Discourse-Oriented German Climate Change Glossary

Anna-Janina Goecke (777707)
University of Potsdam
Text Mining in the Social Sciences - Prof. Dr. Stede

28th September 2021

1 Project Description

This project aims to create a discourse-oriented glossary of German climate change terms with special regard to "loaded" terms in either a positive or negative manner. The main idea was to set a focus on the communities in which the terms are used and for what purpose. Additionally, we want to investigate in which types of context the climate change terms can be found. Within the scope of this project, a "loaded" term is defined as a term which carries a specific connotation and is only used by certain communities or in certain situations and contexts. Since this project aims to focus on climate change terms, German climate change compounds will be the target terms of the analysis and evaluation. The whole project, especially the creation of a glossary, was first meant to be part of a collaboration but then materialized into a single person project as part of a University class on "Text Mining in the Social Sciences". The project report is organized as follows: Section 2 defines the basis of this project with special regard to the initial idea of building a glossary. Section 3 forms the first step of converting the manual approach described in section 2 into the handling of bigger textual data sets by retrieving data from the internet and translate this kind of data into a corpus format. Section 4 comprises the main part of this paper, presenting results of Text Mining methods and their evaluation.

The code including the code documentation[1] and the corpus data can be found in the Github repository[2]. Additionally, the original files from the Web Scraping, all plots, and the final glossary text files are included. Since there was no prior expertise in using R [6], most of the knowledge has been gained within the work on this project by working through various websites and tutorials[3].

[1] For the R Notebook including code and code documentation please see `glossary_notebook.Rmd` or `glossary_notebook.nb.html`

[2] `https://github.com/ajgoecke/climate_change_glossary`

[3] `https://quanteda.io/articles/quickstart.html`,
`http://inhaltsanalyse-mit-r.de/grundlagen.html`,
`https://bookdown.org/yihui/rmarkdown/notebook.html`,
`https://www.tidytextmining.com`

2 Glossary

In order to create a glossary of "loaded" German climate change compounds it was first necessary to get an overview of already existing terms. Therefore, a list of those terms was created via online research and reading of multiple short online articles regarding topics such as "climate change" itself and closely related ones. The online research of German climate change compounds resulted into a list of loaded terms such as "Klimahysterie", "Klimaskeptiker" and "Klimareligion". For the next step of creating a glossary a random subset of these terms was taken into consideration. Since this project aims to evaluate terms with respect to their function within discourse, we were especially interested in the use of the given terms. Accordingly, we wanted to identify the potential context of these compounds and in which communities they are primarily used. Another interesting aspect was to determine if the term is used by the community which it wants to describe itself (internal attribution) or whether it is rather used by the opposed community to describe the other group (external attribution). With respect to this, we decided on manually researching and formulating short paragraphs containing information about a term's context, community and use. The following example illustrates the description of one of the terms[4]:

Climate Fanaticism
Climate fanaticism denotes the "extreme" perspective of climate activists on climate change. This term is predominantly used by right-winged parties in connection with the term "climate madness" to refer to climate activists such as Greta Thunberg or Annalena Baerbock. The term receives a very negative connotation by the radical and aggressive component of "fanaticism" and serves to discredit potential impacts of climate change. AfD members such as Beatrix von Storch and Jörg Meuthen make use of this term to question rational thinking of climate activists. Furthermore, they use the term on social media platforms within the context of climate hysteria and dictatorship.

While working on the identification of the glossary terms, it appeared that the communities using loaded terms can be divided into two major groups: climate activists and climate sceptics (those two groups will be called "activists" and "sceptics" within this project). Certainly there are more subgroups settled within both groups but for simplicity and with respect to the scope of this project, we will focus on these two primary groups.

The full list of loaded climate change terms and its descriptions have been documented on the notion.so webpage[5].

[4]This is a translation of the originally constructed paragraph for the German glossary version.
[5]`https://www.notion.so/invite/3fd00d9bdeea76db4bb9e9654dadaac3af33403b`

3 Corpus

Web Scraping. Since we wanted to dive deeper into the context and the usage of the terms we identified within the glossary creation step, it was necessary to gather big textual data sets containing the loaded terms. In order to create a corpus based on these texts, the Web Scraping tool Trafilatura [1][6] was used. With the help of this tool we could easily extract text data from input webpages. Accordingly, a link of the webpage containing the wanted text data has been feeded into the tool which then created an output folder containing a text file for each of the subpages in a hierarchical manner. The whole process was divided into two subprocesses: i) getting the links of all the subpages within a given webpage, ii) retrieving the text files of all the subpages by inserting the previously created list of subpages. For the purpose of the project, the contents of the webpages EIKE[7] and Fridays For Future[8] were used.

EIKE

The self-appointed institute for climate and energy is one of the best known German associations of the climate change denial community. Its members decline the existence of climate change caused by humans and its effects on nature. In the past, German press and climate scientists have labeled EIKE as the heart of political actions of the climate change denial community and systematic spreading of disinformation. A major part of its members could be allocated within the German right-winged party Alternative für Deutschland (AfD). The members describe themselves as "Klimaskeptiker" (en: "climate sceptics") or "Klimarealisten" (en: "climate realists").[4]

Fridays for Future

The Fridays for Future community is a social movement standing up for fast and efficient climate protection. It aims to call attention for political grievances concerning climate change and to initialize climate protection measures. Unlike EIKE, the Fridays for Future community recognizes climate change as caused by humans.[9]

Corpus Creation

For the next step – the creation of a text corpus – all the data has been loaded into R [6]. We decided on using R within this project given that it provides very easy and intuitive corpus handling and evaluation methods. To be able to perform Text Mining methods and to constitute the corpus itself, the Quanteda [2] library was used. When loading the text files into R, it appeared that the corpus size of the two webpages that were initially thought to represent the activists community (Fridays for Future) and the sceptics community (EIKE) significantly differed in size. While EIKE provided us with 14,000 text files, the German version of Fridays for Future only consisted of 500 texts. Con-

[6] A documentation of the tool can be found here: https://trafilatura.readthedocs.io/en/latest/

[7] https://eike-klima-energie.eu

[8] https://fridaysforfuture.de

Name	Abbreviation	Website	Group
Fridays for Future	fff_de	https://fridaysforfuture.de	Activists
Gerechte 1 Komma 5	gk	https://gerechte1komma5.de	Activists
IKEM	ikem	https://www.ikem.de	Activists
Klimafakten.de	kf	https://www.klimafakten.de	Activists
Klimareporter	kr	https://www.klimareporter.de	Activists
German Zero	zero	https://www.germanzero.de	Activists
EIKE	eike	https://eike-klima-energie.eu	Sceptics

Table 1: List of all websites used for the corpora including their abbreviation in the project and origin.

sequently, we checked for additional climate activists webpages, see table 1, and added those to the activist corpus. Given that the corpus data also included English texts, for each text the language has been identified. Within this data cleaning step all non-German texts have been excluded from the corpus. Finally, we ended up with three corpora, one for the activists (2000 texts), one for the sceptics community (2000 texts), and a combined version of the two corpora. The corpora include, inter alia, the following metadata for each text: Number of sentences, origin (from which webpage the data has been retrieved) and group (activists or sceptics).

Corpus Statistics

Within the cleaning of the corpora (removal of non-German texts) we decided to also have a closer look at the metadata of the corpora. This has been done via plotting of some of the corpus statistics such as the number of sentences.

Figure 1 shows a plot of the number of sentences for the activists corpus (P2000), figure 2 for the sceptics corpus (C2000). The corpora clearly differ in text size: While P2000 has a mean of 24,7 sentences per text, C2000 consists of much longer texts with a mean of 76,3 sentences per text. This fact should be taken into consideration when evaluating the term frequency of the corpora.

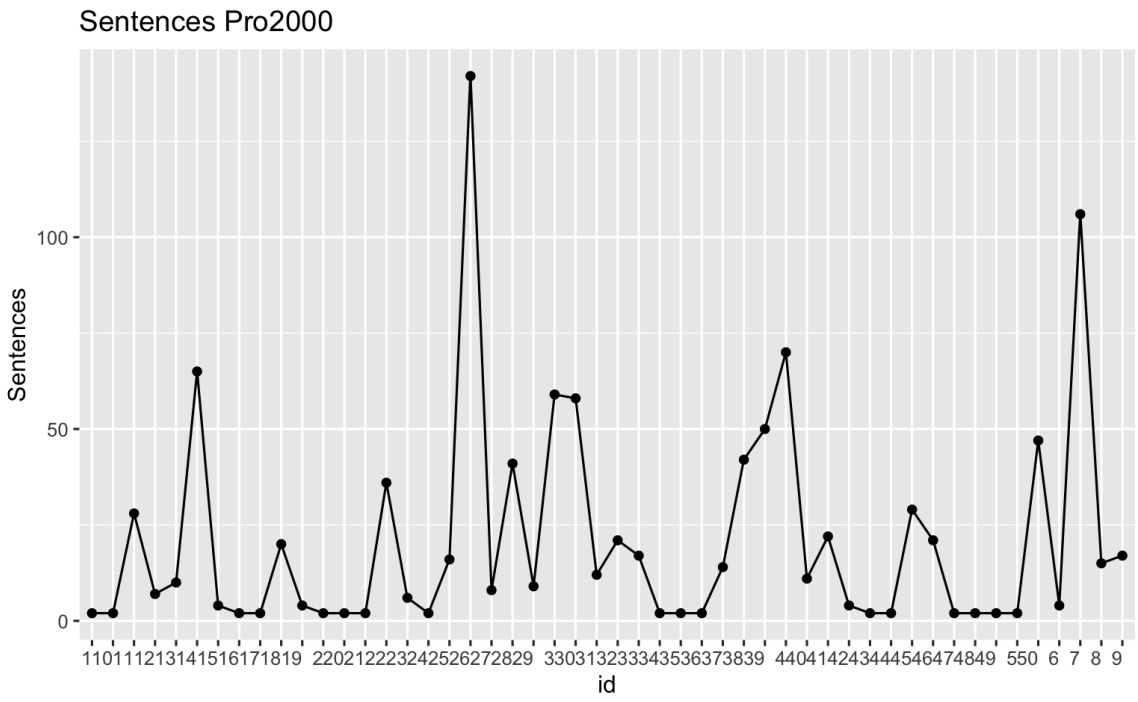

Figure 1: Number of sentences in activists corpus (sample of 50 texts).

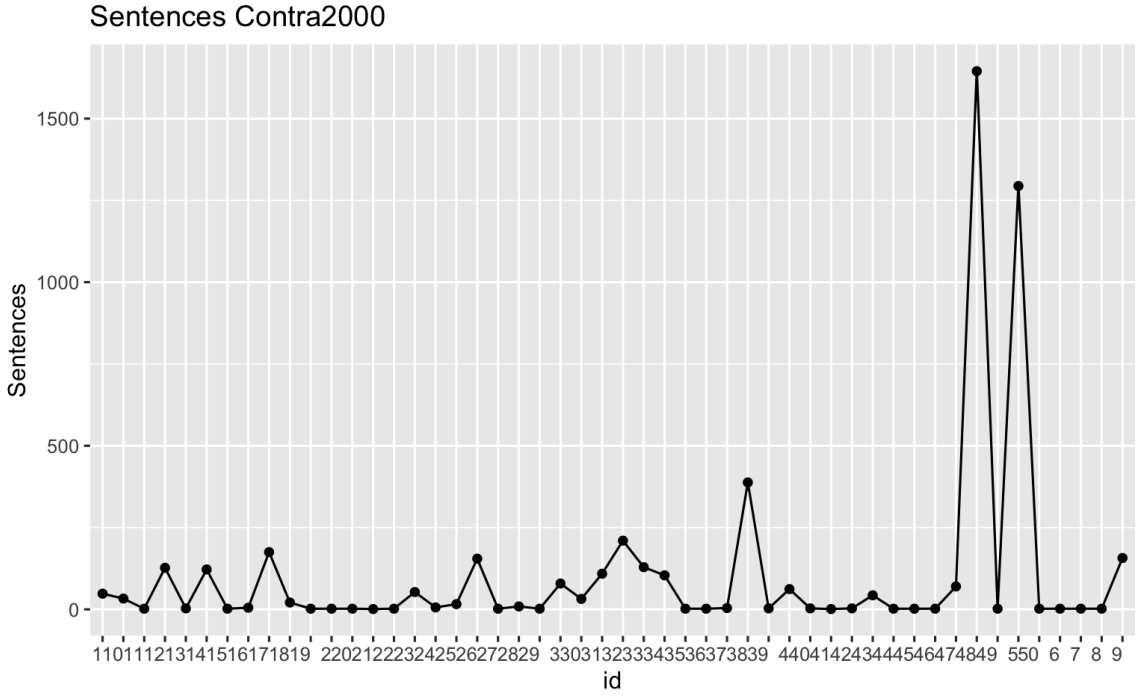

Figure 2: Number of sentences in sceptics corpus (sample of 50 texts).

4 Empirical Work

To get a deeper understanding of the corpora's content and to examine its statistics on a linguistic base, Text Mining methods were applied to the data. Over the last years Text Mining has been proven a useful tool for automatically investigating large sets of textual data. The results then have to be analyzed manually. One of the benefits of applying Text Mining methods is the quick and deep insight on textual data. As one can see in the upcoming sections, we wanted to evaluate word frequencies, check important co-occurrences and explore keywords with special regard to terms connected to climate change. Moreover, one main objective of the application of Text Mining methods to our data was to empirically validate the decision of splitting the corpus into the two groups (activists and sceptics). It will be interesting to identify and analyze differences and similarities between both groups and to investigate certain word occurrences and contextual features.

For the purpose of this project we decided on using R and the Quanteda library - as already mentioned in section 3. Hence, it was necessary to get started with using R and Quanteda. This was done by working through multiple guides[9]. The Quanteda package provided us with very easy-to-start and essential Text Mining methods.

4.1 Cleaning

In a first step, the already created corpus data was pre-processed by removing punctuation and special characters, and by lowercasing all words contained withing the texts. This was necessary for the upcoming stages of Text Mining. Thereafter, a list of stop words was loaded into R[10]. The removal of stop words is one way to improve the examination of word frequencies and to exclude words from the analysis that do not provide semantic information in a text because they are too common (such as function words, e.g. prepositions) to be significant in a frequency analysis. Subsequently, lemmatization of the dataset was performed. Lemmatization is a useful technique to create a more readable set of features. It is used to ensure that words sharing the same base word are all treated as the exact same word within frequency analyses. For instance, the word forms "went", "goes", "gone" would be reduced to their infinitive form "go". The spacyr package[11] was used for the lemmatization. Unfortunately, most German lemmatizers are still very inaccurate when it comes to the lemmatization of compound nouns. Accordingly, most of the climate change compounds that will be investigated and shown in the following plots, are not lemmatized correctly. This is one of the limitations of this project that would have to be solved in future work.

[9] https://tutorials.quanteda.io, http://inhaltsanalyse-mit-r.de/grundlagen.html

[10] As part of the "stopwords" package from R, the German version of the following list was used: `http://snowball.tartarus.org/dist/snowball_all.tgz`

[11] https://github.com/quanteda/spacyr

4.2 Frequency Analysis

4.2.1 Term Frequency

One of the Text Mining techniques applied to the data set was a frequency analysis of the terms contained within the texts. This was done to gain a deeper insight into the corpus' content and to be able to quantify what the texts are about. One important measure is the term frequency (tf) which basically determines how often a term occurs within a text or the whole corpus. Anyway, it is important to say here that the general frequency of a term is not always as informative as one would expect. A lot of words that appear in a text many times may not be important, such as function words (e.g. prepositions, determiners). Since they don't carry meaning they are often excluded from the analysis by applying stop lists to the data. Stop lists consist of stop words, i.e. words that do not add semantic information to a document. Hence, a predefined list of German stop words[12] was applied to the corpus data in this project. Since a lot of cases require the stop lists to be customized to properly fit to the individual textual data, a customization of the stop list was performed within this project[13].

Another approach to identify important words is to inspect the term's inverse document frequency (idf) which weight decreases if a term is very common and increases for terms being very special to a text. The combination of both measures (tf and idf) results in a term's tf-idf score which quantifies the term's importance with respect to the whole collection of texts.[8, p.40f.]
In Text Mining applications, tf-idf is also used to perform keyword extraction. Hereby, the most relevant words of a text are being obtained in an automated process. Keyword extraction is particularly useful for summarization and topic identification of large data sets.

Since the tf-idf is a common numerical statistic to determine most relevant words of a corpus, it is one of the measures we want to apply to the activists and sceptics corpus to pinpoint differences and similarities between both corpora.

Figure 3 shows a plot of the top 50 climate change compounds of the activists corpus. Unsurprisingly, the top position is occupied by the term "Klimaschutz" (en: "climate protection"), followed by words such as "Klimakrise" (en: "climate crisis"), and "klimapolitisch" (en: "climate political"). All these compounds definitely have a connection to the climate activists community and can be located within the supporting community of climate action.

[12] `http://snowball.tartarus.org/dist/snowball_all.tgz`
[13] The customization of the original stop list was done within the code of the R notebook file: `glossary_notebook.Rmd`

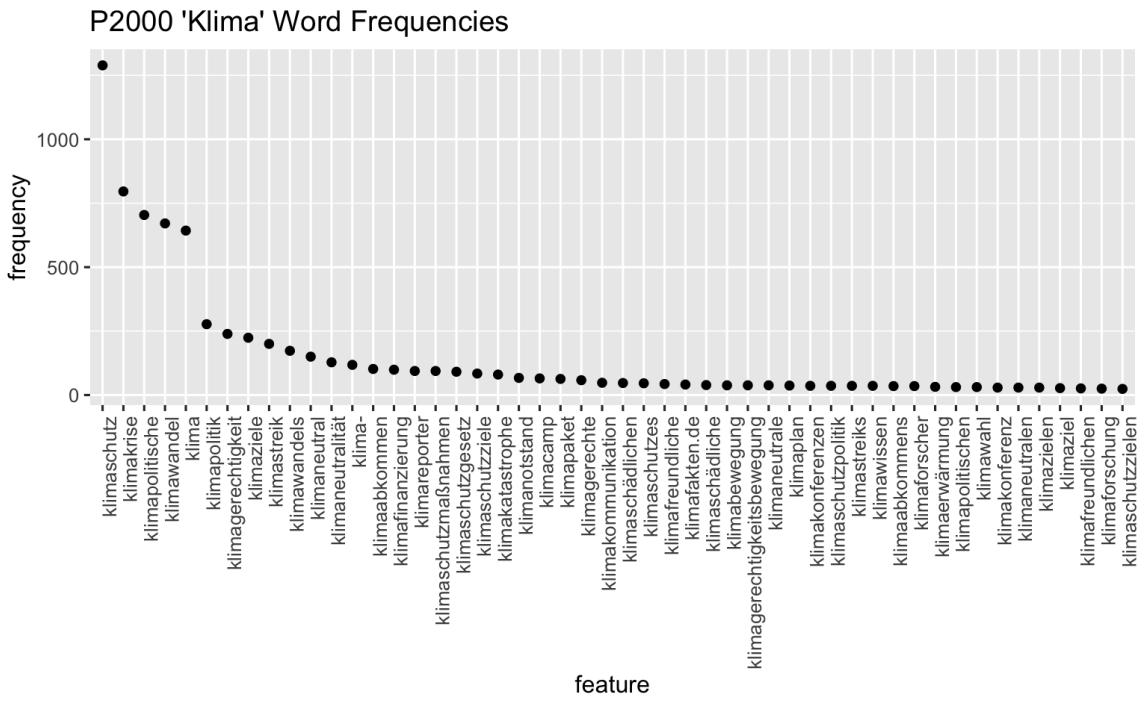

Figure 3: Top 50 "Klima" words for activists corpus.

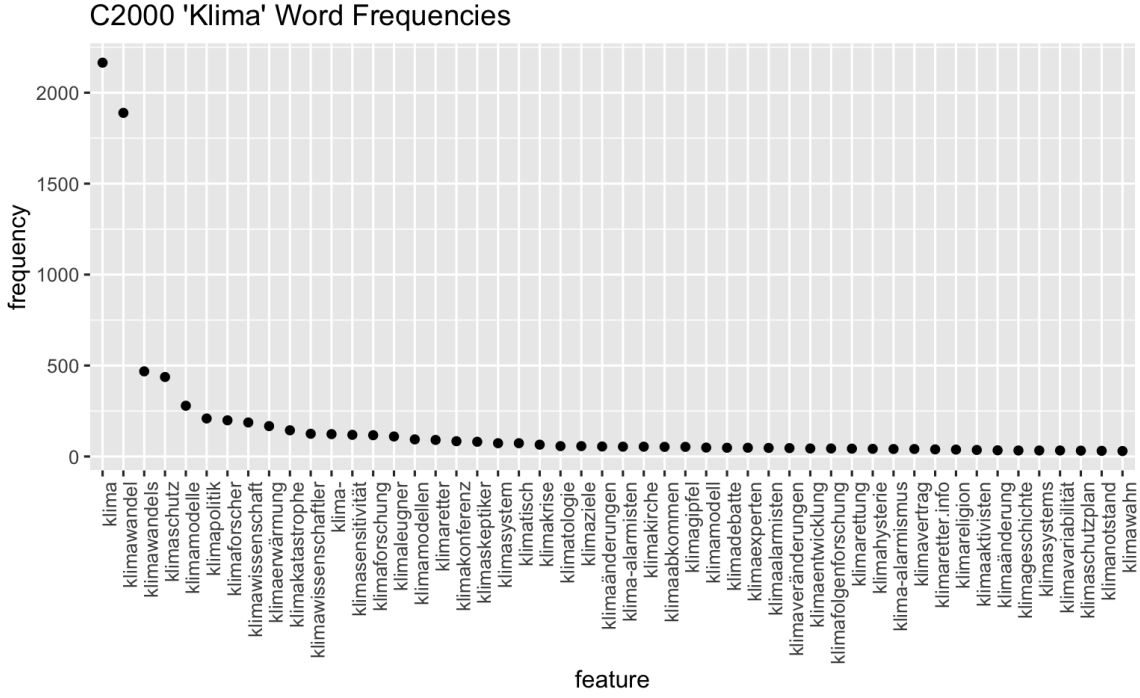

Figure 4: Top 50 "Klima" words for sceptics corpus.

Conversely, the plot of the top 50 climate change compounds of the sceptics corpus (see figure 4), reveals words such as "Klimamodelle" (en: "climate models"), "Klimawissenschaft" (en: "climate science"), and "Klimaforscher" (en: "climate researcher/scientist") within its top 10 terms. This correlates with the manual findings of section 2. A lot of texts from the climate deniers community is about explanations of climate change not being caused by humans but by nature itself, as can be proven by climate scientists.

```
 [1] "klimapolitische"          "klimagerechtigkeit"         "klimastreik"
 [4] "klimaneutral"             "klimaneutralität"           "klimafinanzierung"
 [7] "klimareporter"            "klimaschutzmaßnahmen"       "klimaschutzgesetz"
[10] "klimaschutzziele"         "klimacamp"                  "klimapaket"
[13] "klimagerechte"            "klimakommunikation"         "klimaschädlichen"
[16] "klimaschutzes"            "klimafreundliche"           "klimafakten.de"
[19] "klimaschädliche"          "klimaneutrale"              "klimagerechtigkeitsbewegung"
[22] "klimabewegung"            "klimaplan"                  "klimastreiks"
[25] "klimawissen"              "klimakonferenzen"           "klimaschutzpolitik"
[28] "klimaabkommens"           "klimapolitischen"           "klimawahl"
[31] "klimaneutralen"           "klimazielen"                "klimaziel"
[34] "klimafreundlichen"        "klimaschutzzielen"
```

Figure 5: List of "Klima" words only contained in Top 50 for activists corpus.

```
 [1] "klimamodelle"          "klimawissenschaft"     "klimawissenschaftler" "klimasensitivität"
 [5] "klimaleugner"          "klimamodellen"         "klimaretter"          "klimaskeptiker"
 [9] "klimasystem"           "klimatisch"            "klimatologie"         "klimaänderungen"
[13] "klimakirche"           "klima-alarmisten"      "klimagipfel"          "klimamodell"
[17] "klimadebatte"          "klimaexperten"         "klimaalarmisten"      "klimaveränderungen"
[21] "klimafolgenforschung" "klimaentwicklung"      "klimarettung"         "klimahysterie"
[25] "klima-alarmismus"      "klimavertrag"          "klimaretter.info"     "klimareligion"
[29] "klimaaktivisten"       "klimaänderung"         "klimageschichte"      "klimasystems"
[33] "klimavariabilität"     "klimaschutzplan"       "klimawahn"
```

Figure 6: List of "Klima" words only contained in Top 50 for sceptics corpus.

Certainly, there are terms such as "Klimaschutz" (en: "climate protection") and "Klima" (en: "climate") itself that can be found in both corpora since those are very common terms for the topic of climate change in general. Nevertheless, a closer look at both plots (figure 3 and 4) definitely reflects the use of certain words being popular for only one of the communities. Terms being used very commonly by the sceptics community are, e.g.: "Klimasensitivität" (en: "climate sensitivity"), "Klima-alarmisten" (en: "climate alarmists"), "Klimahysterie" (en: "climate hysteria"), "Klimareligion" (en: "climate religion"), and "Klimakirche" (en: "climate church"). These are mainly negatively loaded terms that also appear in our glossary. Figure 6 shows the climate change terms that could be found within the sceptics top 50 climate change compounds but not in the ones for the activists community. Conversely, figure 5 shows the same list for the terms used by the activists but not by the sceptics community (always with respect to the top 50 climate change compounds). Looking at both lists illus-

trates the impression we already got when manually creating and researching loaded climate change terms. There are certain terms that are more present in the community of climate deniers, particularly with respect to terms being loaded negatively or with an excessive connotation. Whereas the activists terms express more realistic non-overrated sentiments when looking at the words displayed in figure 5.

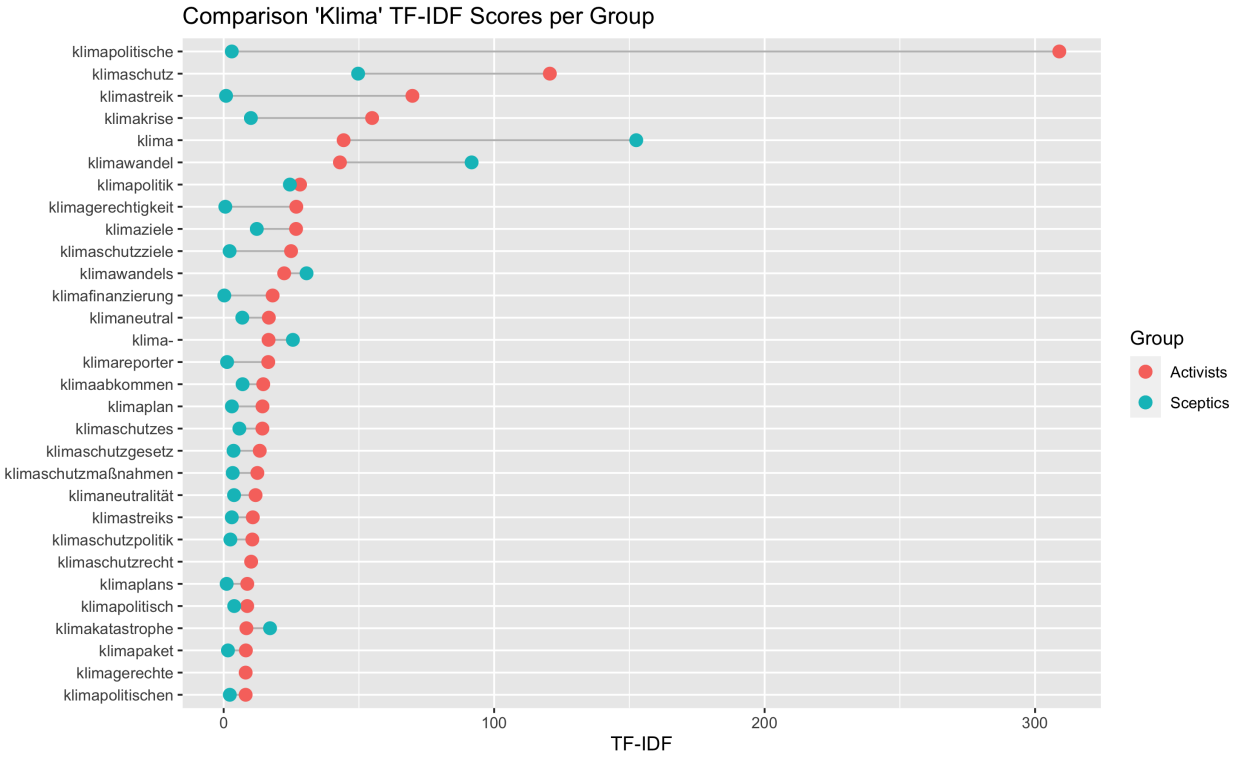

Figure 7: TF-IDF scores of "Klima" words by groups.

Additionally, tf-idf scores for both corpora were calculated to see if the results and impressions that we got from the frequency plots above could be reproduced. In fact, figure 7 showing the tf-idf scores (i.e. relative frequencies) of both corpora's climate change terms, could confirm the findings. Climate change terms connected to climate actions (i.e. "Klimaschutz", "Klimaschutzziele", "Klimaziele") and to climate awareness (i.e. "Klimakrise", "Klimastreik") are rather used by the climate activists group, whereas a loaded term such as "Klimakatastrophe" appears to be used more by the sceptics community. Figure 8 and 9 show the according tf-idf scores for both corpora separately.

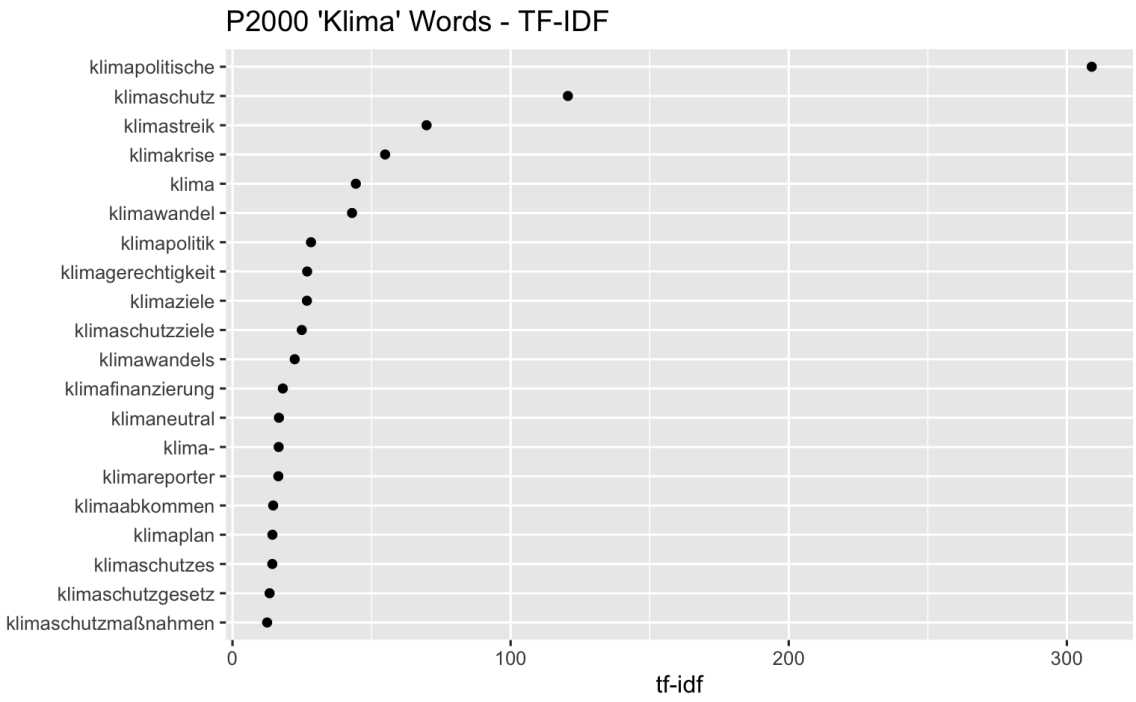

Figure 8: TF-IDF scores of "Klima" words for activists corpus.

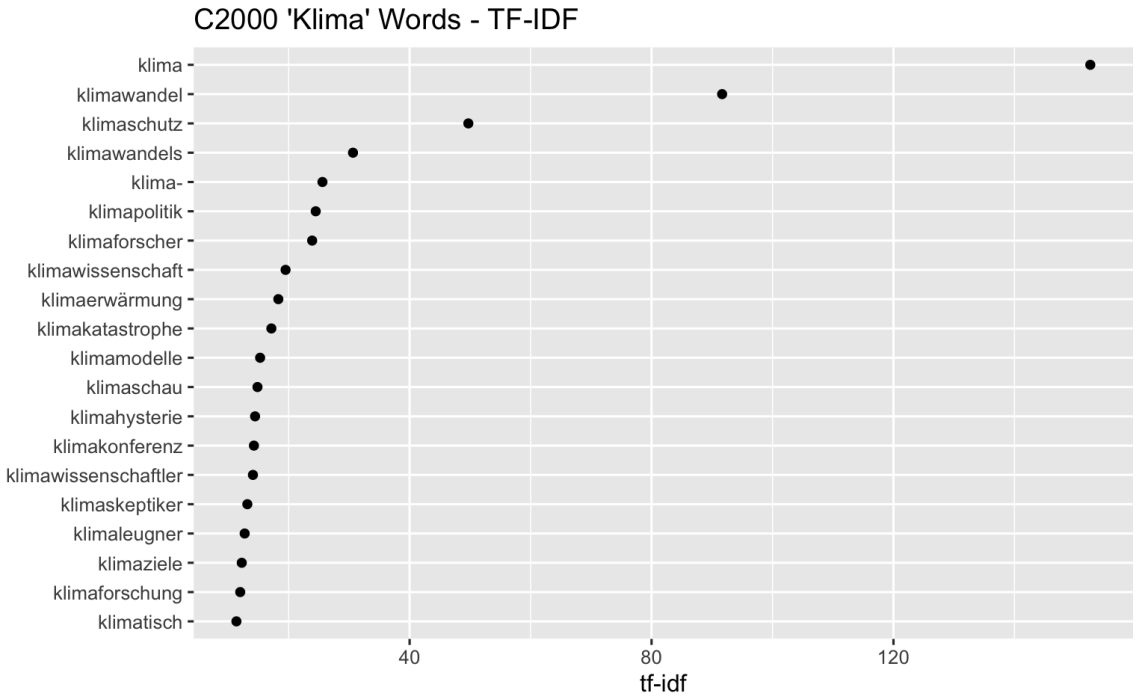

Figure 9: TF-IDF scores of "Klima" words for sceptics corpus.

4.2.2 Keyness

Keywords are words in a text that "occur with unusual frequency" [7, p.236] or in other words, terms that appear more often than one would expect. One way to identify keywords is by carrying out a simple frequency comparison. Another way, for instance, is to perform statistical tests to determine keywords. Those tests compare word frequencies against their expected frequencies (e.g. chi-squared measure[14]). Accordingly, keyness is denoted as an indicator of how statistically typical the use of a term is for a text relative to a reference text [3]. Since it is a textual feature, the keyness of a term can vary - always depending on the context of a document.
The main idea of why we determine keyness within the scope of this project is to help identifying differences in text collections, in our case the two corpora (activists vs. sceptics). If both corpora differ significantly, this could also be reflected by keyness. Quanteda calculates keyness scores by using the chi-squared measure. If the observed value in the target corpus exceeds its expected occurrence, this should manifest in a positive chi-squared value. A term with a negatively indicated chi-squared value can be found less often than expected with respect to the target corpus [5].

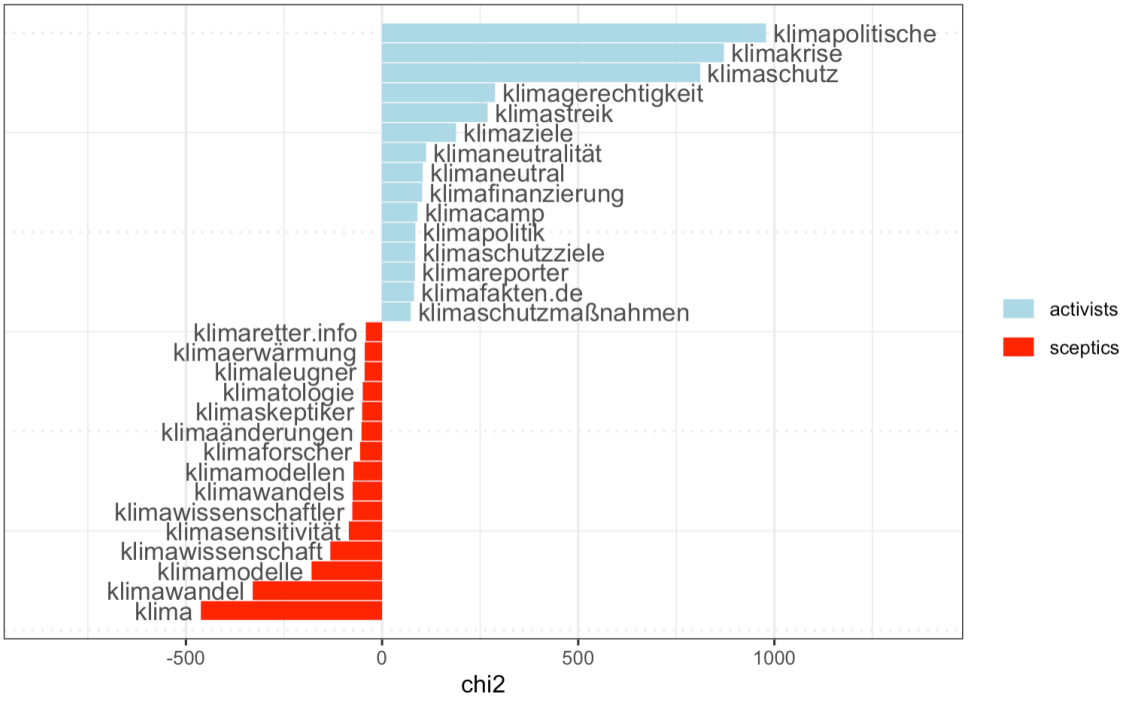

Figure 10: Comparison of keyness for "Klima" words in both groups. Activists were taken as target corpus, sceptics as reference corpus.

As we can see in figure 10 loaded terms such as "Klimaleugner", "Klimareligion" and "Klimasensi-

[14]The chi-squared measure is a method to calculate differences between observed and expected frequencies. It is often used to establish statistically significant differences between the expected and observed frequencies.

tivität" are predominantly used by the sceptics community. These terms appear with a negatively indicated chi-squared value since they are underrepresented in the activists corpus. In comparison, terms with respect to climate actions and climate awareness are rather scored positively with respect to the chi-squared value, given that their actual term frequency in the target corpus, i.e. the activists data, is higher than expected. As we already figured within this paper, loaded terms, especially the negatively loaded ones, seem to be used more often by the sceptics community. This finding could be confirmed with the keyness analysis of figure 10. On the positively scored side of this plot we can identify a lot of words strongly connected to the activist community. However, also the impression that we already got from our manual analysis could be replicated here when looking at the terms being negatively scored: the occurrence of words regarding climate change science and scientists is much higher for the sceptics group than for the activist corpus.

4.3 Context Analysis

4.3.1 Collocations

Another Text Mining method that was taken into account for the climate change data focuses on the notion of context. A well-known technique to explore the context of terms contained within textual data is the extraction of n-grams. This basically means to partition a text into word chunks to get a better understanding of the semantic content of a text. The analysis of n-grams is very useful for examining semantic and syntactic relationships and connections between words [8, p.57].

For the corpora of this project terms have been breaken down into bigrams (two-word combinations). Given that a lot of the most common bigrams in the text consist of at least one very common word (e.g. "of the", "in the") it was necessary to clean the bigram data by applying our customized stop list.

The collocation plots in figure 11 and 12 actually do not show any surprising data. While we can find mentions of climate political topics and indicators of climate change impacts in the activists corpus, there are a lot of bigrams related to renewable energies and fossil fuels in the sceptics plot. However, the collocation analysis has not proven to be as informative as expected.

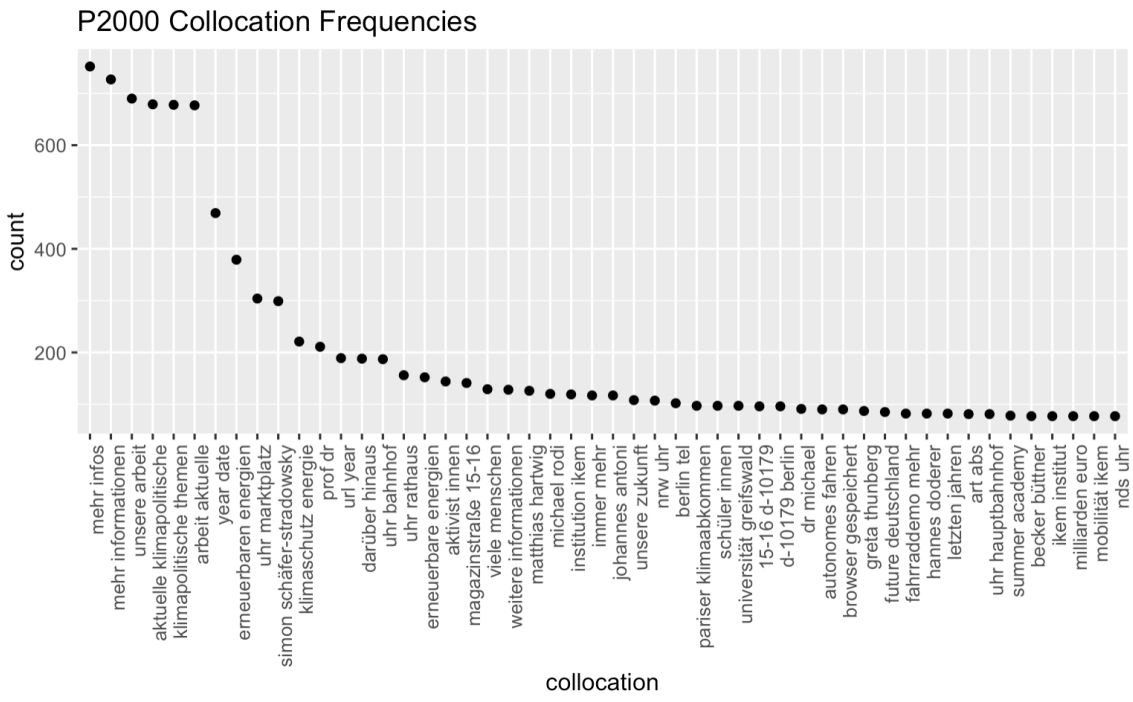

Figure 11: Collocations for activists group.

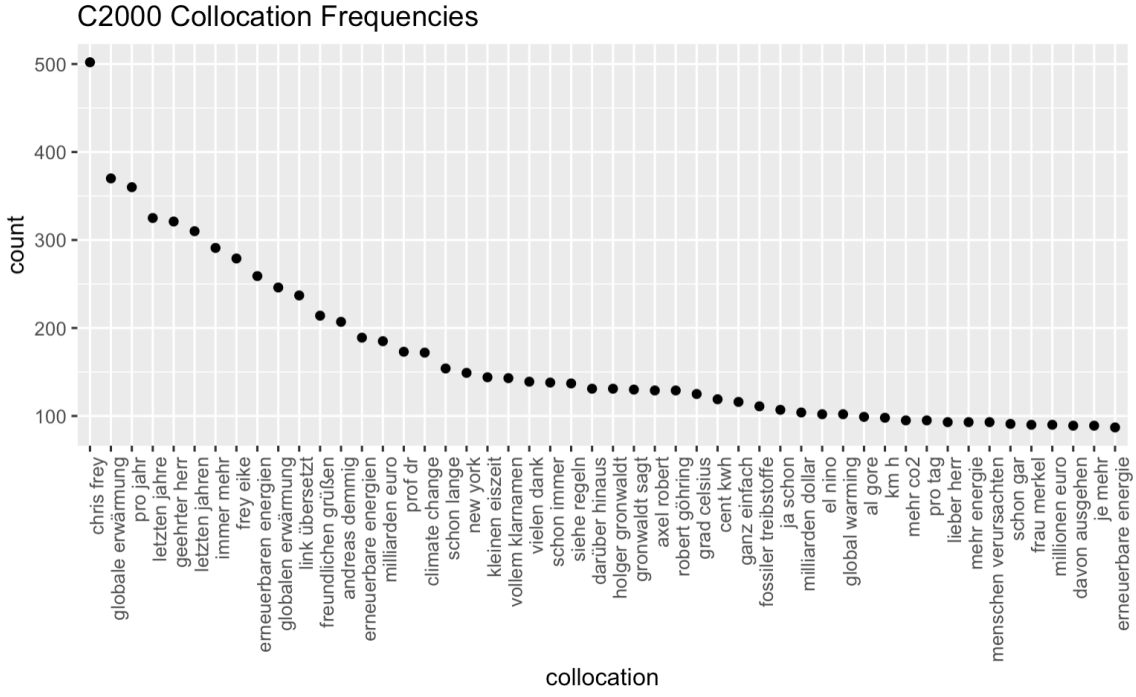

Figure 12: Collocations for sceptics group.

4.3.2 Keyword in Context

As already explained above, keywords are terms of major interest of a text, since they provide elementary information about it. Keyword-in-Context (KWIC) is an automated process to identify the context of a keyword. It basically outputs the preceding and following words of a given keyword. The analysis of KWIC was used within this project to evaluate the context of certain relevant terms, such as the loaded terms that were established within our glossary list. This section aims to investigate the context of two examples of KWIC. It would go beyond the scope of this paper to analyze more words but this technique could indeed be used to form a final glossary entry of a term and give background information about the context in which it is used. For the following analysis a random term from the glossary list was considered as a keyword and analyzed via the KWIC function of the Quanteda library. Accordingly, table 2 and 3 show a keyword with five preceding and five following words.
The context we can see in table 2 about "climate hysteria" clearly approves the impression about the term that we already gained when researching the term for the glossary list. It is used by the sceptics community as a negatively connoted word to denunciate and describe the actions of climate activists. Conversely, the term "climate crisis" in table 3 seems to be used by the activists community to invoke people to be active and aware of the climate. This is in line with the findings of the manual research for the context of the glossary terms.

	Preceding Words	**Keyword**	**Following Words**
1	wie auch die ganze	Klimahysterie	" . Es gibt keinen
2	mir nach diversen Virushysterien die	Klimahysterie	gerade rechtzeitig . So habe
3	es den führenden Protagonisten der	Klimahysterie	keineswegs um das Klima geht
4	zu nehmen . Die geschürte	Klimahysterie	mit ihrer metastatischen Durchdringung inzwischen
5	. Mit dem Siegeszug der	Klimahysterie	in den Regierungen , der
6	die Massenmedien , daß	Klimahysterie	" das Unwort des Jahres
7	dazu verbreiteten die Medien ordentlich	Klimahysterie	, indem sie von kochenden
8	ist . Es hat eine	Klimahysterie	mit religiösen Untertönen geschaffen .
9	der Erderwärmung zu ängstigen .	Klimahysterie	herrscht in China weder in
10	gerade derjenigen , welche der	Klimahysterie	und der desaströsen Energiewende das
11	wurde . Während sich die	Klimahysterie	nur langsam bewegte , in
12	Soros31 mit dem Schüren einer	Klimahysterie	Macht und Reichtum vermehren ,
13	die deutsch-österreichische Automobilindustrie aus der	Klimahysterie	Nutzen zu schlagen . VW
14	Marktwirtschaft auszusetzen . Durch die	Klimahysterie	werden langfristige Wirtschaftspläne eingeführt -
15	Das größte Interesse an der	Klimahysterie	haben jedoch die Militärs .
16	verdient ihr mit euer	Klimahysterie	" ? Bitte seid transparent
17	verdient ihr mit euer	Klimahysterie	" ? Bitte seid transparent
18	usw .) und die	Klimahysterie	hier noch lange weiterlaufen wird
19	Mortalitäten Liste von Petitionen gegen	Klimahysterie	97 % Um nun den
20	gegen die Erderwärmung " Die	Klimahysterie	beruht im Wesentlichen auf der

Table 2: Keyword-in-Context for "Klimahysterie" (sceptics corpus).

	Preceding Words	Keyword	Following Words
1	macht gerade eindrücklicher auf die	Klimakrise	aufmerksam als der Planet selbst
2	die gesellschaftliche Debatte über die	Klimakrise	dürfen nicht davon abhängen ,
3	Zeichen dafür , dass die	Klimakrise	in der Mitte der Gesellschaft
4	sich ganz zentral um die	Klimakrise	drehen wird - eine Tatsache
5	diejenigen über den Ausgang der	Klimakrise	entscheiden , die ein politisches
6	Ein Kommentar zu Die	Klimakrise	wartet nicht auf einen Impfstoff
7	dass niemand mehr an der	Klimakrise	vorbeikommen wird . Ob jung
8	in der Coronakrise eskaliert die	Klimakrise	bisher ungebremst . Wir können
9	allerdings ausreicht , um die	Klimakrise	abzuhalten , ist dann doch
10	für Klimagerechtigkeit - denn die	Klimakrise	wird im Moment auf dem
11	sich vor den Folgen der	Klimakrise	zu schützen . Dass rassistische
12	im Globalen Süden durch die	Klimakrise	stärker betroffen sind und zugleich
13	Belastung durch Umweltschäden und die	Klimakrise	. Unter anderem deshalb ,
14	den Hitzesommern , die die	Klimakrise	bringen wird , eine gefährliche
15	besonders unter den durch die	Klimakrise	heißer werdenden Sommern und Extremwetterereignissen
16	dass manche Menschen durch die	Klimakrise	stärker betroffen sind als andere
17	zu den gender-spezifischen Auswirkungen der	Klimakrise	auf nicht-binäre Menschen . Es
18	Klimafolgen leiden . Durch die	Klimakrise	wird die Anzahl von Naturkatastrophen
19	sich vor den Auswirkungen der	Klimakrise	zu schützen . Eine Studie
20	die geschlechtsspezifische Gewalt durch die	Klimakrise	zunehmen wird . Grund dafür

Table 3: Keyword-in-Context for "Klimakrise" (activists corpus).

4.4 Glossary Enrichment

In a last step results from the frequency analysis were taken to create lists of words containing German climate change compound nouns. Two lists[15] were built, one for the activists corpus and one for the sceptics corpus. This was done to be able to compare those two word lists to the primary glossary that has been constructed manually in section 2. With this step, we wanted to enrich the glossary as well as empirically validate the existence of the climate change compounds that have already been added manually.

The final glossary list contains 2967 terms (activists: 826, sceptics: 2141).

5 Conclusion

5.1 Limitations of the project

Since the project has been carried out within a limited time frame and as a proof of performance of the course "Text Mining in the Social Sciences" at the University of Potsdam there are some limitations

[15] All lists can be found in the folder `glossary_lists`

of the project which should be eliminated in future work. So far, within the scope of this project only the EIKE website was taken for creating the corpus data for the climate sceptics community. It would definitely be recommendable to create the corpus from various sources to allow for more significant findings. At the moment, we can only tell what is happening in the sceptics community from a very restricted perspective. This fact could be improved in next steps by researching more websites for the sceptics corpus and creating another random subcorpus from that data. Another aspect of the project that has led to minor errors and a lot of additional manual work is the German lemmatization provided by the spacyr package. During the evaluation of the results it became clear that the lemmatization of the German climate change compounds is not as accurate as one would wish. Accordingly, it would either be necessary to manually create a list of lemmata to improve the outcomes of frequency and context analyses or to find another lemmatization library which takes care of the compounds automatically. Furthermore, due to the time limit, no manual revision of the automatically created glossary lists has been performed. This, as well as a comparison of the manually and the automatically created list, would be a necessary step for building a final glossary.

5.2 Outlook

In future work it would be very interesting to dive deeper into the characteristics of both corpora. For instance, a Sentiment Analysis of the texts could reveal information about each communities' emotional status about the examined topic, in this case climate change. This idea could be extended to an emotion analysis which could break the two main sentiments "positive" and "negative" further down into single emotions which could co-occur with certain climate change compounds. This would add supplementary information to the glossary and could differentiate it from other already existing glossaries.

References

[1] A. Barbaresi. Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 122–131. Association for Computational Linguistics, 2021.

[2] K. Benoit, K. Watanabe, H. Wang, P. Nulty, A. Obeng, S. Müller, and A. Matsuo. Quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30):774, 2018.

[3] M. Bondi and M. Scott. *Keyness in texts*, volume 41. John Benjamins Publishing, 2010.

[4] A. Brunnengräber. Klimaskeptiker in Deutschland und ihr Kampf gegen die Energiewende. 2013.

[5] C. Gabrielatos. Keyness analysis: Nature, Metrics and Techniques. In *Corpus Approaches to Discourse*, pages 225–258. Routledge, 2018.

[6] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[7] M. Scott. PC analysis of key words—and key key words. *System*, 25(2):233–245, 1997.

[8] J. Silge and D. Robinson. *Text Mining with R: A Tidy Approach*. O'Reilly Media, Inc., 1st edition, 2017.

[9] M. Sommer, D. Rucht, S. Haunss, and S. Zajak. Fridays for Future: Profil, Entstehung und Perspektiven der Protestbewegung in Deutschland. 2019.