

Aalborg Universitet



Using BERT to Study Semantic Variations of Climate Change Keywords in Danish News Articles

Meier, Florian Maximilian

Published in:

Poster @ DHNB 2024 Digital Humanities in the Nordic and Baltic Countries 8th Conference

Creative Commons License
CC BY 4.0

Publication date:
2024

Document Version
Early version, also known as pre-print

[Link to publication from Aalborg University](#)

Citation for published version (APA):

Meier, F. M. (2024). Using BERT to Study Semantic Variations of Climate Change Keywords in Danish News Articles. In *Poster @ DHNB 2024 Digital Humanities in the Nordic and Baltic Countries 8th Conference*

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal -

Take down policy

If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.

Using BERT to Study Semantic Variations of Climate Change Keywords in Danish News Articles

Meier Florian¹

¹*Department of Communication and Psychology, Aalborg University, Copenhagen*

Abstract

The terms greenhouse effect, global warming, and climate change are often used synonymously in everyday conversations about the warming planet. Utilizing Danish BERT, the study employs masked language model tasks to uncover semantic shifts and variations of these climate change-related keywords in Danish news articles from 1990 to 2021. The findings offer insights into contextual understandings and framing nuances by journalists, contributing to a deeper comprehension of these terms in the Danish media discourse on CC.

Keywords


Climate change, Newspaper, BERT, Semantic shift,


1. Introduction and background

The terms greenhouse effect (GE), global warming (GW), and climate change (CC) are often used interchangeably to describe "human activities that alter the composition of the global atmosphere" (UNFCCC 1992). Each term, however, is associated with different aspects of this reality. GE refers mainly to causes of climate change; GW concerns the effects of changes to the global atmosphere and the long-term warming of the planet. CC encompasses GW and describes a broader range of changes, including melting glaciers, rising sea levels and severe weather events. While these terms carry their nuanced meanings, their interpretations in public communication have yet to be thoroughly examined. Due to their interchangeability, these terms are often treated as synonymous or utilized as concise descriptors for anthropogenic CC. They serve as de-facto standard keywords in data retrieval for CC news studies (Hase et al. 2021). Consequently, many studies need to pay more attention to the historical patterns, semantic variances, and communicative implications that underlie these three keywords. While we are not the first to study public perception and differences in the meaning of these terms, previous work has mostly relied on qualitative methods or costly to collect survey data (Schuldt 2016; Soutter and Möttus 2020).

In this article, we report on the outcome and experience of the use of large language models (LLMs), specifically Danish BERT (Bidirectional Encoder Representations from Transformers), to capture word senses and semantic variation in the usage of CC keywords in Danish news

 fmeier@ikp.aau.dk (M. Florian)

 <https://vbn.aau.dk/da/persons/142274> (M. Florian)

 0000-0001-9408-0686 (M. Florian)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

articles (Devlin et al. 2019). BERT has revolutionized natural language processing (NLP) because of the bidirectional training approach, among other reasons. BERT processes input data in both directions, which enables the model to capture contextual information from both preceding and following words, allowing for a better understanding of the overall context of a word in a sentence. It has also undergone pre-training using the masked language model task (fill-mask), where a certain percentage of words in the input text are randomly selected and replaced with a special [MASK] token. The goal of the model is then to predict the original words that were replaced. Inspired by previous work, we take advantage of BERT's capabilities in predicting masked tokens to study semantic shifts and variations of the three CC keywords (Lucy et al. 2023). By masking the keywords, we can both learn a) what Danish BERT has learned about them and b) how journalists have used the different keywords in their reporting.

While multilingual BERT models exist, these models usually perform poorly for languages such as Danish or Norwegian because of the underrepresentation of these languages in the training data. Thus, in our context, we use Danish BERT, a BERT model that is pre-trained on 1.6 Billion Danish tokens from Danish Wikipedia and other Danish texts (Møllerhøj, Rodriguez, and Fabrin 2021; Hvingelby et al. 2020).

2. Data

Our analysis relies on a dataset obtained programmatically through API access to *Infomedia*, one of the largest media archives in the Nordic Countries. We gathered articles related to CC and associated keywords (GE and GW)¹ spanning the years 1990 to 2021. The articles were sourced from all currently circulating printed newspapers and two online news platforms operated by public service corporations, Danmarks Radio (dr.dk) and TV2 (tv2.dk). These websites rank as the fourth and fifth most visited in Denmark (Similarweb 2023). Our focus is on core CC articles, encompassing all news items containing one of the keywords at least twice. In total, we work with a sample of 32,214 articles. For this study, we extract the sentences in the articles that mention one of these keywords.

Due to the rate limitations of the Hugging Face Inference API and a lack of CC reporting in the early years of our dataset's time frame, we focus on a subset of these sentences. We separate the dataset into five periods and randomly sample 5% of sentences for each period and keyword. This leaves us with 2303 sentences (GE 161, GW 785, CC 1357). Figure 1 shows the number of occurrences for each keyword per publishing period.

3. Approach

We study the semantic variation of CC keywords by performing masked word prediction (fill-mask) using Danish BERT via the Huggingface Inference API. Hugging Face is an open-source community and provider of machine-learning tools for NLP. One of its main features is the *Transformers library*, which allows researchers, developers, and practitioners to easily use and fine-tune various pre-trained models for tasks such as text classification or language translation.

¹The original Danish query was: klimaforandring* OR "global* opvarmning" OR drivhuseffekt*.

However, solid programming knowledge in a language like Python is necessary to use models from the library. Hugging Face’s Inference API, on the other hand, is a service that allows users to make predictions using pre-trained models hosted by Hugging Face without downloading or managing the models locally. The Inference API provides a simple and efficient way to make predictions by sending HTTP requests to the Hugging Face servers, where the models are hosted. Users can send text or input data to the API and receive model-generated responses in return. This process only requires knowledge of how APIs work and can be done using basically any programming language.

In our approach, we take the subsample of 2303 sentences, mask each occurrence of a CC keyword, and send the sentence to the inference API to retrieve a prediction from Danish BERT for what it thinks which token got masked. For example:

- **Original sentence:** Der er således stærke beviser på, at isbjørnene er påvirket af klimaforandringerne.
- **CC keyword masked:** Der er således stærke beviser på, at isbjørnene er påvirket af [MASK].
- **Predictions:** aber, natur, mennesker, is, naturen

After a successful call, the API returns five potential word candidates for the [MASK] token, including a confidence score representing the models likelihood estimate (0-1) that this is the correct word. In total our approach results in 11,515 (2303×5) predicted words and confidence scores.

4. Findings

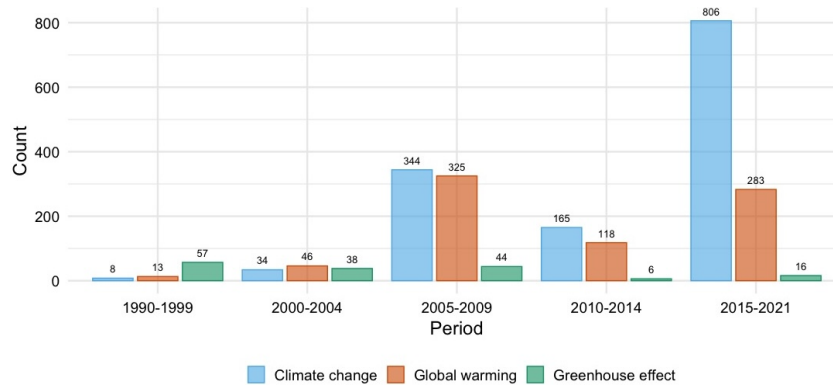


Figure 1: Keyword occurrence across the five publishing periods.

From Figure 1, we learn that in the 1990s, when CC reporting in Denmark was still in its infancy, GE was the most frequently used term. From the year 2000 onwards, both GW and CC began to be used more often by Danish journalists. It suggests that the 2000s was a transitory decade, with more terminological diversity in public discourses on CC. In the most recent period,

CC has been the dominant term, with GE losing more and more importance. While this analysis already gives some insights into how news reporting on CC has changed over the years, it does not yet tell us how these concepts have been discussed and framed differently in the press. For this, we look into the masked token prediction.

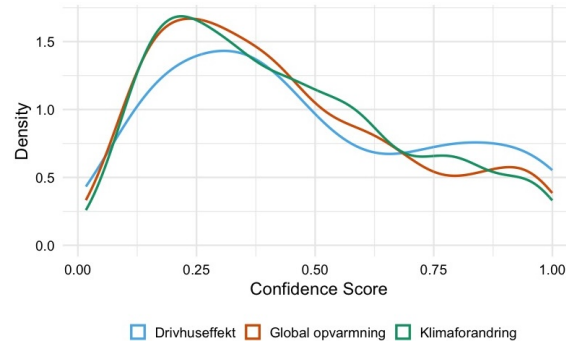


Figure 2: Distribution of confidence scores for the first (most likely) token for each keyword.

Figure 2 shows the distribution of confidence scores for the first predicted token. We can observe that GE follows a flatter curve and lies above GW and CC for very high confidence scores (0.75 - 1). Moreover, the average confidence score for GE is 0.477, and thus slightly higher than for the other two keywords (GW: 0.432, CC: 0.437), i.e. the model has more confidence in predicting a token when GE is masked. One explanation could be that the contexts in which the word GE is used — in other words, the frames journalists use to write about it — are more similar and leave less room for ambiguity.

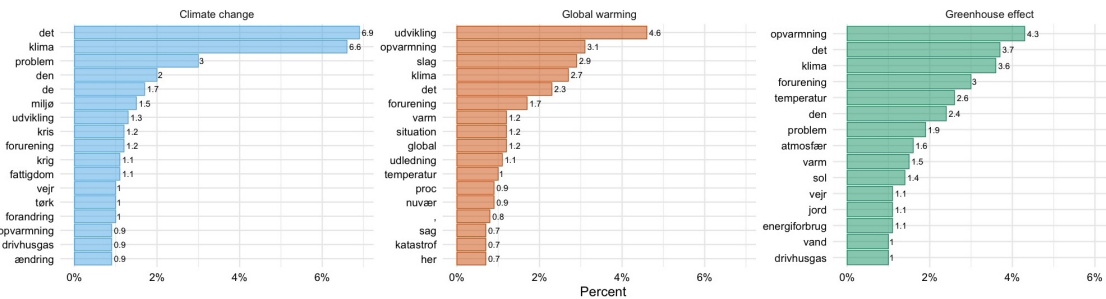


Figure 3: Top 15 most predicted words for each keyword. The words have been stemmed to account for different word forms.

Figure 3 shows the share of the 15 most frequently predicted tokens for each keyword. First, we can look at the cases where the model predicted the correct token. For GE, the model never predicted the correct token (drivhuseffekt). Given that we only masked the word *opvarmning* for GW, the model predicted the correct token in 3.1%. Moreover, the type-token ratio of the predicted terms tells us something about how varied the predictions were. The higher the ratio, the more unique terms, i.e. diverse predictions, were made. The type-token ratio is highest for GE (0.50), while for GW (0.28) and CC (0.22), it is comparably low. While we would expect that GE shows a lower value for the type-token ratio due to less variation in how it has been used in

the media, the metric is sensitive towards sentence length, i.e. the number of total tokens. As the number of total sentences for GE (161) is low, a higher number is to be expected.

Looking at the top 15 predicted words, it shows that each keyword is associated with different scientific and social discourses while, at the same time, a certain congruence is visible. For example, the words climate (klima), warming (opvarmning) and pollution (forurening) are among the top 15 words for each keyword. However, striking differences emerge, too. GE, for example, is mainly linked to natural science terms like atmosphere (atmosfær), energy consumption (energiforbrug) or greenhouse gas (drivhusgas). GW follows a different trajectory. Whereas it was a relatively technical notion in the 1990s, it became increasingly associated with the (social) effects of climate change in the 2010s. Of the three keywords, CC is the most diverse, dominated by words that emphasize the continuous change of the climate (forandring, ændring), its visible consequences (tørk, miljø and vejr) and the severity of the climate crisis (krise). Moreover, we can see that CC is framed similarly to other social issues, e.g. war (krig) or poverty (fattigdom). Using some basic sentiment analysis (AFINN), we also see that more negatively connotated words are predicted for CC (problem, krise, krig, fattigdom) than for the other two.

5. Discussion and conclusion

Our analysis tells us something about a) what Danish BERT has learned about CC as represented via the three keywords and b) how journalists have framed the different keywords in their reporting. Those two aspects are inseparable as the model makes predictions based on the text material it has been trained on and given the specific context at hand, i.e. the words surrounding the [MASK] token.

In this study, we are using the vanilla version of Danish BERT, which means we do not perform any fine-tuning. Consequently, we get less accurate predictions, as BERT can only use the context from the data it has been pre-trained on. However, this also says something about the training data itself and in what contexts the three keywords have appeared in the training material. As previously mentioned, the model predicted the correct token only in around 10% of the time. The fact that it never predicts the token *drivhuseffekt* hints at a lack of training data containing that token. Fine-tuning the model on our dataset would very likely increase this percentage by a large degree. However, our goal is not to increase its accuracy in this task but to learn about differences in meaning using sentences from news articles.

While the three terms are sometimes interchangeably used when speaking about CC in everyday life, our results show that journalistic writing in which these three keywords appear is similar yet different. GE, for example, is mainly used in a natural science context, while CC is the most diverse term, often framed similarly to other social issues.

One limitation of the approach is that only one word can always be masked. In the case of global warming, a two-word phrase in Danish (global opvarmning), only *opvarmning* could be masked. However, this did not increase the overall success of the model.

Finally, working with the Hugging Face Inference API was far from unproblematic. Currently, the API documentation (at least for the open, non-enterprise version) is insufficient and does not inform users about rate limits, which can result in unpredictable 429 (Too Many Requests)

errors. Thus, to be as gentle as possible, we limited the number of requests to 10 per minute. For smaller proof-of-concept studies, this seems acceptable. However, this approach does not scale if one has a bigger sample size. For this, using the Transformer library via Python is more suitable and recommended.

References

- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” In *Proc of ACL 2019: Human Language Technologies, Volume 1 (Long and Short Papers)*, edited by Jill Burstein, Christy Doran, and Thamar Solorio, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>.
- Hase, Valerie, Daniela Mahl, Mike S. Schäfer, and Tobias R. Keller. 2021. “Climate change in news media across the globe: An automated analysis of issue attention and themes in climate change coverage in 10 countries (2006–2018).” *Global Environmental Change* 70. <https://doi.org/10.1016/j.gloenvcha.2021.102353>.
- Hvingelby, Rasmus, Amalie Brogaard Pauli, Maria Barrett, Christina Rosted, Lasse Malm Lidegaard, and Anders Søgaard. 2020. “DaNE: A Named Entity Resource for Danish.” In *Proc. of LREC’12*, edited by Nicoletta Calzolari and et al., 4597–4604. Marseille, France: European Language Resources Association. ISBN: 979-10-95546-34-4. <https://aclanthology.org/2020.lrec-1.565>.
- Lucy, Li, Jesse Dodge, David Bamman, and Katherine Keith. 2023. “Words as Gatekeepers: Measuring Discipline-specific Terms and Meanings in Scholarly Publications.” In *Findings of the ACL 2023*, edited by Anna Rogers and et al., 6929–6947. Toronto, Canada: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.findings-acl.433>.
- Møllerhøj, Jens Dahl, Dario Rodriguez, and Henrik Fabrin. 2021. *Nordic BERT*. https://github.com/certainlyio/nordic_bert. Accessed: 2024-01-24.
- Schuldt, Jonathon P. 2016. ““Global Warming” versus “Climate Change” and the Influence of Labeling on Public Perceptions.” *Oxford Research Encyclopedia of Climate Science*, <https://doi.org/10.1093/acrefore/9780190228620.013.309>.
- Similarweb. 2023. *Top Websites Ranking in Denmark*. <https://www.similarweb.com/top-websites/denmark/>. Accessed: 2024-01-11.
- Soutter, Alistair Raymond Bryce, and René Möttus. 2020. ““Global warming” versus “climate change”: A replication on the association between political self-identification, question wording, and environmental beliefs.” *Journal of Environmental Psychology* 69:101413. ISSN: 0272-4944. <https://doi.org/https://doi.org/10.1016/j.jenvp.2020.101413>.
- UNFCCC. 1992. *ARTICLE 1 DEFINITIONS*. <https://unfccc.int/resource/ccsites/zimbab/conven/text/art01.htm>. Accessed: 2024-01-26.

6. Online Resources

The R code, especially the function to call the Hugging Face Inference API are available via GitHub.