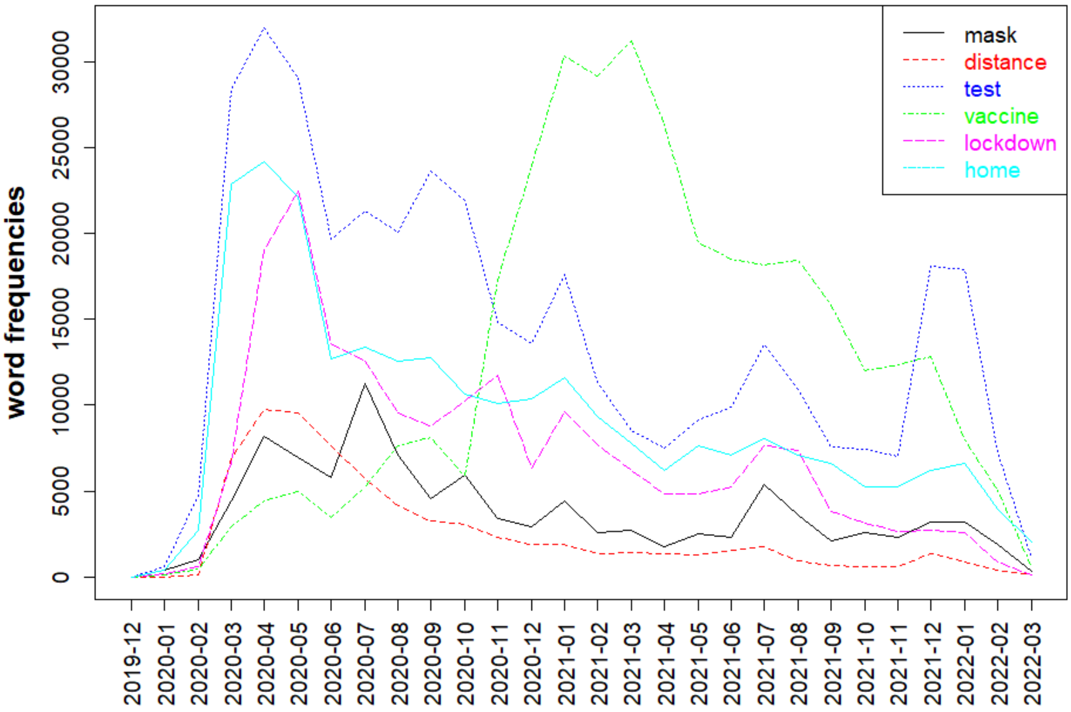

Figure 13: Word frequencies over time (measures)

A.3. Word frequencies Vaccines

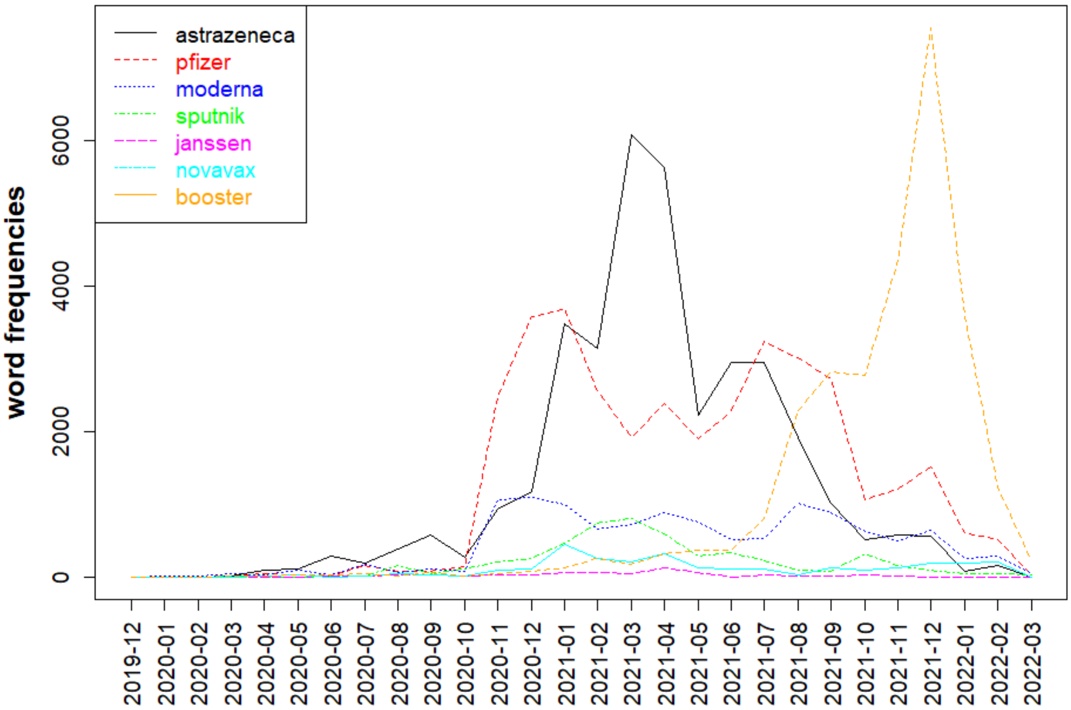

Figure 14: Word frequencies over time (vaccines)

A.4. Co-occurrence Network - covid-19 (complete)

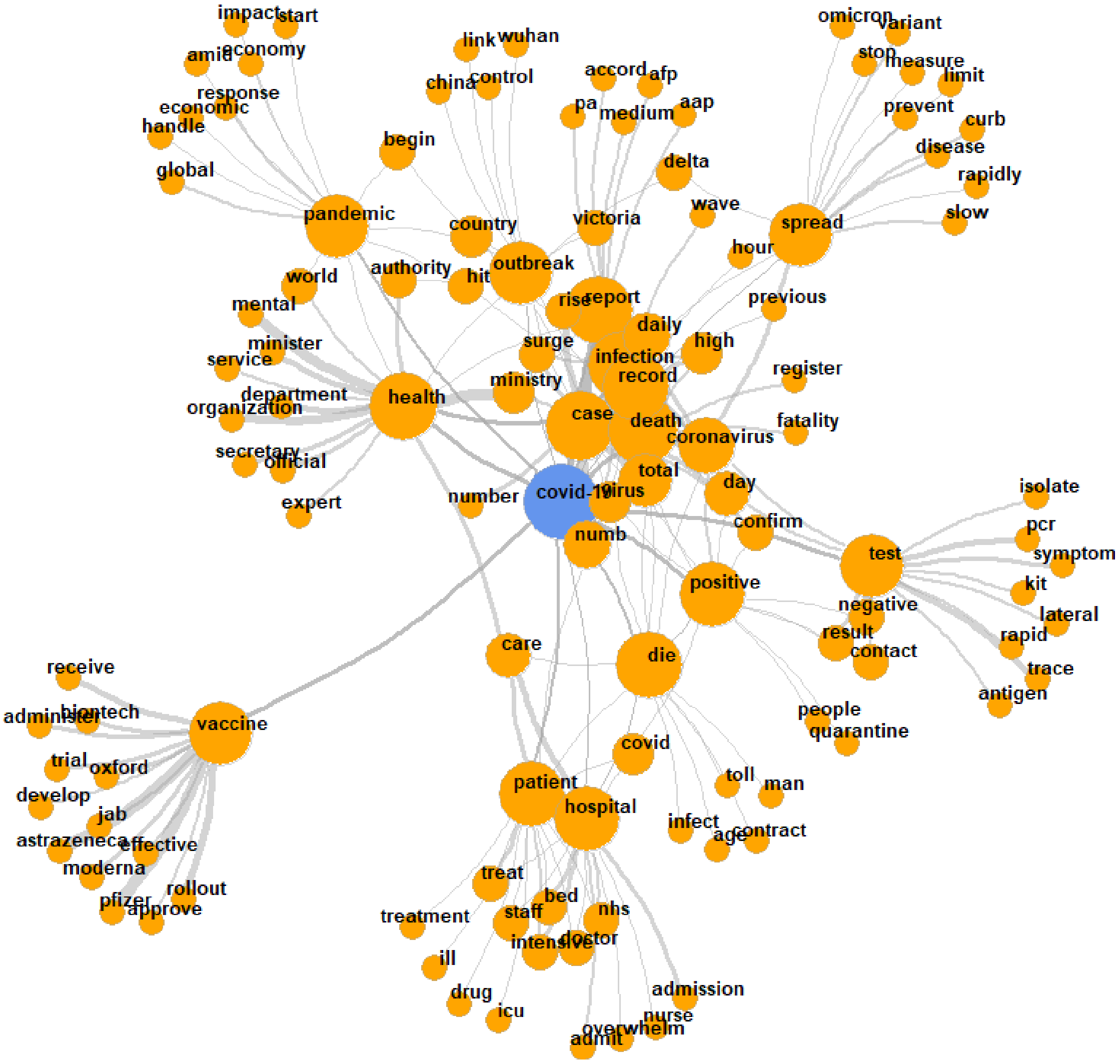

Figure 15: Co-occurrence network over the entire period (complete)

A.5. Co-occurrence Network - covid-19 (Dec2019 to Dec2020)

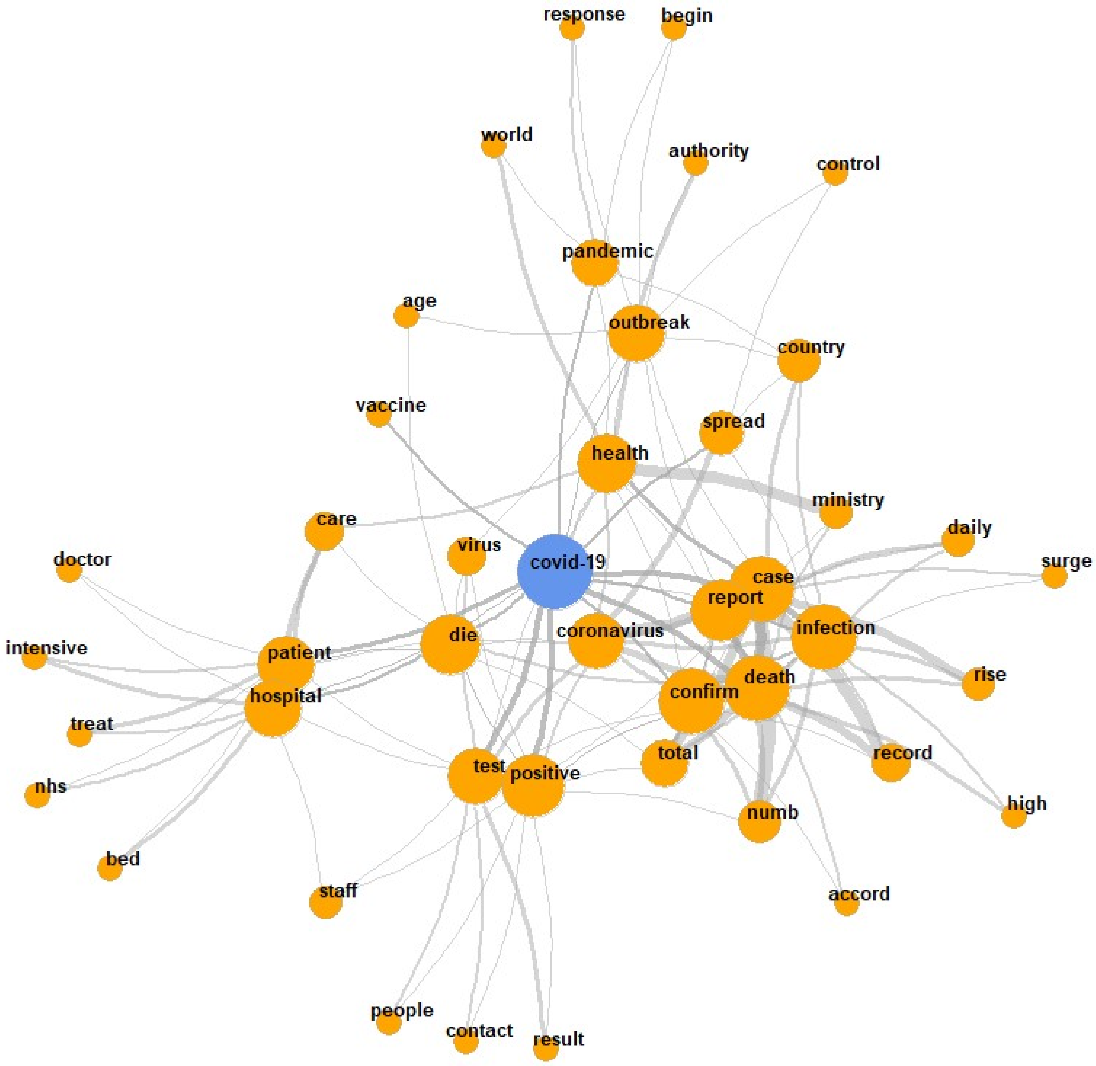

Figure 16: Co-occurrence network for the period from Dec2019 to Dec2020

A.6. Co-occurrence network - covid-19 (Jan2021 to Mar2022)

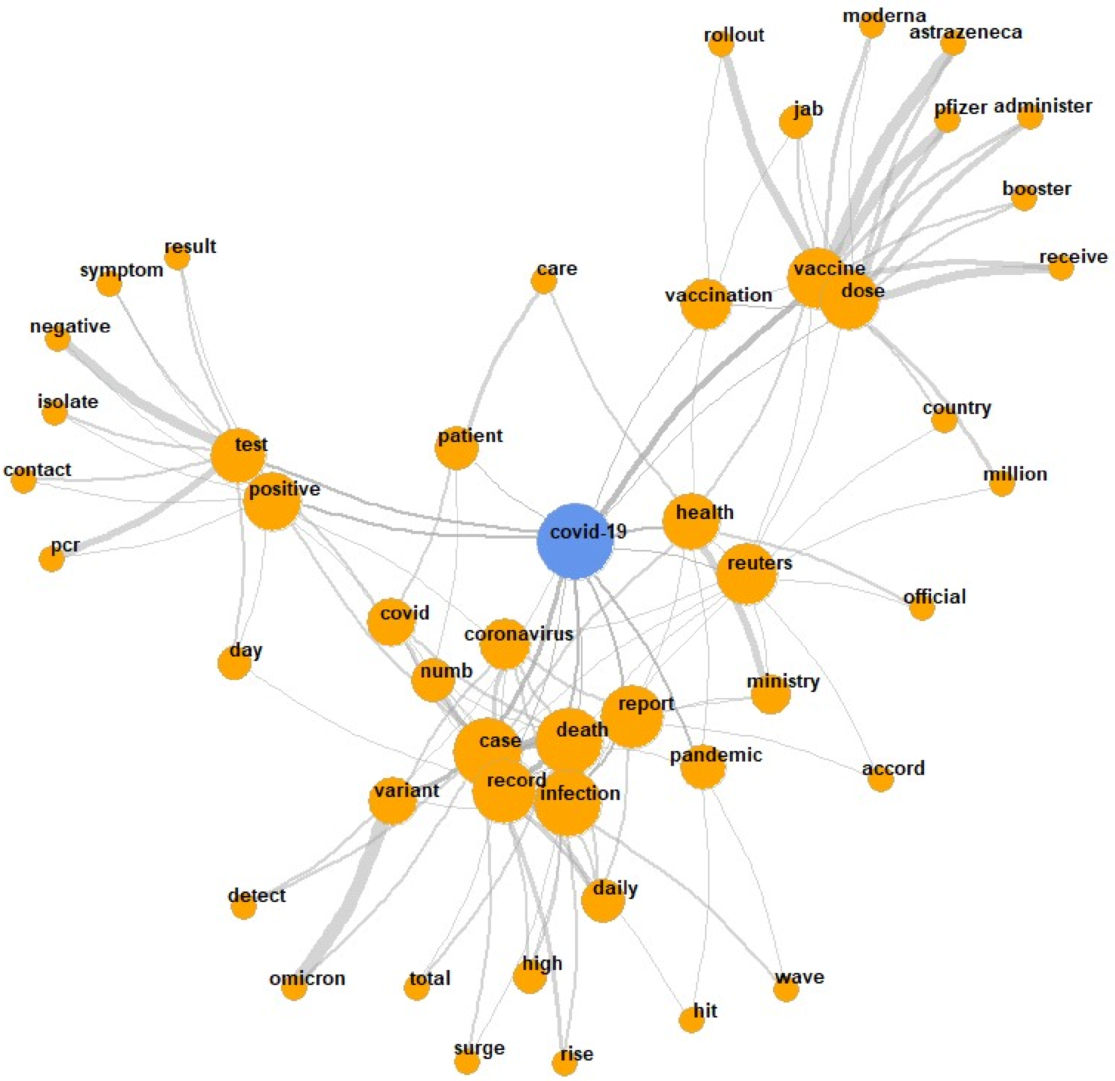

Figure 17: Co-occurrence network for the period from Jan2021 to Mar2022

A.7. Co-occurrence Network - vaccine (complete)

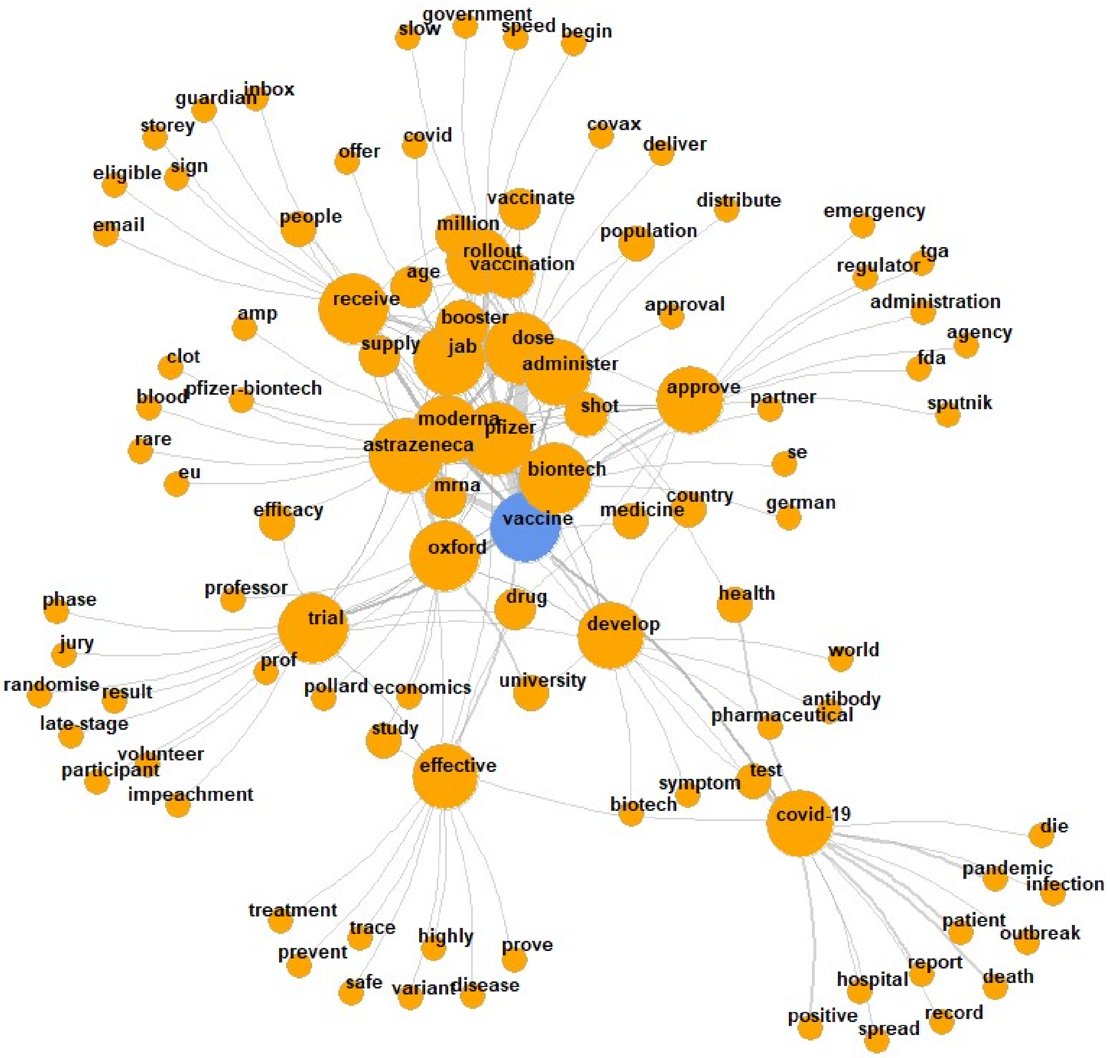

Figure 18: Co-occurrence network over the entire period (complete)

A.8. Topic Modelling: LDA-Tuning

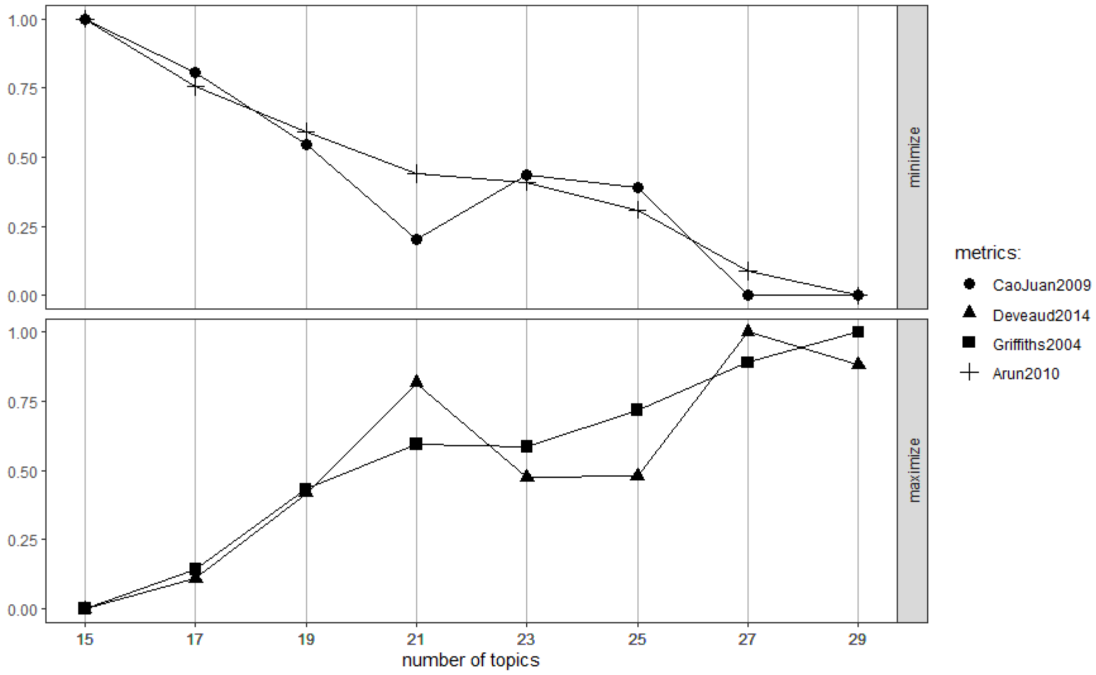

Figure 19: LDA-Tuning to find the best number of topics

A.9. Topic Modelling: proportions over time (2 time periods)

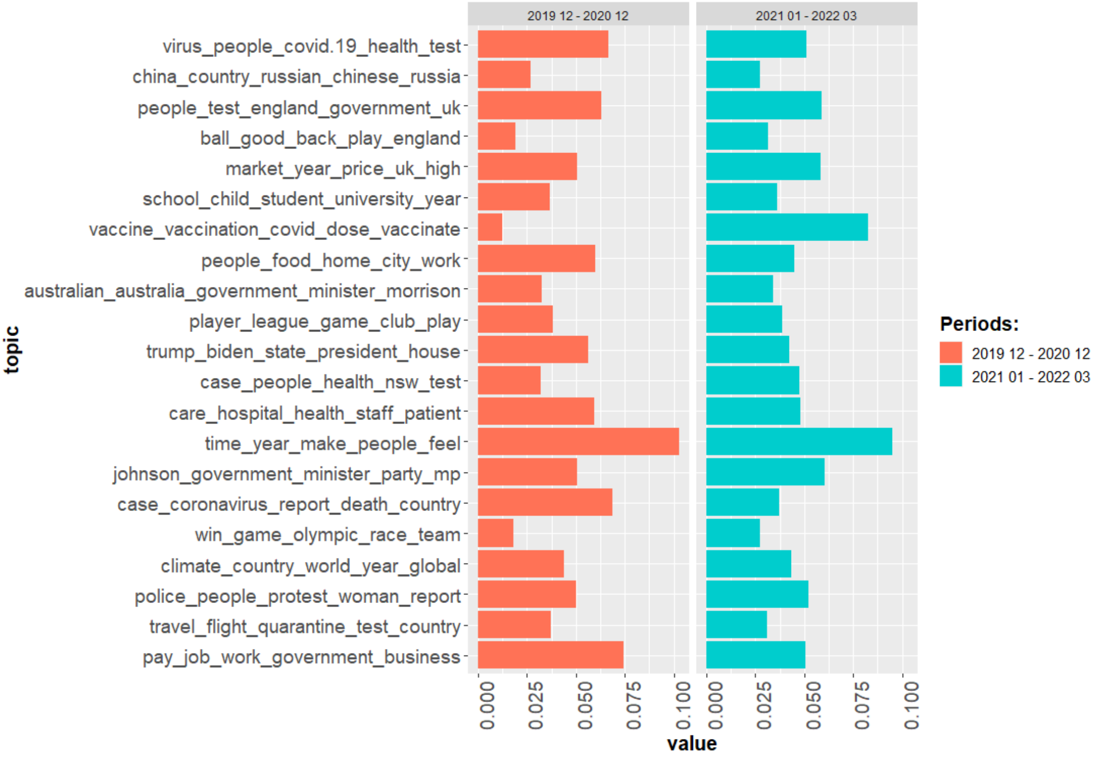

Figure 20: Topic proportions over time (2 periods)

A.10. Topic Modelling: proportions over time (4 time periods)

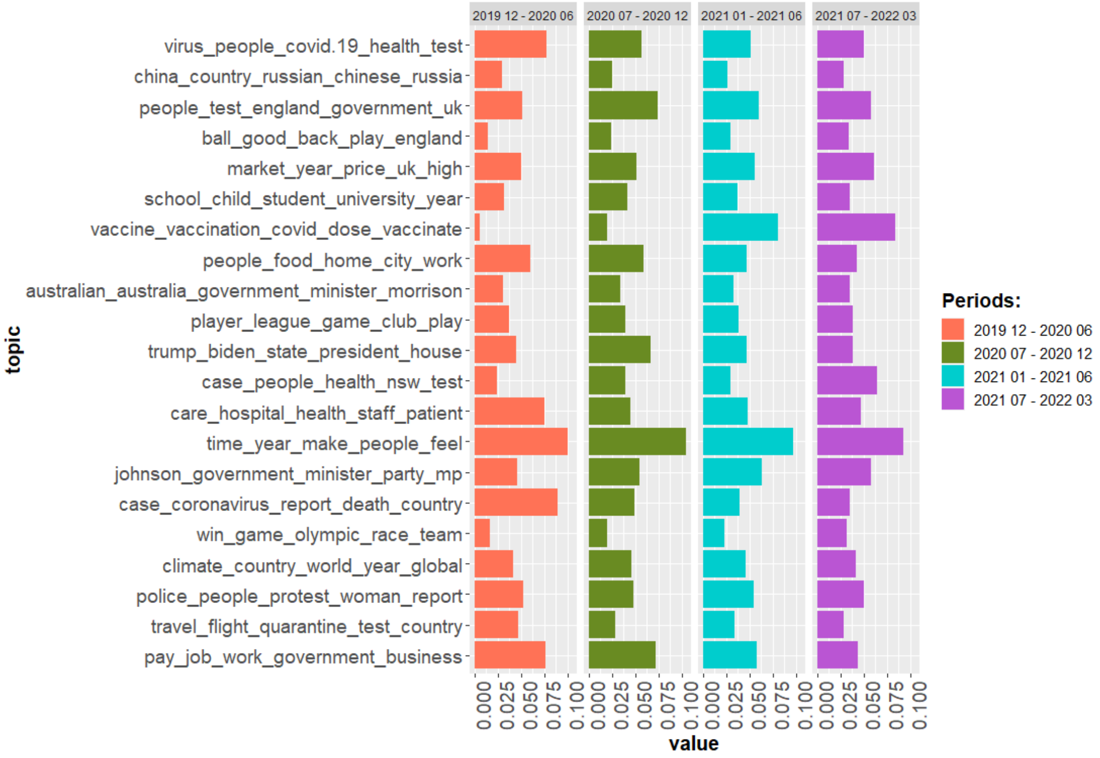

Figure 21: Topic proportions over time (4 periods)

A.11. Used R packages

The following R packages were used in this work:

- guardianapi[5]: To query and retrieve guardian articles
- quanteda[6]: Text data management and analysis
- dplyr[7]: Data manipulation
- stringr[8]: Easier work with strings
- ggplot2[9]: Creating graphics
- wordcloud2[10]: Data visualization
- topicmodels[11]: Fitting topic models
- igraph[12]: Creating and manipulating graphs and analyzing networks
- tidyverse[13]: Collection of R packages data science
- RColorBrewer[14]: Provides color schemes for graphics

[5] https://cran.r-project.org/web/packages/guardianapi/guardianapi.pdf
[6] https://quanteda.io/
[7] https://dplyr.tidyverse.org/
[8] https://www.rdocumentation.org/packages/stringr/versions/1.4.0
[9] https://ggplot2.tidyverse.org/
[10] https://cran.r-project.org/web/packages/wordcloud2/vignettes/wordcloud.html
[11] https://cran.r-project.org/web/packages/topicmodels/index.html
[12] https://igraph.org/r/
[13] https://cran.r-project.org/web/packages/tidyverse/index.html
[14] https://cran.r-project.org/web/packages/RColorBrewer/index.html

List of Figures

List of Tables

Bibliography

Manar D. Samad Ahamed Sabber. Information mining for COVID-19 research from a large volume of scientific literature. *CoRR*, abs/2004.02085, 2020. URL https://arxiv.org/abs/2004.02085.

Greg Hill Marco Carvalho Marco Arguedas Thomas C. Eskridge James Lott Rodrigo Carvajal Alberto J Canas, Roger Carff. *Concept Maps: Integrating Knowledge and Information Visualization*, pages 205–219. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005. ISBN 978-3-540-31962-7. doi: 10.1007/11510154_11. URL https://doi.org/10.1007/11510154_11.

Albert Sechehaye Charles Bally. *Kapitel V. Syntagmatische und assoziative Beziehungen*, pages 147–152. De Gruyter, 2011. doi: doi:10.1515/9783110870183.147. URL https://doi.org/10.1515/9783110870183.147.

Corel Corporation. Knowledge-map, was ist eine wissenslandkarte. https://www.mindmanager.com/de/features/knowledge-map/, 2022. Online; Acces date: 2022-03-30.

E. Lynn Usery Dalia E. Varanka. The map as knowledge base. *International Journal of Cartography*, 4(2):201–223, 2018. doi: 10.1080/23729333.2017.1421004. URL https://doi.org/10.1080/23729333.2017.1421004.

Tau Ming Liew Jing Xuan Koh. How loneliness is talked about in social media during covid-19 pandemic: Text mining of 4,492 twitter feeds. *Journal of Psychiatric Research*, 145:317–324, 2022. ISSN 0022-3956. doi: https://doi.org/10.1016/j.jpsychires.2020.11.015. URL https://www.sciencedirect.com/science/article/pii/S0022395620310748.

Herbert A. Neumann. *Die Entstehung der Virologie*, pages 135–144. AW Wissensverlag, 2019. ISBN 9783940615596. doi: doi:10.1515/9783110870183.147.

Imtiaz Mahmud Nafi Md. Musleh Uddin Hasan Niaz Mahmud Zafri, Sadia Afroj. A content analysis of newspaper coverage of covid-19 pandemic for developing a pandemic management framework. *Heliyon*, 7(3):e06544, 2021. ISSN 2405-8440. doi: https://doi.org/10.1016/j.heliyon.2021.e06544. URL https://www.sciencedirect.com/science/article/pii/S2405844021006472.

Charles A. Perfetti. The limits of co-occurrence: Tools and theories in language research. *Discourse Processes*, 25(2-3):363–377, 1998. doi: 10.1080/01638539809545033. URL https://doi.org/10.1080/01638539809545033.

Martin Schweinberger. Analyzing Co-Occurrences and Collocations in R kernel description. https://slcladal.github.io/coll.html#1_Extracting_N-Grams_and_Collocations, 2022a. Online; Access date: 2022-04-01.

Martin Schweinberger. Text analysis and distant reading using r. https://slcladal.github.io/textanalysis.html#What_is_Text_Analysis, 2022b. Online; Access date: 2022-04-01.

Martin Schweinberger. Topic modeling with r, 2022c. URL https://slcladal.github.io/topicmodels.html. Online; Access date: 2022-03-18.

George W. Furnas Thomas K. Landauer Richard Harshman Scott Deerwester, Susan T. Dumais. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990. doi: https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9. URL https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-4571%28199009%2941%3A6%3C391%3A%3AAID-ASI1%3E3.0.CO%3B2-9.

Laura Tilton Taylor Arnold. *Humanities Data in R*. Quantitative Methods in the Humanities and Social Sciences. Springer, Cham, 1 edition, 2018. doi: 10.1007/978-3-319-20702-5. URL https://doi.org/10.1007/978-3-319-20702-5.

Valneva. Covid-19 – vla2001. https://valneva.com/research-development/covid-19-vla2001/, 2022. Online; Access date: 2022-03-30.

H. Ping Tserng Yu-Cheng Lin, Lung-Chuang Wang. Enhancing knowledge exchange through web map-based knowledge management system in construction: Lessons learned in taiwan. *Automation in Construction*, 15(6):693–705, 2006. ISSN 0926-5805. doi: https://doi.org/10.1016/j.autcon.2005.09.006. URL https://www.sciencedirect.com/science/article/pii/S0926580505001317. Knowledge Enabled Information System Applications in Construction.