Climate Trends in the Political Discourse of the Danish Parliament

Eisuke Okuda
ITU, Copenhagen
eiok@itu.dk

Andreas F. F. Olsen
ITU, Copenhagen
frao@itu.dk

Anders Hjulmand
ITU, Copenhagen
ahju@itu.dk

Abstract

The evidence for human-induced climate change is overwhelming. There is an increasing urgency for implementing policies that mitigate the consequences. However, the extent to which climate policies are discussed among politicians remains unclear. This paper examines trends in the climate discourse in the Danish parliament *Folketinget* using publicly available meeting transcriptions from the period of 2007 to 2023. The meeting transcriptions contain multiple agendas that include speeches from politicians to address the given agenda. This paper utilises annotations on the agenda titles and subsequently trains a classifier to label all speeches as climate related or non-climate related. The study finds a general increase in the prevalence of climate related discussions from 2007 to 2023. Moreover, left-wing parties engage more in climate related discussions than right-wing parties. The compiled dataset provides opportunities for future work to conduct extensive analysis on the general political discourse from 2007 to 2023.

Code & processed data is available on Github. Raw data is available on OSF.

1 Introduction

The rise in greenhouse gas emissions is accelerating and is expected to escalate further without effective political interventions (Füssel, 2009). In addition, the loss of biodiversity is happening at an alarming rate (Cardinale et al., 2012). As a consequence, the damages associated with climate change are becoming increasingly comprehensible. Political adaptation to the climate changes is therefore unavoidable (Berrang-Ford et al., 2011).

While there appears to be a growing political focus on implementing climate policies, such as investments in the renewable energy sector or introducing fees on greenhouse gas emissions, the exact extent to which climate are discussed among politicians remain yet to be examined. Our paper uses publicly available transcriptions from the Danish Parliament to analyse the climate trends in the political discourse.

The Danish Parliament *Folketinget* consists of 179 members that are elected for a period of maximum four years. The primary discussions of Folketinget take place in the parliament hall where meetings occur daily in the weekdays (except on Mondays). The attendees are primarily the elected politicians, but people from outside the political parties can also attend occasionally. The majority of the members in Folketinget represent a political party, although there are some members with no party-affiliations. The political parties are often placed along a spectrum that reflect their core political values. Left-winged parties, such as *Alternativet*, *Enhedslisten* and *Socialistisk Folkeparti*, have traditionally been engaged in climate related policies. On the other hand, right-winged parties, such as *Dansk Folkeparti*, *Liberal Alliance* and *Konservative Folkeparti* have historically been more active in non-climate related policies, such as immigration.

To explore how the Danish politicians engage in discussions about the climate, this paper investigates the following hypotheses:

- **H1:** The frequency of political discussions addressing the climate is generally increasing from 2007 to 2023.
- **H2:** The frequency of political discussions addressing the climate is higher in the time-period preceding an election compared to the time-period proceeding an election.
- **H3:** Left-wing parties engage more in discussions about the climate than right-wing parties.

To examine the above hypotheses, this paper utilises meeting transcriptions from the Danish parliament hall from 2007 to 2023, resulting in a dataset of 1.684 meetings and 369.762 individual speeches.

2 Data collection

2.1 Parliament hall meetings

The meetings in the parliament hall of Folketinget are transcribed and publicly available on the website of Folketinget[1]. The transcriptions were collected in the period of October 3rd 2007 to September 7th 2023 available in the format of HTML. There also exist transcriptions from October 5th 2004 to September 11th 2007, but these are only available in PDF format hence not collected.

Meeting transcriptions are structured as in Figure 1. Each meeting is divided into *m* agenda items, each denoting a self-contained discussion. Each agenda item contains *k* speech items, which denotes a speech by a single person. Speech items within the same agenda items are inherently related to another. Speech items contain information of the name, party, role (member of Folketinget, minister, etc.), and title of the speaker as well as time-stamps for the start and end of the speech. In every meeting before the first agenda item and after the last agenda item there is a section called *Announcements from the moderator*, containing practical matters on the daily operations of Folketinget. These sections were not included as they did not contain any political information.

The content of the meetings were scraped using the Python libraries `scrapy`, `requests`, and `beautiful soup`, resulting in a dataset comprising of the 3 tables: *meeting-data*, *agenda-data*, and *speech-data*. The 3 tables are linked via a composite key `[meetingID, agendaID, speechID]`.

Meeting-data comprises one row per meeting, and includes information of the date, day and at which time the meeting began. There was a total of 1.684 meetings.

Agenda-data includes information on the title and type of agenda item. There are three types of agenda items: Regular (which are often the processing of legislative proposals), questions for ministers, and questions for the prime minister. There was initially a total of 23.856 agenda items.

Speech-data consists of one row per speech item, and includes information of the speaker (name, party, role and title) as well as duration of the speech and the speech text itself. All speech items belonging to the moderators were removed as they did not contain any political content. This resulted in a total of 369.762 speech items.

2.1.1 Missing values in the datasets

The *agenda-data* had 8.620 entries with no speech items, e.g. when an agenda item was a vote on a legislative proposal or when a question for a minister was withdrawn. These agenda items were removed resulting in 15.236 rows in *agenda-data*. The *speech-data* contained 1.409 rows with missing title of the speaker. The speech content in these rows was meta-information of the meeting status, such as *”The meeting is postponed”* and therefore removed. In addition, there were 66.095 speech items with missing speaker party values whereof 1.057 belonged to members of Folketinget and 65.038 belonged to ministers i.e. *missing at random*. The 1.057 missing speaker party values of members were imputed using the title of the speaker. However, the parties of the ministers were not included in the transcriptions and were therefore *cold deck* imputed by scraping additional data from Wikipedia.

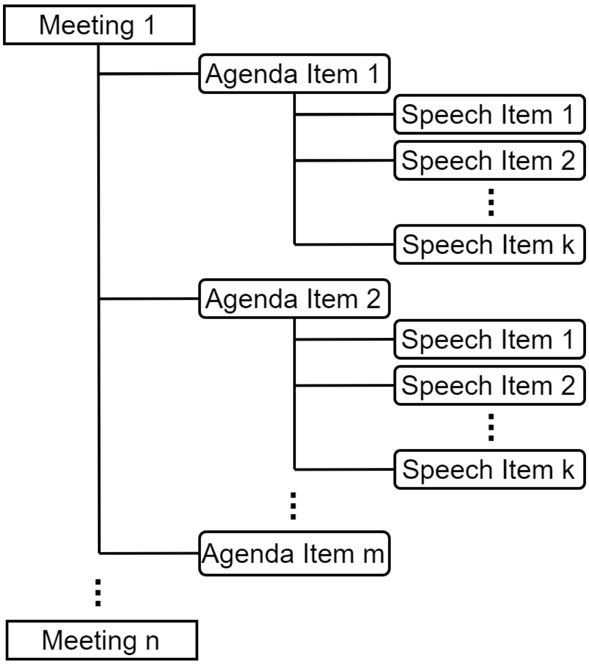

Figure 1: Structure of meeting transcriptions from the Danish Parliament *Folketinget*. Each meeting contains *m* agenda items, and each agenda item contains *k* speech items.

2.2 Members of Folketinget

The names, parties and time-period of members in Folketinget from 2005 to 2023 were collected from

[1] https://www.ft.dk/da/dokumenter/dokumentlister/referater

Wikipedia[2] using `request` and `beautiful soup`. Members might belong to different parties at different points in time. The resulting dataset *parliament-members* consists of 1.074 rows corresponding to 6 instances of Folketinget.

The missing values in speaker party of *speech-data* were initially imputed with data from *parliament-members* based on matching name and time-period. However, there were still 23 unique speakers with missing affiliations in this initial imputation. 5 of these speakers were missing in *parliament-members*, and the remaining 18 did not match the time-period in *parliament-members*. The party and time-period of these 23 speakers were scraped from Wikipedia[3] and added to *parliament-members*. This resulted in no missing values of speaker party in *speech-data*.

2.3 Summary statistics of the datasets

The average duration of meetings was 4 hours and 31 minutes, containing on average 7 agenda items and 199 speech items. Out of the 369.762 speech items, 304.410 belonged to members of Folketinget, the remaining 65.352 belonged to ministers, and all the speech items had an average duration of 51 seconds.

Table 1 shows summary statistics of the three types of agenda items. Questions for ministers and prime ministers are shorter than regular agenda items.

Type	n	Speech items	Words	Duration
Regular	7,767	25	6,348	41:18
Question for M	7,108	8	644	7:49
Question for PM	361	8	647	7:34

Table 1: Summary statistics of the 3 types of agenda items. Number of speech items and words are averages. Duration is given in minutes:seconds. Abbreviations: M; minister, PM; Prime Ministers.

3 Methodology

3.1 Annotation

The main purpose of this study was to label speech items as climate related or non-climate related to investigate the climate trends in the Danish political discourse.

An initial examination of 200 speech items revealed that climate related speeches were infrequent $\sim 5\%$. In addition, the average number of words in speech items were 151 compared to the average length of agenda item titles of 11. It was chosen to annotate agenda item titles based on the assumption that climate related agendas also contained climate related speech items.

The 1.542 agenda items ($\sim 10\%$) were sampled while stratifying for year and type of agenda item, to reduce the risk of representation bias. The annotation task was to label agenda item titles as either climate related: **C**, non-climate related: **NC**, or **Unknown**. The annotators followed predefined guidelines that included two sets of climate related topics (see appendix A). When encountering a topic from the first set, the agenda title was labeled **C** regardless of context, whereas topics belonging to the second set were context dependent. If an agenda item did not oblige to any of the sets it was labeled **NC**. Despite these guidelines, the notion of what makes an agenda item climate related or non-climate related differs from individual to individual based on their implicit biases.

There were three male annotators aged from 24 to 26, two of Danish origin and one of Japanese origin, for whom the Danish texts were machine translated into English. The translation could lead to measurement bias. All agenda items were annotated by all three annotators to ensure quality, whilst items with at least one **Unknown** label, was subsequently annotated in plenum. The individual **C** and **NC** annotations were finally aggregated using majority voting.

To further evaluate the annotation results, the inter-observer variability was calculated using Cohen's Kappa statistic as follows:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{1}$$

where,

- p_o is the probability of actual agreement (observed agreement),

[2] For example the Folketing of 2005.

[3] For example Jeppe Kofod.

- p_e is the probability of chance agreement (expected agreement).

3.2 Text processing

To analyse the speech items of all meetings, the speech texts were tokenized using the `spaCy` model `da_core_news_sm3.7.0` that was trained on two different datasets `Dane` and `UD Danish DDT` (Hvingelby et al., 2020; Brogaard Pauli et al., 2021; Johannsen et al., 2023; Honnibal et al., 2020). The `spaCy` model included a set of danish stop-words that was deployed as an initial filtering as well as removing words with alpha numeric values but specifically selecting any version of the word *CO2*.

To compare semantics of the speech items numerically, all tokens were mapped to a 500 dimensional embedding space utilizing a skip-gram `word2vec` model trained using the library `gensim`. The model published by the Society of Danish Language and Literature was trained on a corpus consisting of more than $1B$ words including texts from parliament transcripts and the danish dictionary (Nimb and Sørensen, 2018; Sørensen et al., 2023; Mikolov et al., 2013). To represent a single speech item, an average vector of all word vectors comprising the speech item was calculated as follows: Let t_i be the set of tokens in $speech\_item_i$, such that $t_i = \{w_0, w_1, ...w_n\}$ and $speech\_item_i, w \in \mathcal{R}^{1\times 500}$

$$speech\_item_i = \frac{1}{|t_i|}\sum_{j}^{|t_i|} t_j \tag{2}$$

3.3 Sets of initial words

To transition from annotated agenda items to annotated speech items, two sets of words that characterize climate related and non-climate related speeches were created. Speech items were sampled from all the annotated agenda items, resulting in 2.271 speech items from **C** agenda items and 35.423 speech items from **NC** agenda items. All unique words from the selected speech items were extracted, and the word frequencies were calculated as the relative frequency to the total number of words in **C** and **NC** respectively, denoted wf_C and wf_{NC}. The odds of a word appearing in either **C** or **NC** were calculated as follows:

$$odds_C = \frac{wf_C}{wf_{NC}} \tag{3}$$

$$odds_{NC} = \frac{wf_{NC}}{wf_C} \tag{4}$$

Table 2 shows an example of two words with high $odds_C$, two words with high $odds_{NC}$ and two words with odds around 1. The mean (SD) of $odds_C$ and $odds_{NC}$ was 1.99 (12.9) and 0.46 (1.39).

Word	$odds_C$	$odds_{NC}$
nuclear	312.8	0.001
nature plan	387.3	0.002
crime	0.01	104.8
health sector	0.01	36.7
thanks	0.94	1.06
time	1.09	0.91

Table 2: Examples of words with high $odds_C$ and $odds_{NC}$, and odds around 1. Words were translated from Danish to English.

When using a cut-off point in odds to define a set of salient words, there exists a trade-off between set-size and quality of words in the set. Increasing the cut-off point will result in a smaller set-size but higher quality words and vice versa.

In order to find a reasonable cut-off point a subset of words was created with $odds_C > 1$ and $wf_C > 1$ and another subset of words with $odds_{NC} > 1$ and $wf_{NC} > 1$. These words were divided into 20 percentiles based on their odds. The average cosine similarity ACS within each percentile group $k = 1, 2, ..., 20$ was then calculated using the function `similarity` from the library `gensim`.

$$ACS_k = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n} \sum_{j=i+1}^{n} cos(word_i, word_j) \tag{5}$$

where n is the number of words belonging to percentile k.

Figure 2 shows ACS for each odds percentile groups. The ACS for **C** and **NC** appear to be stationary for percentile groups 1 to 15, and generally increasing for percentile groups 16-20. In addition, the ACS for **C** increases more than for **NC**, meaning that subsets of words with high $odds_C$ are more semantically similar than subsets of words with high $odds_{NC}$. This could be because the **NC** group consists of a heterogeneous set of topics,

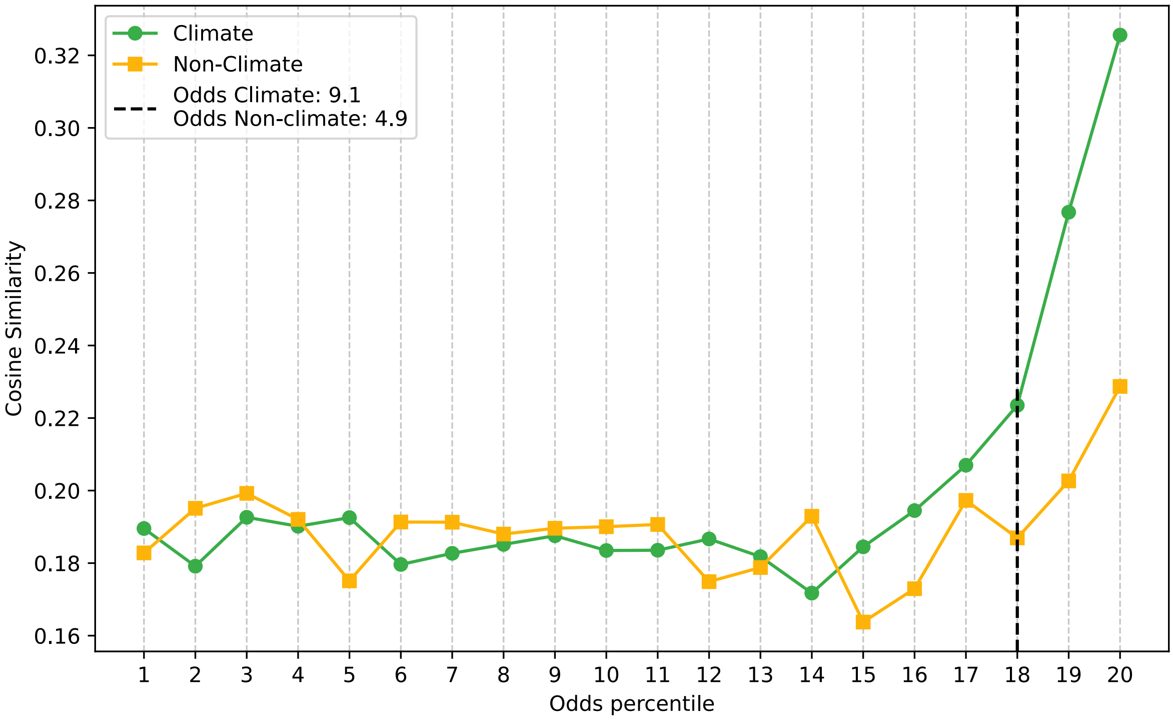

Figure 2: Average cosine similarity for odds percentile groups.

whereas the **C** only refers to set of topics within climate.

A cut off point of 18 was chosen, thereby selecting words within the highest 15% $odds_C$ and $odds_{NC}$. This resulted in a set of 174 **C** words a set of 154 **NC** words. Among these words, the minimum $odds_C$ and $odds_{NC}$ was 9.1 and 4.9 respectively.

To verify that the set of 174 **C** words were meaningful and coherent, a "semantic filter" was applied. The 500 dimensional embedding space of words from both sets were initially reduced to 2 dimensions using `t-SNE` with perplexity 20 (Van der Maaten and Hinton, 2008). A `K-Means` algorithm with 2 clusters was subsequently deployed to group words by their semantic characteristics. The words from both sets were generally clustered to their respective groups, but 66 out of the 174 **C** words were clustered as **NC**, including words such as *cars*, *area*, and *forest-politics*. These words were removed from the set of **C** words, with the exception of a few manually chosen words. In addition, a group of context dependent **C** words from the **C** cluster were manually removed. This resulted in a set of 108 **C** words.

3.4 Sets of augmented words

The 2 sets of initial words were based entirely on the speech items of the annotated agenda items ($\sim 10\%$ out of all agenda items). To obtain a better representation of salient **C** and **NC** words, the initial two sets were augmented using the remaining 332.068 speech items. These were split into **C** and **NC** subsets based on each speech item having a word count higher in one group than the other from the initial sets of **C** and **NC** words. The two sets of augmented words was then created by following the same procedure as above in section 3.3 (see Appendix B for results on average cosine similarity). This resulted in a set of 58 augmented **C** words a set of 156 augmented **NC** words. The sets of augmented words did not overlap with the sets of initial words.

3.5 Set of generic words

In addition to the sets of **C** and **NC** words, a set of generic words was created by extracting all unique words from the 332.068 non-annotated speech items. Words appearing with a wf_C and a wf_{NC} in the top 5 percentile and an $odds_C$ and $odds_{NC}$ between 0.9 and 1.1 were selected, resulting in a set 190 generic words. These generic words were occurring often across all speech items, and were equally likely to appear in **C** and **NC** speech items. The set of generic words thus serves as a custom-made stop-words list. All speech items containing one or more generic word had them removed before computing the speech item vector in equation 2, such that the signal of the remaining words was enhanced within the speech item representation.

3.6 Classifier

Investigating the prevalence of climate related speeches in the parliament transcriptions was treated as a supervised binary classification task with classes; **C** and **NC**.

A training set was selected based on the sets of **C** and **NC** words. If a speech item contained 2 or more words from the **C** word set and had a ratio of **C** to **NC** words > 0.5, it was given a pseudo label C and oppositely if a speech item had one or more **NC** related words but the ratio was ≤ 0.5, meaning the frequency of **NC** words compared to **C** words was less than double, it was given a pseudo label NC. This approach was based on the assumption that if more than two **C** words, such as *Sustainable*, *Windmillpark*, and *Energy*[4] appeared, it was deemed a climate related speech item. The less strict threshold of **NC** words stems from the NC class being inherently more heterogeneous. Some of the speech items that received pseudo labels were manually assessed before adopting them as "true" labels.

Excluding the annotated dataset, this process resulted in a training set consisting of 197.595 speech items with a 1:10 ratio of C to NC all represented by their vectors as input features $\in \mathcal{R}^{1\times500}$ and a target value C or NC.

To perform the classification task, a `LightGBMClassifier` model was employed with the library's default hyper parameters (Ke et al., 2017). To mitigate the risk of overfitting, `StratifiedKFold` was deployed with $k = 5$ using `SciKit Learn` (Ojala and Garriga, 2009). The average prediction of the five folds was used to classify all 369.762 speech items.

To evaluate the model performance, the 1.687 annotated agenda items consisting of 37.694 speech items was used as a test set. The agenda items in the test set were labeled C if more than 50% of its speech items were predicted as C and vice versa i.e. a majority voting.

3.7 Identifying climate topics

To further examine the climate related topics frequently discussed in Folketinget, the 27.846 C-classified speech items was extracted. The average vectors of these speech items were recalculated by using a new set of **C** words. This set includes words with $odds_C > 2$, such that the vectors are highly sensitive to climate related topics.

Uniform Manifold Approximation and Projection (`UMAP`) (McInnes et al., 2018) was then deployed to reduce the dimensionality into 3 dimensions. While `UMAP` and `t-SNE` are both popular dimensionality reduction techniques, `UMAP` arguably preserves more of the global structure. An inspection of the initial visualization revealed three distinct clusters. A `K-Means` algorithm with 3 clusters was further applied to categorize speech item vectors by their semantic characteristics. To label each cluster, their centroids were fed into the `most_similar` function from the library `gensim`, to retrieve the top k words by their cosine similarities as seen in Table 3.

Word	**Cosine Similarity**
fossil-free	0.768
low-energy society	0.768
the energy systems	0.766
energy future	0.757
the data centers	0.752

Table 3: Top 5 words that had the highest similarity to the vector representation of "Energy" cluster centroid from Figure 5. Words were translated from Danish to English.

4 Results

4.1 Annotation results

Table 4 shows the annotation results from the 3 annotators. The results demonstrate that **C** agenda items were infrequent. While there was a low variability among annotator results in **NC** labels, the variance among **C** and **Unknown** labels was higher. In particular, annotator 1 who was of Japanese origin annotated fewer **C** than annotator 2 and 3, which could be an effect of implicit biases, in particular the confirmation bias.

The Cohen's Kappa scores were calculated on each pair of annotators resulting in $\kappa_{1,2} = 0.71$, $\kappa_{1,3} = 0.56$, and $\kappa_{2,3} = 0.67$. These numbers indicate that while annotator 1 and annotator 2 agreed the most on their annotations, annotator 1

[4] Translated from Danish to English

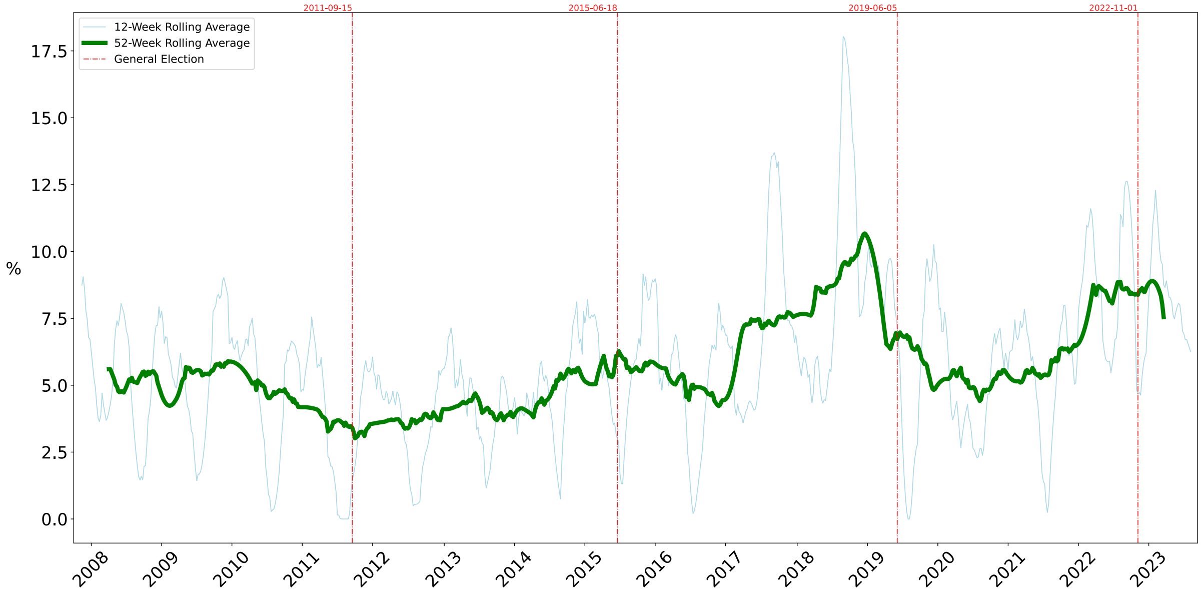

Figure 3: Percentage of climate related speech items from 2008 to 2023 along with time-stamps for general elections. Data is aggregated by week and visualised using a centered moving average with a window of 12 weeks (blue) and 52 weeks (green).

Label	Annotator 1	Annotator 2	Annotator 3
NC	1433 / (0.93)	1418 / (0.93)	1404 / (0.91)
C	58 / (0.04)	99 / (0.06)	124 / (0.08)
Unknown	51 / (0.03)	25 / (0.01)	14 / (0.01)
Total	1542	1542	1542

Table 4: Annotation results from each Annotator. Annotator 1 used English translations.

and annotator 3 agreed poorly. These results also indicate the existence of confirmation bias. The aggregated annotator results are shown in Table 5.

Label	Count
NC	1438 / (0.93)
C	104 / (0.07)

Table 5: Aggregated annotation results.

4.2 Classifier results

To investigate the model performance on the 1.542 true annotated agenda items, Table 6 shows the majority voting results. Given the speech item vectors the model was able to predict 57 out of the total 104 agenda items that were **C** agenda items. The false positive C predictions indicate that the model picked up on some features that insinuated **C** speech items but actually were **NC**. This is reflected in Table 7 with a precision of 0.626. This is also reflected in the recall of 0.548 as the model predicted 47 actual **C** to be **NC**, suggesting that the majority voting threshold may be higher than the proportion of **C** speeches within an agenda item. The $F1$ score of 0.584 seen in Table 7 indicates the ability to identify the presence of both classes, as opposed to overfitting to class NC and achieve a similar overall accuracy.

	Predicted C	**Predicted NC**
C	57	47
NC	34	1404

Table 6: LightGBM Confusion Matrix

	Precision	**Recall**	**F1**
Values	0.626	0.548	0.584

Table 7: Evaluation Metrics

5 Main findings

The aim of this paper was to examine how much politicians in Folketinget engage in climate related discussions. It was found, that the overall propor-

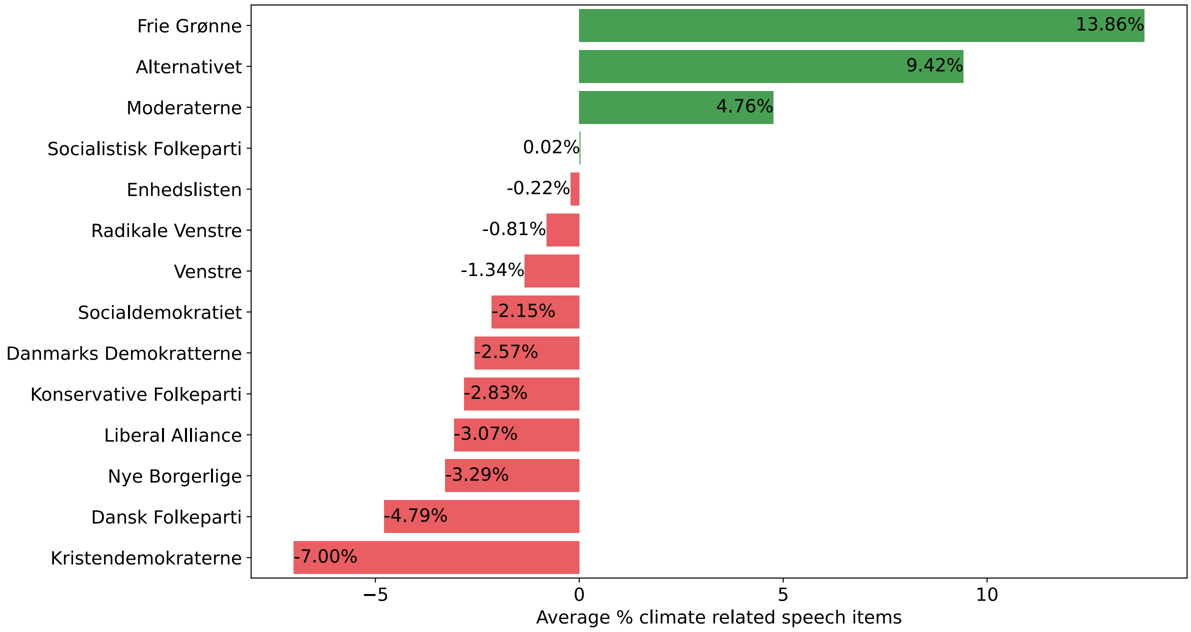

Figure 4: Percentage of climate related speech items for each political party relative to the average percentage of climate related speech items across all political parties.

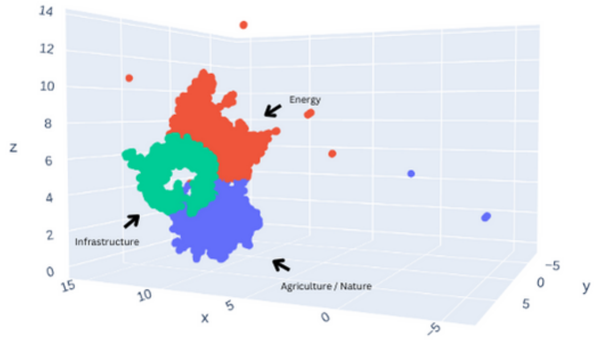

Figure 5: **C** labeled speech items visualized in a three dimensional space. Three distinct clusters were identified with the categories: *Energy*, *Agriculture/Nature*, and *Infrastructure*.

tion of climate related speech items in Folketinget from 2007 to 2023 was 7.5%. In addition, it was found that the amount of political discussions about climate generally increased from 2008 to 2023, thus providing evidence for **H1** (see Figure 3).

Furthermore, a visual inspection of Figure 3 suggests that the amount of political discussion about climate was higher in the time-period preceding an election compared to the time-period proceeding an election, with the exception of the 2011 election. This provides some evidence for **H2**. This trend might be a result of the climate crisis being one of the most important topics for Danish voters (vox).

It was found that left-wing parties generally engage more in climate discussions than right-wing parties as seen in Figure 4. This result provides evidence for **H3**. It is important to note that some political parties were only established after 2007, and since there was generally more climate related speech items in later years, the high percentages of *Frie Grønne* and *Moderaterne* should be interpreted with caution. Although left-wing parties were found to generally engage more in climate discussions, it was *Liberal Alliance*, a right-wing party, that had the largest increase in climate related speech items (Figure 6).

In addition to the initial hypotheses, frequently discussed climate related topics in Folketinget were identified. Clustering revealed 3 categories within the climate related speeches: *Energy* (48%) *Agriculture/Nature* (33%), and *Infrastructure* (19%) (see Figure 5). These results indicate that topics related to renewable energy were most-frequently discussed, while conversations about infrastructure were less common.

6 Discussion and limitations

6.1 Classifier

The classifier results demonstrates the models ability to distinguish some agenda items from climate related to non-climate related based on the majority voting. These positive climate signals are re-

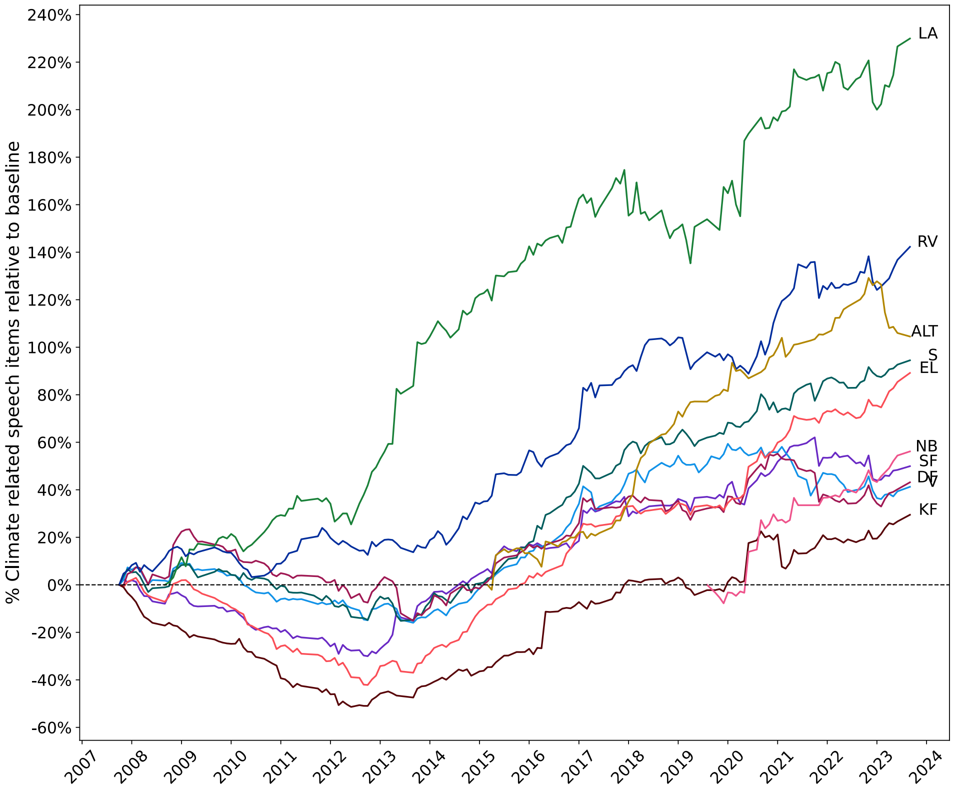

Figure 6: Percentage of climate related speech items relative to the baseline for each political party. The baseline is the percentage of climate related speech items in the first month. Data is aggregated by month and visualised using a centered moving average with a window of 24 months. The political parties *Moderaterne*, *Kristendemokraterne*, *Danmarks Demokraterne*, and *Frie Grønne* are removed due to lack of data.
Abbreviations: LA; Liberal Alliance, RV; Radikale Venstre, ALT; Alternativet, S; Socialdemokratiet, EL; Enhedslisten, NB; Nye Borgerlige, SF; Socialistisk Folkeparti, DF; Dansk Folkeparti, V; Venstre, KF; Konservative Folkeparti.

flected in the F1 score being greater than a random learner at 0.5 and neither show tendency to predict all agenda items to be non-climate indicated by the precision and recall.

Since the annotated true labels only exist on the agenda item level, it raises the questions of which threshold of the proportion of climate related speech items in an agenda item, to be labeled as climate related. The majority voting labels agenda items containing 45% predicted climate related speech items to still be non-climate, which adds an additional error step in the classifying process.

In addition to the majority voting, the numerical representation of the speech items also limits the classifier. Each word vector loses it signal in the process of representing a speech item as the average of all words present in the particular speech item. The signal loss is mitigated by statistical heuristics that removes stop words in a speech item to enhance the signal of the remaining words, pulling each speech item vector in either climate or non-climate direction in the 500 dimensional embedding space.

6.2 Annotations

The main limitation of this project was that annotations were conducted on agenda items, whereas the analysis and classification were based on speech items. Future work could leverage crowd-sourcing to obtain sufficient speech item annotations.

In addition, the distinction between climate related and non-climate related agenda item titles is inherently ambiguous. This creates an element of subjectivity in the annotation task, despite the annotators following the same set of guidelines. Moreover, some agenda titles may be difficult to label without contextual knowledge. This may partly explain why annotator 1, who used English translations of the titles, only annotated 58 as **C** compared to the others (99 and 124 **C**).

The annotation process was the first part of classifying speech items as climate related or non-climate related. The subsequent steps, from creating sets of salient words to implementing a classifier was heavily dependent on the quality, diversity and size of the annotation dataset. As an example, the odds were heavily dependent on the speech items from the annotated agenda items. A different sample of annotated agenda items, could have resulted in different sets of words.

6.3 Sets of words

There are several shortcomings concerning the sets of words. Although the average cosine similarity was generally increasing as the odds increased, there was no clear way of defining a reasonable cut-off point to distinguish salient words from trivial words. It was chosen to include odds percentile groups 18, 19, and 20, but lower odds percentile groups, could also have been selected.

In addition, it is unlikely that the set of non-climate words encompass the entire space of ”true” non-climate speech items. The non-climate category is by definition diverse, covering topics from education to immigration. Since the set of non-climate words was ultimately based on the annotated agenda items, it is unlikely that they faithfully represent the entire space of topics.

The definition of pseudo-labels was based entirely on heuristics utilising the sets of words, for example, that pseudo label **C** was given to speech items with two or more words from the **C** word set. This limitation could have been avoided by annotating speech items.

7 Conclusion

This study examined the extent to which politicians in the Danish parliament *Folketinget* engage in discussion about the climate. A dataset of meeting transcriptions from Folketinget in the period of 2007 to 2023 was created, resulting in 369.762 individual speeches. A classifier was subsequently trained to label speeches as climate related or non-climate related.

The results of this paper showed a general increase in the prevalence of climate related discussions. Furthermore, left-wing parties engaged more in climate related discussions than right-wing parties. The results suggested that climate policies are becoming a more frequently discussed topic among Danish politicians, but that there is a substantial difference between the engagement of the parties.

It is important to interpret the findings with caution, as the classifier lacked ground truth labels for speech items throughout the training and validation process. Future work could improve the classifier by using annotated speech items, although this is a comprehensive annotation task.

Future work could also leverage the 3 climate related categories; *Energy*, *Agriculture/Nature*, and *Infrastructure*, to further examine climate trends in the political discourse of the Danish Parliament.

References

Voxmeter. http://www.voxmeter.dk/om-os/.htm. Accessed: 2023-12-10.

Lea Berrang-Ford, James D Ford, and Jaclyn Paterson. 2011. Are we adapting to climate change? *Global environmental change*, 21(1):25–33.

Amalie Brogaard Pauli, Maria Barrett, Ophélie Lacroix, and Rasmus Hvingelby. 2021. DaNLP: An open-source toolkit for danish natural language processing. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa 2021)*.

Bradley J Cardinale, J Emmett Duffy, Andrew Gonzalez, David U Hooper, Charles Perrings, Patrick Venail, Anita Narwani, Georgina M Mace, David Tilman, David A Wardle, et al. 2012. Biodiversity loss and its impact on humanity. *Nature*, 486(7401):59–67.

Hans-Martin Füssel. 2009. An updated assessment of the risks from climate change based on research published since the ipcc fourth assessment report. *Climatic change*, 97(3-4):469–482.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Rasmus Hvingelby, Amalie Brogaard Pauli, Maria Barrett, Christina Rosted, Lasse Malm Lidegaard, and Anders Søgaard. 2020. DaNE: A named entity resource for Danish. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4597–4604, Marseille, France. European Language Resources Association.

Anders Johannsen, Héctor Martínez Alonso, and Barbara Plank. 2023. Ud danish-ddt. Universal Dependencies version 2.10.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Sanni Nimb and Nicolai H. Sørensen. 2018. Word2dict – lemma selection and dictionary editing assisted by word embeddings. In *Proceedings from Euralex 2018*, Ljubliana, Slovenia.

Markus Ojala and Gemma C. Garriga. 2009. Permutation tests for studying classifier performance. In *2009 Ninth IEEE International Conference on Data Mining*, pages 908–913.

Nicolai Hartvig Sørensen, Society of Danish Language, and Literature. 2023. Word2vec-model for danish.

A Annotation: Sets of topics

Two sets of topics were defined for the annotation process. If an agenda item title includes any words from the first set of topics, it should be labeled **C** regardless of the context. On the other hand, if an agenda item title includes topics in the second set, it should only be labeled **C** if the topic appears in a climate context. As an example, many discussions about the agriculture sector are about finances and not about the climate.

The first set of topics includes:

- CO2
- Nuclear
- Green
- Plant based
- Nature Nurturing
- Biodiversity
- Windmills
- Recycling
- Nature & Enviroment

The second set of topics includes:

- Electricity
- Agriculture
- Energy
- Infrastructure

B Average cosine similarity for sets of augmented words

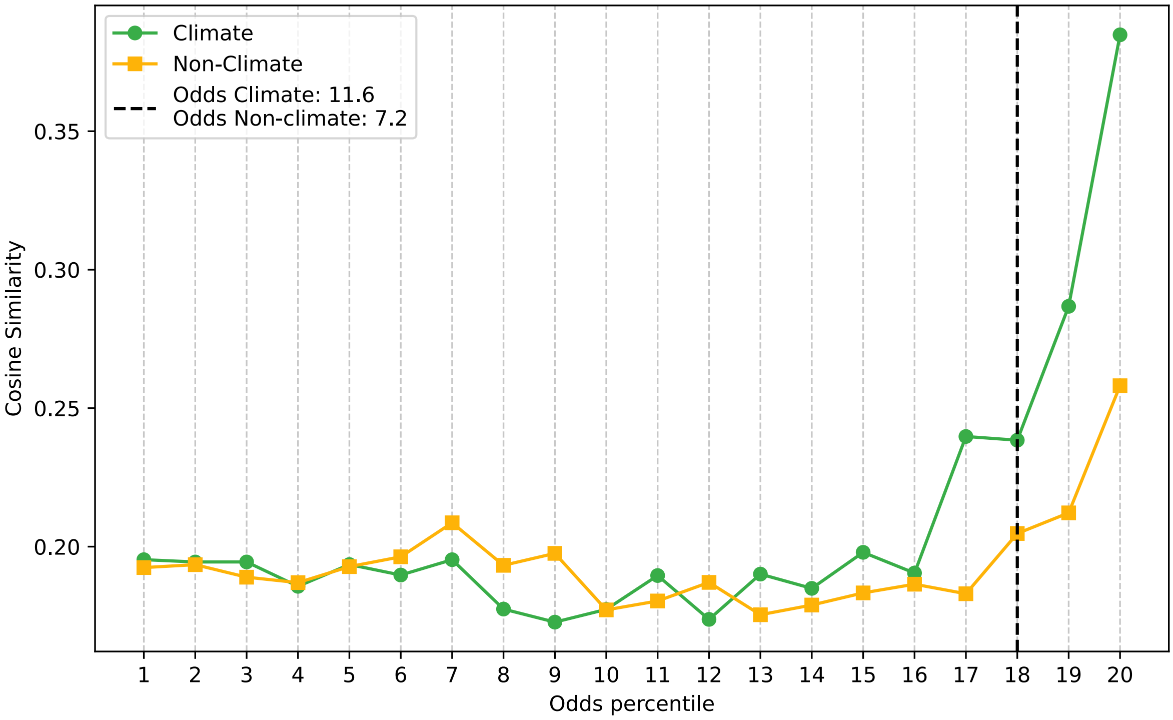

Figure 7: Average cosine similarity for odds percentile groups in the sets of augmented words.