Social Data Science

Summer School 2019

The Danish Climate Debate

- From A Social Data Science Perspective

1 Introduction

Since the industrial revolution in the late 19. century actions of humanity have caused a prolonged heating of the Earth's atmosphere. This is mainly due to emission of greenhouse gasses and known as global warming.

Climate change on the other hand refers to an unusual change of the average weather pattern locally as well as globally (2). Over the last three decades the world have experienced a global rise in temperature of 0.43 °C while Denmark has had a temperature rise of 1.03 °C over the same period (11). This has caused floodings, droughts, cloudbursts and heat records to be beaten over and over, but the list of consequences is even longer. As Barack Obama, the former president of the United States, stated in his speech the 31th of August 2015 in Anchorage, Alaska, at The International Arctic Conference (10):

> *" ... climate change is no longer some far-off problem. It is happening here. It is happening now. Climate change is already disrupting our agriculture and ecosystems, our water and food supplies, our energy, our infrastructure, human health, human safety – now. Today. And climate change is a trend that affects all trends – economic trends, security trends. Everything will be impacted. And it becomes more dramatic with each passing year."*

With this in mind, we want to examine how the climate change has been mentioned in the media particularly in the news articles of The Danish Broadcasting Corporation (DR) and in which manner the rise in temperature over the last decade - from 2009 to 2019 - has affected both the dimension and the tone in the climate debate.

Therefore we scrape DR for articles containing the climate buzzwords 'Klima', 'Klimaforandringer', 'Klimatosse', 'CO2', Miljøaktivist' and 'Temperaturstigninger' for the period 1st of January 2009 to the 25th of August 2019. Specifically, we want to investigate how the frequency of these climate related articles has developed during this period and if there is a correlation between this development and the unusual temperature rise in Denmark over the past decade. Furthermore, we want to determine the tone in the Danish climate debate through a sentiment analysis of the scraped articles. This sentiment analysis will be executed by using the lexical approach, AFINN, in both Danish and English on the given (translated) articles. By comparing these two we search to find the limitations of the lexical approach. We will additionally discuss whether or not the DR articles containing the buzzword 'Klimatosse' has been used in a positive or negative manner during the Danish election in June 2019 and afterwards.

From our analysis we conclude that there has been an increased awareness about the climate due to a sizable increase in published articles with a climate related content from DR as well as a positive, enlarging of the overall tone in these articles. We find systematic deviations between the sentiment scores of the articles in Danish and English respectively which could be caused by methodological issues. Finally, we discover a positive correlation between temperature anomalies for Copenhagen and the number of articles with a climate content for the last decade. This indicates that the temperature anomalies and the resulting extreme weather conditions - potentially caused by climate change - increase the general public's interest for articles regarding the climate.

2 Considerations

2.1 The selected climate buzzwords

The climate buzzwords selected for this analysis are 'Klima', 'Klimaforandringer', 'Klimatosse', 'Temperaturstigninger', 'Miljøaktivist' and 'CO2'. These buzzwords are randomly chosen by us. First, we brainstormed on all climate buzzwords that came to our minds. Afterwards, to the best of our abilities, we went through the list of buzzwords making sure that we chose a broad spectrum of buzzwords from the past ten years. Also, with the sentiment analysis in mind, we chose somewhat neutral buzzwords to see how these would be represented either in a 'positive' or 'negative' way by DR, which is an objective broadcast and news institution.

2.2 Data Ethics

During the process of finding a suitable data source and topic for this paper, we used 'The Principles-based approach' (1). This approach says that we as researchers should adapt our researching process to the existing rules when for instance scraping web pages. We did consider scraping articles for our project from Danish newspaper websites such as Politiken or Berlingske. However, most of these articles were not available without a subscription. By subscribing we would agree on terms which may involve not being able to web scrape without breaking the conditions of these terms. Therefore, we chose to web scrape the website of DR which is public, meaning that it is legal to web scrape. This ensures that our research is replicable.

But just because a website such as DR is public and therefore legal to scrape, we still have to consider the principle of 'Respect for Persons' (1). This principle implies treating people and their sensitive and identifiable information properly so they remain anonymous. We handled this potential issue by using only non-personal data for this paper i.e. only the headlines, release dates and bodies of the given articles from DR and not for instance the authors' names. This is also in line with the principle of 'Respect for Law and Public Interest' (1), where we used compliance by respecting the law of for example GDPR. In addition, we used the transparency-based accountability approach by being clear about our goals for the scraping process, the methods used for analysis in this paper as well as handling our results. With this transparency approach we prevent that someone is suspecting us for doing something illegal.

When it comes to ethics things get more difficult, because ethics is not based on specific rules. This makes it harder for us as researchers to know which approach would be best to use in the data collecting process. Therefore, it requires common sense to navigate in. As a consequence, we made a risk/benefit analysis of our research project and ended up concluding that the benefits of scraping data from DR is greater than the belonging risks. This is in line with 'The principle of Beneficence' (1) as we consulted the teaching assistants of The Social Data Science course in order to secure an appropriate ethical balance and objective point of view on our research project.

3 Data Processing

3.1 The Data Collection Process

In order to answer the research questions, we scraped the articles from DR. The scraping process consists of 5 steps: Connecting, mapping, parsing, storing and logging (3). We chose to scrape articles from DR because - as mentioned above - its website is public and therefore legal to scrape. Additionally, DR has an archive for articles, which is very useful for scraping data as it makes it possible for us to extract articles from past years. We have specifically chosen articles from the 1st of January 2009 to the 25th of August 2019.

3.2 Scraping

Before we could connect to the DR server, we had to install some required third-party libraries such as 'Numpy', BeautifulSoup', 'Pandas' and 'Requests'. Installing these libraries, makes it possible for us to access the HTML content from DR's web page (HTTP) (3). We were very conscious about only scraping links for the articles on 'Nyheder' in the search box on DR's web page. Hence, this would result in a much 'cleaner' data set meaning that we would not have to drop links from for example radio programs containing some of our buzzwords.

The way DR's web page is constructed made it possible for us to build the URL using a recognizable pattern in the articles due to the fact that the URL for each article had the same structure (12). In addition, we modified the URL slightly because the web page does not have numbered pages when going through the article archive. Therefore we wrote the URL of the web page with the following 'format' code "...=Newestpagesize=10page=".format(page)" saying each search page would consist of 10 article links.

Furthermore, we created a while loop to control the data scraping process. This loop took the current year (2019) as input and scraped our article data all the way back until 2008 which insured, we would scrape articles from exactly the 1st of January 2009 to the given scraping day in 2019. We sat the rate limit (time sleep) to 0.5 second. This guaranteed that we would not send too many calls to DR's server and risk getting blocked.

After scraping the article links, we parsed through these links with the libraries 'BeautifulSoup' and 'html.parser' to scrape for our chosen, relevant variables from DR's web page. Using these libraries makes it easier for us the navigate through DR's HTML searching for either a tag, a name or an attribute (12).

Once these links were located, we could search within the hyperlinks for our desired variables (12). Specifically, we scraped for the following three variables from DR's web page: the headline, the published date and the body of a given article. Because we were only interested in the frequency of the selected buzzwords, we did not scrape variables such as word length or author. Lastly, we stored our raw data as a csv-file and kept a data log (see section 3.3). This data process was repeated six times. One for each of the buzzwords so we ended up with six data frames. Then we merged all the data frames except the one for 'Klimatosse' into one big data frame. We omitted the data for 'Klimatosse' in the joint data frame because the articles for 'Klimatosse' only occurred from 1st of January 2019 and forth. Hence, it would be impossible to show how

the sentiment or number of articles have developed for 'Klimatosse' on an annual basis from 2009 to 2019.

3.3 Data log

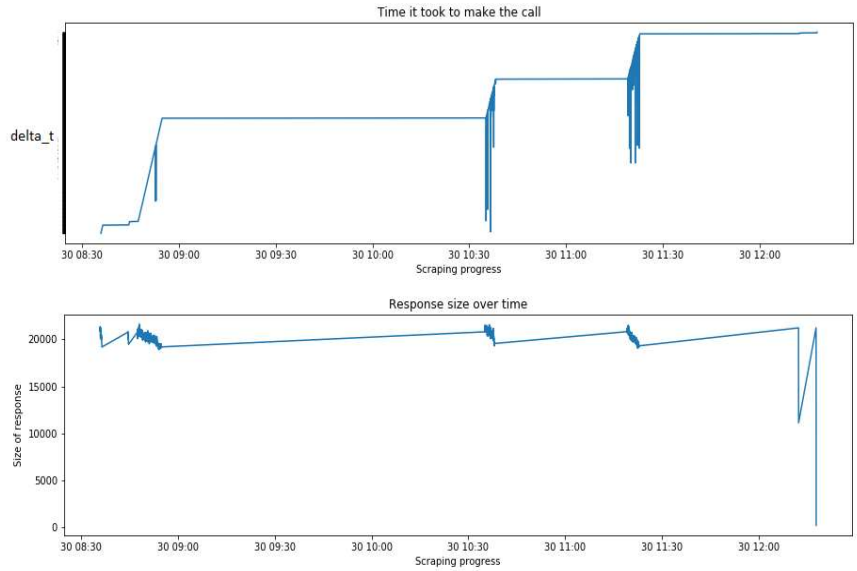
Keeping a data log is a way to document the data process. It clarifies that data is scraped objectively. The data log is a way to document the data quality and is important for a developer in order to keep track of the data process (5). Logging the data process gives other developers an understanding of the data process which insures reliability of the given developer (12).

Web scraping is still considered hacking in some cases. Therefore, having a data log makes the data process more trustworthy and transparent to keep track of the data flow and save all access and 'connecting calls' from the IP's to for example the server of DR (4). Also, it makes it easier to detect what is going on if an error occurs or if the program crashes (5). A data log makes the scraping replicable. Additionally, the data log helped us through the data process, creating an overview of our data process (4).

Going through our data log there is no data which is systematic missing (see the attached zip-file). We know it is important to store the log while scraping, but because this was our first-time web scraping, our log somehow overwrote itself when we scraped for the different buzzwords. This means that we only have the log for 'Klima' from our original scraping day, the 25th of August 2019, which is shown in figure 8.1 and 8.2 in the appendix. Therefore, we had to repeat the whole scraping process again on the 30th of August 2019 in order to create an adequate log. This log and the resulting data are only used for analyzing in this section as we are using the data, we originally scraped the 25th of August 2019 for the rest of this paper.

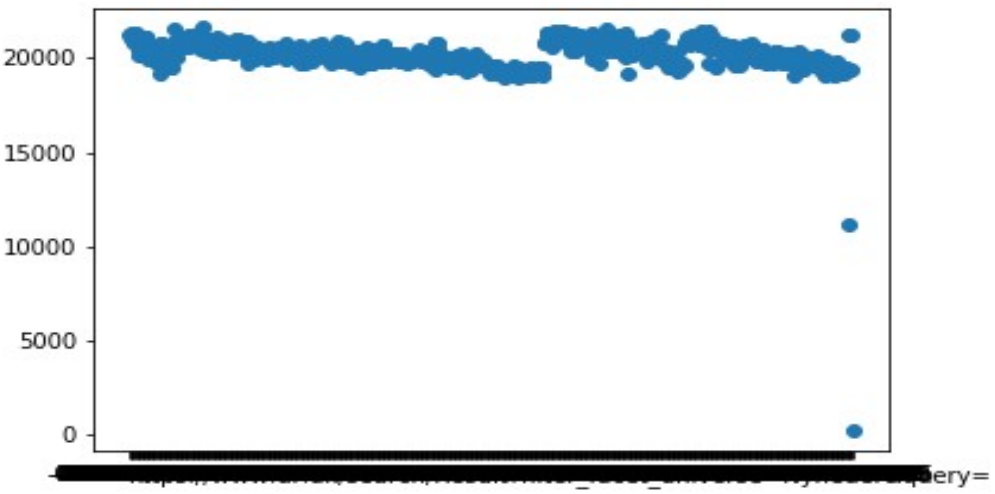
Figure 3.1 shows our scraping process from the 30th of August 2019 for all of our buzzwords. Specifically, we plot the time it took to make the call to the servers of DR as well as the size of the response over time. Unfortunately, Python kept making the y-axis in the top figure and the x-axis in the bottom figure black. Most likely it is due to a formatting issue in Python. But we notice that the black axes as well as the appearance of the figures correspond to the ones shown in our lecture and therefore, we assume that the plots are correct despite the potential formatting issue (12).

Figure 3.1: Plots for whole scraping process



From figure 3.1 we find that it took around 4 hours to scrape the entire data set. The call time process looks decent and has an increasing trend even though it is divided into three stages or curious peeks. The size of the response time plotted in the lower part of figure 3.1 is approximately 20 KB throughout the whole period disregarding the last peek, which is heading vertically towards zero. This could be an indication of large responses in the scraping process which perhaps is a consequence of us having re-cached an asset-heavy page after the program was restarted (7). The dramatic drop in the response size could be caused by us having stopped caching the page. This could explain why the response size dropped in the end of the process as it is close to 0 KB (7).

Figure 3.2: Plot of δ_t against $response_{size}$ for whole scraping process



As anticipated, there is a significantly correlation between δ_t and the size of the response in figure 3.2 as there are only 3 potential outliers (12). We therefore conclude that our web scraping process was executed with a satisfying outcome.

3.4 Cleaning the data

After extracting the articles, it was obvious that there were missing values. We deleted 4 rows from the 'Klimaforandringer' data frame, 6 rows from the 'Klima' data frame, 3 rows from the 'CO2' data frame, 2 rows from the 'Temperaturstigninger' data frame and lastly, we dropped 1 row from the 'Miljøaktivist' data frame. These missing values could appear because some of the links were constructed in a different way after all. In total we dropped 16 rows from our merged data frame.

3.4.1 The final data set

With the cleaning process done we still had 7702 articles left to analyze. The number of articles for the given buzzwords varies, as does the frequency of these. Through the analysis of the data we ended up with the following columns in our merged data frame: index, title, date, body, year, buzzword, sentiment (Danish), body in English, sentiment English and lastly, the difference in the sentiment scores. This means that our final merged data frame consist of 7684 listings and in addition we have 18 listings in the data frame for 'Klimatosse'.

For visualization of the figures we used libraries in Python such as 'Matplotlib', 'Seaborn', "Numpy' and 'Pandas'. The data we extracted was represented with a multi-index so we had to tidy it, using the 'groupby' function, in order to make the plots. The resulting figures will follow in the analysis below in section 4.

3.5 Sentiment analysis

For the sentiment analysis we used a lexicon-based approach, i.e. an unsupervised technique, to investigate the tone in the scraped articles from DR. The lexicon-based method is a type of classification that is used to construct a sentiment lexicon (8). By applying the sentiment lexicon to a given string it determines if the string in total has a positive or negative polarity (6). The values attached to each word in the lexicons are predetermined. Generally, the goal for the sentiment analysis is to gauge the attitude and emotions of a text based on the computational treatment of subjectivity (6).

Practically, we first transformed the text bodies of the articles into strings. Then we applied the lexicon to these strings. The lexicon attaches a sentiment score to each word in the string and assign each article string with a total sentiment score (8). E.g. if an article contains more words with a negative sentiment score than a positive sentiment score it is determined as having a total negative sentiment score.

3.5.1 AFINN lexicon

The AFINN lexicon was originally created by Finn Årup Nielsen, a professor at DTU, with the purpose of analyzing the sentiment in micro blogs such as Twitter (9). The AFINN lexicon consist of a vocabulary with 2477 unique words all assigned with a score between -5 (very negative) to +5 (very positive). (9) The words are assigned manually by the Finn Årup Nielsen himself.

We are using the AFINN lexicon because it is the only lexicon available in Danish as well as English. Even though it was created to read tweets, with a limit of 140 to 280 character, the articles from DR are still short compared to the length of an article from any proper newspaper.

4 Analysis

4.1 Analysis of the climate debate based on articles from DR

We start our analysis by focusing on DR's articles on the climate buzzwords 'CO2', 'Klima', 'Klimaforandringer', 'Miljøaktivist' and 'Temperaturstigninger'. More specifically we investigate how the number of articles in total and for the specific climate buzzwords have developed over the past decade.

Figure 4.1 shows the aggregated number of published DR articles on the given buzzwords from 2009 to 2019. We have extrapolated the data for 2019, because we only scraped articles until the 25th of August 2019, in order to compare the number of published articles in 2019 with the remaining years. The extrapolated number of articles in 2019 is obtained by the following calculation, where we use that the scraping day, the 25th of August 2019, is day number 237 of 2019 and the number of released articles until this date is 1103:

$$\text{Number of articles in 2019 - extrapolated} = \frac{\text{Number of days in 2019}}{\text{Scraping day as day number of 2019}} \cdot \text{Summarized number of articles on the scraping day} \Longrightarrow$$

$$\text{Number of articles in 2019 - extrapolated} = \frac{365}{237} \cdot 1103 = 1698,71 \approx 1699 \text{ articles}$$

Note, that we only extrapolate for this figure. For the rest of this paper we use the raw data as it is.

According to figure 4.1 the extent of articles with a climate focus given the climate buzzwords has increased from 16 in 2009 to 1698 in 2019. This remarkable increase is especially caused by the large increases of around 400 articles per year from respectively 2009 to 2010, 2012 to 2013 and from 2017 to 2019. Although the development has an upwards trend for the period it has been uneven with decreases in the number of published articles between -13 to -121 in both 2011, 2012, 2015 and 2016 respectively. We note as well that within the last two years the number of articles has almost doubled from 882 published articles in 2017 to 1698 in 2019, corresponding to a 92.5 pct. increase for the period. This indicates that the focus on climate in the news articles of DR given our buzzwords has enlarged substantially since 2017 ceteris paribus.

Figure 4.1: Articles from DR on selected climate buzzwords 2019 - extrapolated, 2009-2019

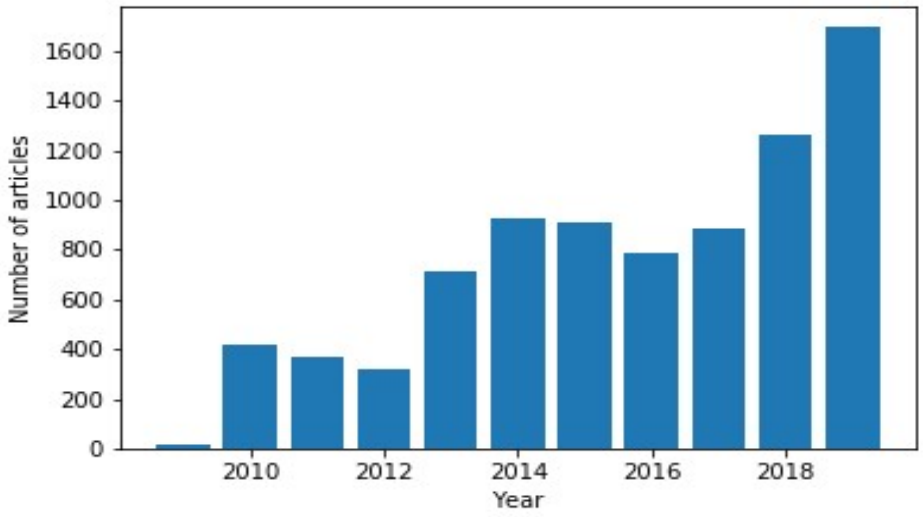


Figure 4.2 divides the overall development of the buzzwords from figure 4.1 into the specific news article development for each buzzword. Since we only have data for 2019 until the 25th of August 2019, we disregard 2019 in the following analysis. In 2009 the 16 articles on the climate buzzwords included only articles for 'CO2', 'Klimaforandringer' and 'Klima'. The first articles on 'Miljøaktivist' and 'Temperaturstigninger' appeared in 2012 and 2010 respectively, but in a very small scale as well. The general trend from 2009 to 2018 for the buzzwords has been positive, where especially 'Klimaforandringer', 'CO2' and 'Klima' has attached a large amount of attention in the news articles of DR, as they cover 94.5 pct. of the buzzword articles in 2018.

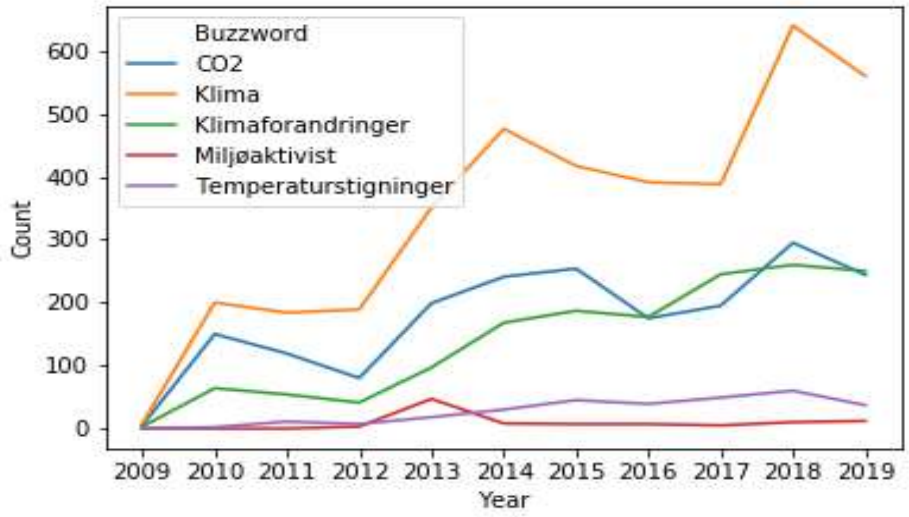
The development pattern of articles covering the buzzwords 'CO2', 'Klima' and 'Klimaforandringer' has mainly been the same with increases during the periods 2009-10, 2012-15 and 2016-18 while there has been decreases in the remaining subperiods. It is worth noticing that the absolute value of annual decreases in volume of the articles has been much smaller relatively to the annual increases. Since 2016 these three

buzzwords have experienced an increase in media exposure between 47 and 69 pct. Just below half of the articles published in 2018 are about 'Klima', corresponding to 640 articles, while articles regarding 'CO2' and 'Klimaforandringer' make up around one fifth each.

Articles about 'Temperaturstigninger' has experienced a gradual rise from 2 articles in 2010 to 60 articles in 2018, but in spite of that the buzzword only makes up less than 5 pct. of the articles in 2018. Additionally, there has on average been 12.4 articles on 'Miljøaktivist' each year for the period 2012-2018 with a potential outlier of 47 articles in 2013.

The potential outlier may be explained by the focus on the arrest of the Greenpeace activist Anne Mie Roer Jensen, among 29 other activists, on the 19th of September 2013. They had demonstrated against the Russian oil drillings in the Arctic Ocean. Anne Mie Roer Jensen was charged for piracy, which can give up until 15 years of prison in Russia. Therefore, the increase of written articles in 2013 on the word 'Miljøaktivist' may not be an expression of an increasing interest in the climate debate but simply because of one single event that caused a huge media exposure. Hence, the increased number of articles containing the buzzword in 2013 (14).

Figure 4.2: Published articles of the climate buzzwords, 2009-2019



4.2 Sentiment analysis

In this sentiment analysis we look at the entire scraping period from 2009 to 2019. Even though we only have data until the 25th of August for 2019. We assume that the average sentiment scores for the given buzzwords applies for the entire year.

From figure 4.3 we find that the sentiment scores for all the buzzwords systematically are more positive when using the English sentiment score in respect to the Danish one, except for the buzzword 'Miljøaktivist'. Here the English sentiment score is lower than the Danish for the entire period apart from 2018, where the English

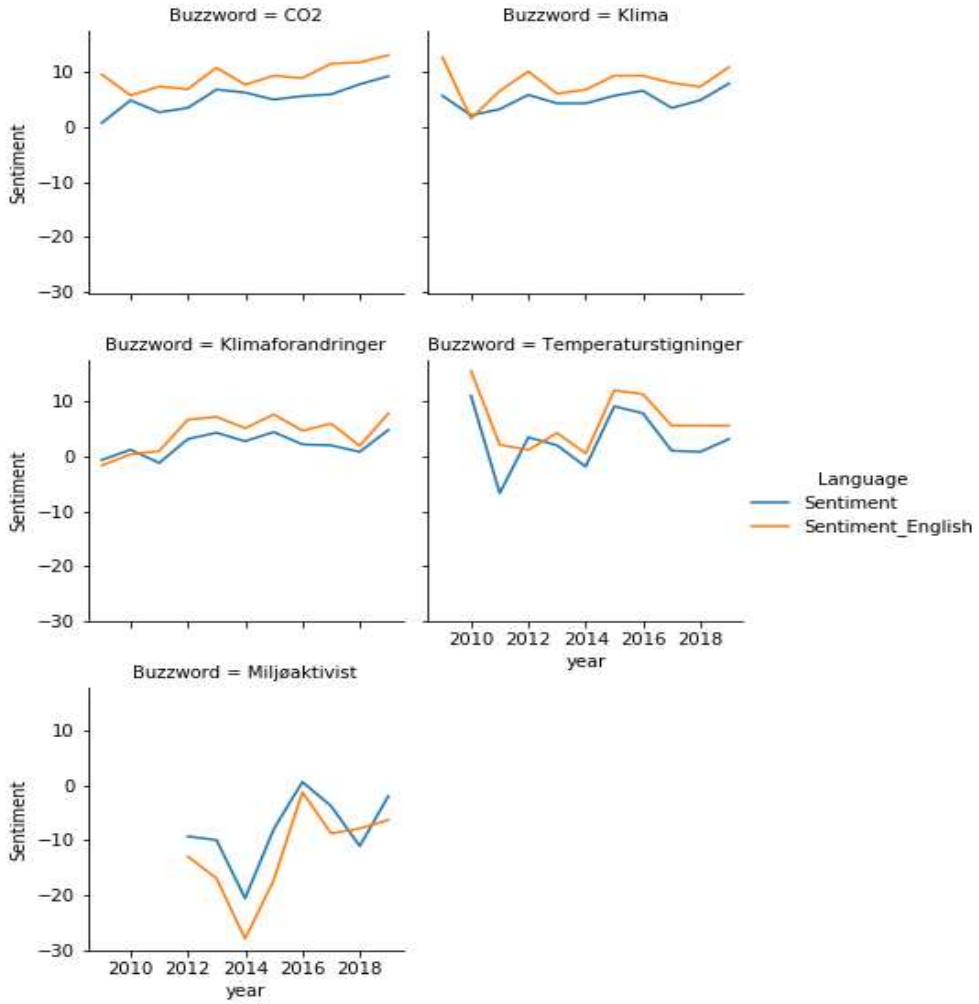
sentiment score is five score points below the Danish.

The overall trend for articles including the buzzwords 'CO2', 'Klimaforandringer' and 'Klima' have been slightly positive with an average Danish sentiment score of 6.3, 2.7 and 5.2 respectively. For 'CO2' this trend could be explained by an increasing focus on how to reach a certain emission goal, where the goal has changed over the period. 'Klima' is a rather neutral word in the context we are considering it. It could also refer to a word such as indoor climate which is not the context we are looking for. Therefore, it does not seem surprising that the average Danish sentiment score for the whole period is positive as this buzzword can refer to other contexts as well. The buzzword 'Klimaforandringer' has the lowest average, positive sentiment score of the three mentioned buzzwords which may be because this buzzword has been put in a slightly more negative context in the media in the period than the two other buzzwords that are somewhat more neutral.

The buzzword 'Temperaturstigninger' has had an overall slightly negative trend for the entire period. However, with some fluctuations from 2010-2011. These fluctuations were decreasing from an average Danish sentiment score of 11 to -6,7, which is a drop by almost 18 score points. In the period 2011-2013 the trend was increasing, but with single a drop in 2014. In 2015 the average sentiment score again increased by almost 11 score points relative to 2014. From 2016 until 2019 the trend was decreasing, however almost stagnant from 2017. One possible explanation for the trend being this volatile could be that the buzzword 'Temperaturstigninger' does not include that many articles as described in section 4.1. This in turn can have the effect of the scores on average fluctuates more than if you had an average of more articles. Hence, a very large or a very low average score will have a larger effect on average. Then if there was a small number of articles. We were expecting the average sentiment score to be negative in the most recent periods as 'Temperaturstigninger' from our knowledge has been put in a negative context. However, it is interesting that the average is decreasing at the same time there has been written more articles including the 'buzzword'.

The first thing to notice when examining the development in the buzzword 'Miljøaktivist' is that there are no observations and hence articles, until 2012. The average Danish sentiment score in 2012 is -9,3 and decreases to -20,6 in 2014. This is a drop of around 11,3 score points. This may be explained as described in section 4.1 by the publicity Anne Mie Roer Jensen got when she was accused for piracy in Russia. From 2014 to 2016 the sentiment increases with 21,2 score points to a slightly positive average sentiment score of 0,6. Hereafter, the trend is decreasing once again until 2018 where the trend has a turning point and increases continuously. In the period from 2016 to 2019 the sentiment score is negative. Hence, 'Miljøaktivist' is the only buzzword where the average Danish sentiment score is negative (almost) throughout the entire period.

Figure 4.3: Average Danish and English sentiment score of the climate buzzwords, 2009-2019



4.3 The Sentiment Analysis of the buzzword 'Klimatosse'

We have specifically chosen the buzzword 'Klimatosse' for a more qualitative sentiment analysis. The word 'Klimatosse' can be translated as a climate freak, describing a person who cares (too much) about the climate. It has had a remarkable impact on the tone of the Danish general election in 2019 and therefore it has had an impact on the tone in the Danish debate climate now. However, this word did not occur in Danish articles or the debate until May 2019. Therefore, figure 4.4 shows the average sentiment score for 'Klimatosse' from week 21 to week 33 in 2019. Further, it indicates the week where the EP election took place as well as the Danish general election. 'Klimatosse' was used at Danish People Party's EP election party by the (former) chairman of the Danish Parliament, Pia Kjærsgaard (15). However, Kjærsgaard's use of the word was not received positively by the public. Interestingly, it ended up being used in a positive matter as the people who were concerned about the climate used it to promote themselves as a 'Klimatosse'. They took back the 'ownership' of the word and turned it into a positive thing to be a 'Klimatosse'. Thereby they changed the

tone evolving around the buzzword.

Danish People Party's EP election party took place the 26th of May 2019 in week 21 which is the week our analysis begins (16). In table 4.1 we have replicated the average sentiment for 'klimatosse' and the number of articles in the given week.

Table 4.1 shows that in the two weeks up until the general election the word 'Klimatosse' gets a slightly more positive score as illustrated in figure 4.4. This is the weeks from the EP election to the Danish general election. One general observation of the data is, that in the week 22 and 23 there was brought a lot more of attention to the word. This is shown as there was written 13 articles in this period as opposed to week 24-33 (which also is a longer period of time). Opposite to the previous week, where there was just written 4 articles. This can also be a reason for some of the rather substantial fluctuations in the trend as these average sentiment scores are based on only one article while the articles in week 22-23 is based on an average of more articles. Hence, in week 22-23 there could have been both positive and negative perspectives on the word 'Klimatosse' which on average is slightly positive.

The week after the election (week 24) there was only one article written which had a sentiment score of -22. This makes the trend decrease substantially. This decrease may illustrate the public's reaction to Danish People Party's rather bad election result which could have been affected by Pia Kjærsgaard's statement of 'Klimatosse' in the context described above. From week 25 to 33 the trend of the average sentiment score is slightly increasing. This could be explained by the focus from the general election and by the word 'Klimatosse' being used in a rather positive manner in the following weeks. Now the word 'Klimatosse' is used by the general Danish population as a positively charged word in a relatively positive matter which is underlined by figure 4.4.

Figure 4.4: Average sentiment for klimatosse, 2019

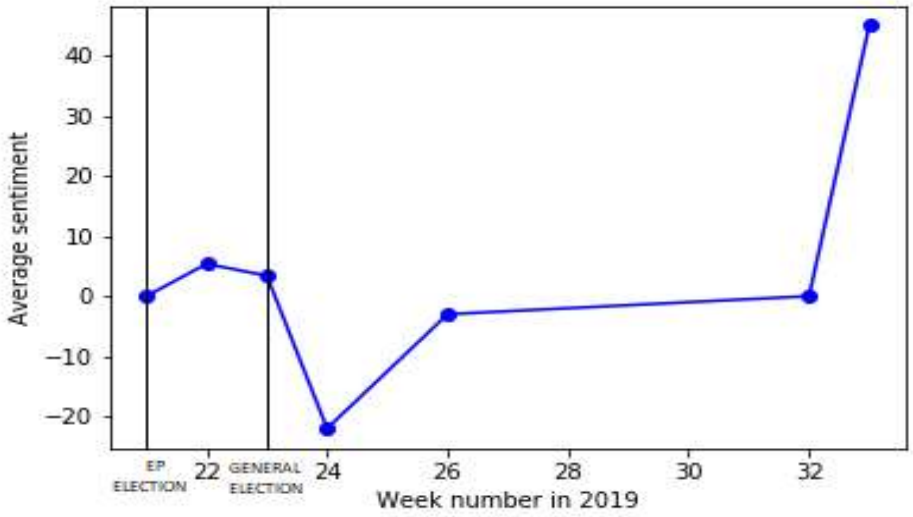


Table 4.1

Average sentiment for 'klimatosse', 2019

Week number	Sentiment score	Number of articles
21	0,0	1
22	5,3	8
23	3,4	5
24	-22,0	1
26	-3,0	1
32	0,0	1
33	45,0	1

5 Discussion

5.1 Self-critique – the climate buzzwords

For this paper we chose the climate buzzwords by brainstorming, i.e. we chose the climate related words that were clearest in our minds. Hence, the words we thought would be best for describing and analyzing the climate tendency in a neutral manner.

The approach could be criticized for creating bias because by choosing climate buzzwords this way we risked choosing words that were hot topics in the media at the moment. However, other relevant buzzwords may have been hot topics in the media earlier in the inspected period of time but these we did not remember when we were brainstorming. So, our analysis is obviously extremely dependent on our chosen buzzword and a way to solve the potential bias problem could be by using machine learning to choose the buzzwords in a more objective way. But given our time frame we chose the used approach to narrow our analysis and, in that way, make a manageable project.

5.2 Temperature anomalies and the number of DR articles with climate content

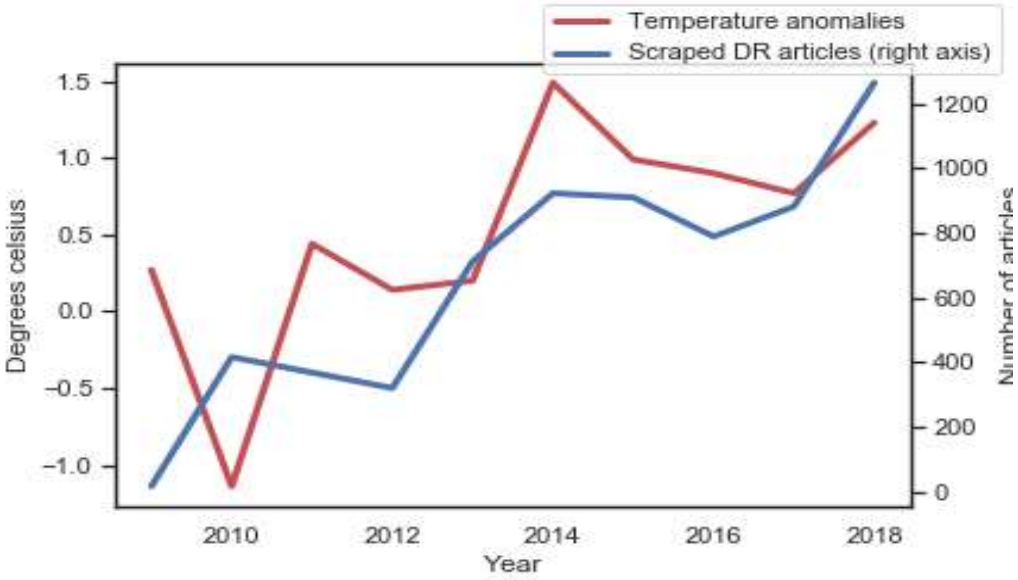
In order to investigate if there is a correlation between temperature anomalies for Copenhagen and the amount of published DR articles regarding the climate buzzwords, we use data from NOAA (The National Oceanic and Atmospheric Administration) (13). We chose to look at temperature anomalies for Copenhagen because this is where the head quarter of DR is situated. The temperature anomalies for 2009 to 2018 are calculated on an annual time basis relative to the base period from 1981 to 2010.

Figure 5.1 shows that the temperature anomalies for Copenhagen on an annual basis have been positive for

the entire period with an average of 0,53 °C per year except for 2010 where the temperature was -0,14 °C below the average of the base period. The number of scraped climate related articles from DR corresponds to figure 4.1 in the analysis above. When comparing these two graphs it seems that they follow the same development pattern with an upwards trend since 2012.

Overall figure 5.1 therefore indicates that there since 2012 has been a positive correlation between temperature anomalies in Copenhagen and the volume of articles published by DR with a climate related content. Hence, the larger (positive) temperature deviation from the base period average the more articles from DR with a focus on climate have been published.

Figure 5.1: Temperature anomalies for Copenhagen and number of scraped articles from DR, 2009-2018



Source: NOAA: "Climate at a Glance - Global Time Series", https://www.ncdc.noaa.gov/cag/global/time-series/55.676,12.568337/land$_{o}cean/ann$/7/1910 − 2019.

5.3 Methodological issues regarding the AFINN lexicon

On a different note, it is also worth discussing that there is a potential uncertainty in the way our articles have been translated. We used the translation in Google sheets, i.e. Google Translate, and the translation was not perfect. Furthermore, the translation tool had its difficulties with translation of the Danish letters 'æ', 'ø' and 'å'. However, we assessed that the translator was a perfectly fine computational tool for translating our 7684 articles given the time frame. If we had more time another translation approach might have been to either get a proper translator or to import the translation library in Python.

With that in mind, it is still interesting to see the different sentiment analysis plotted against each other for comparison as in figure 4.3. As mentioned in section 4 the English sentiment scores are constantly rating the tone of the articles more positive than the Danish one, except for the articles regarding 'Miljøaktivist'. The explanation may be that the English AFINN lexicon is a better sentiment lexicon than the Danish one

because it is more refined. Otherwise, it could be due to mistranslation. This would result in the AFINN lexicon not understanding the context of the articles and therefore misjudge the tones of the articles. The potential mistranslation could also explain why the Danish sentiment scores consistently rated the articles concerning 'Miljøaktivist' more positive than the English sentiment score.

5.3.1 Other Lexicons

As interesting it was to see the (in)difference in the sentiment analysis comparison between the Danish and the English AFINN lexicons just as interesting would it have been to use a different kind of English lexicon. The AFINN lexicons are very limited because they only assigned a value to each word and do not comprehend figurative speech (6). So, to get a better understanding of the tone in the article it could be interesting to use a different English lexicon - like 'Vader'.

'Vader' is a rule-based lexicon. It has a 'deeper' level of understanding text in comparison to the AFINN lexicon. For instance, 'Vader' understands typical negations and use of contractions as negations as well as it understands degree modifiers to alter sentiment intensity which is opposed to the AFINN lexicon (17). 'Vader' can also detect differences in all caps and emoticons. The latter would not be relevant for scraping articles on the web page of DR, but maybe for scraping micro blogs like Twitter (see further research).

5.4 Supervised approach to sentiment analysis

On further point, in order to improve the performance of our sentiment analysis we could combine a lexicion-based method as AFINN with a supervised method like Machine Learning (8). Or if we had more time at hand, we could make our own supervised model. This would require us to get training data in order to classify our articles. Maybe this would have improved our classification of the articles and thereby also the results obtained from the analysis.

6 Further Research

Another approach that could be interesting to consider is to perform an equivalent sentiment analysis on Twitter posts. Twitter is a platform where buzzwords are used and 'literal movements' are created. Therefore, it could be intriguing to investigate how an equivalent sentiment analysis of Twitter would differ from the one we made for DR. Especially because the articles from DR can refer to Twitter posts and the hashtags used there. In this way we would be able to cover both the platform where news is distributed i.e. DR, but also the platform where the news refers to certain buzzwords i.e. Twitter. Another interesting aspect is that DR is a provider of news and has to be somewhat neutral in their way of using different buzzwords. On Twitter people can use almost every buzzword in exactly the context that they want. Hence, the actors on Twitter can make their statement as dramatic as they want.

7 Conclusion

In this paper we examined how the climate change has been mentioned and framed in the news articles of DR considering the period from 2009 to 2019. We found that there has been a significant increase in articles with a climate related content from 16 articles in 2009 to 1698 articles in 2019. Further we found that there is a positive correlation between temperature anomalies for Copenhagen and the published amount of DR articles regarding the climate since 2012. This indicates that temperature anomalies and the resulting extreme weather - possibly caused by climate changes - awakens the public's interest for climate in general.

Our sentiment analysis showed that articles regarding all of the chosen climate buzzwords except 'Miljøaktivist' on average had a positive sentiment score while 'Miljøaktivist' on average had a negative score based on the whole period considered. The English sentiment score was systematically more positive for the buzzwords than the Danish one except for the buzzword 'Miljøaktivist'. Here the average English sentiment score was below the Danish for the entire period apart from one year. These differences might be due to mistranslation of the articles from Danish to English or the fact that the AFINN lexicon has its limitations when it comes to e.g. understanding figurative speech and negations.

Finally, we found from our sentiment analysis of the buzzword 'Klimatosse' that the tone in the news went from being slightly negative around the period of the Danish general election 2019 to being slightly positive in the following weeks. This change might reflect that the climate concerned people took ownership of the word and turned it into something positive to be a 'Klimatosse'.

References

[1] Salganik, Matthew J. (2018). *Bit By Bit: Social Research in the Digital Age*. Princeton University Press.

[2] NASA, *Overview: Weather, Global Warming and Climate Change*, https://climate.nasa.gov/resources/global-warming-vs-climate-change/?fbclid=IwAR2SylPm9fNghDblLMoEzywfqIuq8KKHasD6s3yrJwRODaF145JiPjc4wSA (Last assessed: 26-08-2019)

[3] GeeksforGeeks, *Implementing Web Scraping in Python with BeautifulSoup*, https://www.geeksforgeeks.org/implementing-web-scraping-python-beautiful-soup/ (Last assessed: 25-08-2019)

[4] Real Python, *Logging in Python*, https://realpython.com/python-logging/ (Last assessed: 24-08-2019)

[5] GeeksforGeeks, *Logging in Python*, https://www.geeksforgeeks.org/logging-in-python/ (Last assessed: 24-08-2019)

[6] Medium, *What really is sentiment analysis and how does it work ?*, https://medium.com/@northof41/what-really-is-sentiment-analysis-and-how-does-it-work-c812b962f643 (Last assessed: 28-08-2019)

[7] LogDna, *How to visualize your log data* https://logdna.com/blog/how-to-visualize-your-log-data/ (Last assessed: 30-08-2019)

[8] JOURNAL OF INFORMATION, KNOWLEDGE AND RESEARCH IN COMPUTER ENGINEERING, *A COMPARATIVE STUDY OF SENTIMENT ANALYSIS TECHNIQUES*, https://pdfs.semanticscholar.org/3f10/b006bab60c7f363bc03e72ad405d264b8d42.pdf (Last assessed: 28-08-2019)

[9] DTU, *A new ANEW: Evaluation of a word list for sentiment analysis in microblogs*, http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/6006/pdf/imm6006.pdf (Last assessed: 30-08-2019)

[10] The White House, *Remarks by the President at the GLACIER Conference – Anchorage, AK*, https://obamawhitehouse.archives.gov/the-press-office/2015/09/01/remarks-president-glacier-conference-anchorage-ak (Last assessed: 30-08-2019)

[11] TV2, *Kraftig temperaturstigning i Danmark - Peter Tanev kalder tal for 'skræmmende'*, https://vejr.tv2.dk/2019-04-02-kraftig-temperaturstigning-i-danmark-peter-tanev-kalder-tal-for-skraemmende (Last assessed: 26-08-2019)

[12] Snorre Raalund, *Lecture 10 (SDS 2019), lecture note* (Last assessed: 30-08-2019)

[13] NOAA, *Climate at a Glance - Global Time Series*, https://www.ncdc.noaa.gov/cag/global/time-series/55.676,12.568337/land_ocean/ann/7/1910-2019. (Last assessed: 30-08-2019)

[14] DR, *Holland bringer greenpeace sag til international domstol*, https://www.dr.dk/nyheder/udland/holland-bringer-greenpeace-sag-til-international-domstol. (Last assessed: 29-08-2019)

[15] TV2, *Pia Kjaersgaard: Høj stemmeprocent skyldes klimatosser*, https://nyheder.tv2.dk/2019-05-26-pia-kjaersgaard-hoej-stemmeprocent-skyldes-klimatosser. (Last assessed: 29-08-2019)

[16] Social- og indenrigsministeret, *Valg til Europa-Parlamentet 2019*, https://valg.sim.dk/valgmyndigheder/valg-til-europa-parlamentet-2019/. (Last assessed: 29-08-2019)

[17] Github, *VaderSentiment*, https://github.com/cjhutto/vaderSentiment. (Last assessed: 29-08-2019)

8 Appendix:

Figure 8.1: Plots of the original scraping process for 'Klima'

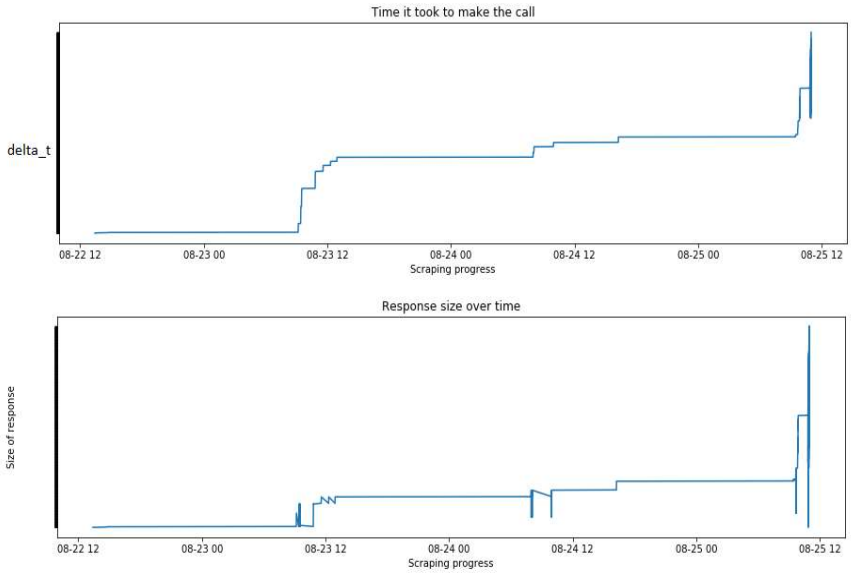


Figure 8.2: Plot of $delta_t$ against $response_{size}$ for 'Klima' from the original scraping process

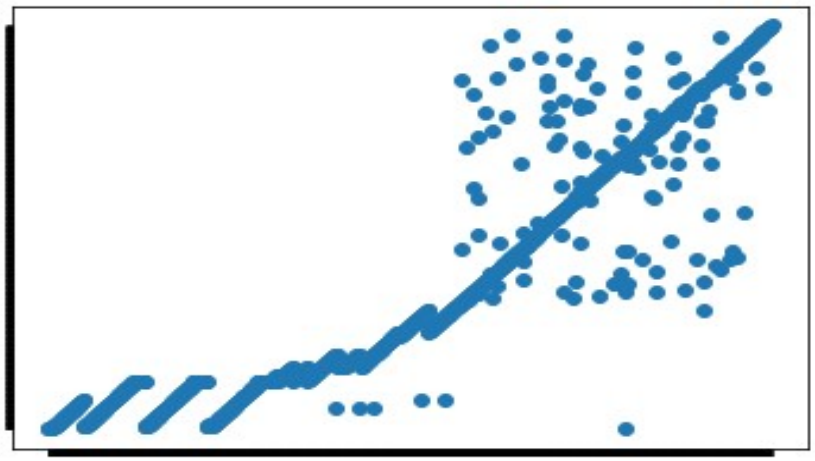


Figure 8.3: Average Danish sentiment score of the climate buzzwords, 2009-2019

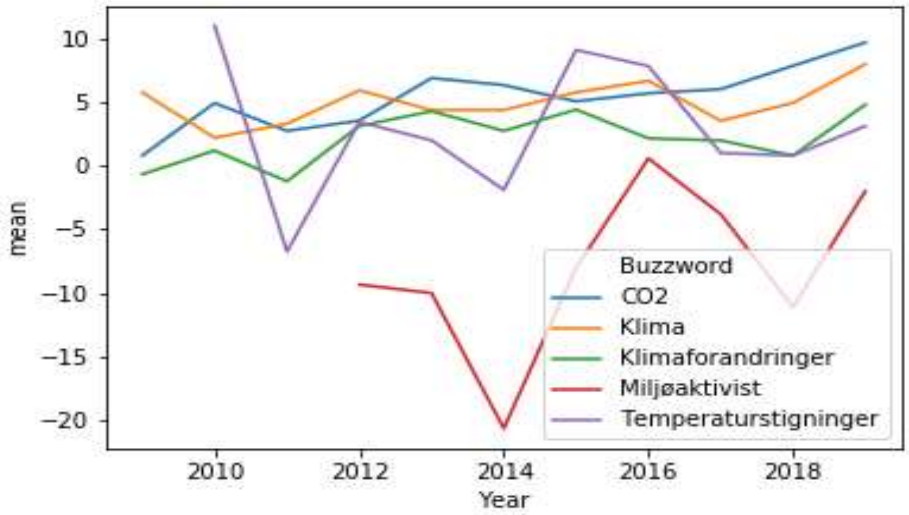


Figure 8.4: Average English sentiment score of the climate buzzwords, 2009-2019