Analyzing Online Discourse on Climate Change Using Word2Vec Embeddings

Deep Learning for Social Sciences

Tobias Weitzel (1358993), Jannik Wirtheim (1398359), Fabian Mahner (1005680)

Repository: `https://github.com/tobiaswtzl/dlss-project24`

27 August 2024

1 Introduction

In recent years, online discourse surrounding climate change has garnered increasing attention, particularly on social media platforms like Reddit (Treen et al., 2022). This study explores the evolution of climate change discussions on Reddit by first analysing data from 2010 to 2022 to provide descriptive insights. Secondly, shifts in the thematic discourse are identified through the use of different word embedding models, with the outputs subsequently compared and discussed. Third, three hypothesis are formulated based on scientific literature and tested using the previously described data and models.

2 Data Collection and Preliminary Analysis

2.1 Data Collection

The Reddit Climate Change Dataset is employed, which includes all Reddit posts containing the terms "climate" and "change" from 2010 until 1. September 2022 (Pavellexyr, 2020). The dataset was selected on the basis of its accessibility, the substantial number of posts it encompasses (620,908), its extensive temporal scope, and its subreddit structure, which facilitates sophisticated analysis across subreddits. Subreddits are online communities that are created and moderated by users, with a specific focus on a particular topic. The analysis focused only on posts[1]. Additionally, the comments were dropped, as including them would have resulted in a 16-fold increase in data size, which would have made model training more complex, particularly given the limited computational resources available.

2.2 Data Cleaning and Preprocessing

Given that the majority of Reddit users are from English-speaking countries (Statista, 2024), only English posts were retained. Additionally, usernames were replaced with "*username*" to reduce vocabulary size and links were removed. Stopwords were retained to preserve the full context for word embeddings, as their removal would strip away crucial contextual elements. For instance, the sentence "*I don't know how to solve this problem.*" without stopwords is "*know solve problem.*", changing the context completely. Furthermore, the texts are converted to lowercase and then tokenized using SpaCy (ExplosionAI, 2024), keeping only alphabetic characters, question marks and exclamation marks[2]. Additionally, this preprocessing step eliminates all emojis. However, the loss of information is minimal since emojis are seldomly utilized on Reddit[3]. SpaCy is designed to preserve semantic integrity by respecting linguistic boundaries and ensuring that words are tokenised in a way that retains their meaning and context, taken language-specific details into account as well as preserving contractions and punctuation. In contrast, Byte-Pair Encoding (BPE), Uniform and Wordpiece employ subword tokenisation[4], which can introduce noise into word-level embeddings due to the splitting of words into smaller units that may not retain their

[1] A posts consists of a title and optionally also a body of text, a link, image or video. The dataset includes solely text data and no images or videos.

[2] This is crucial as in particular question marks and exclamation points are more often employed to emphasize calls to action (Haim et al., 2021).

[3] The reasons are e.g. discussed here `https://tinyurl.com/265wz9pz`.

[4] See Appendix A.2 for vocabulary sizes.

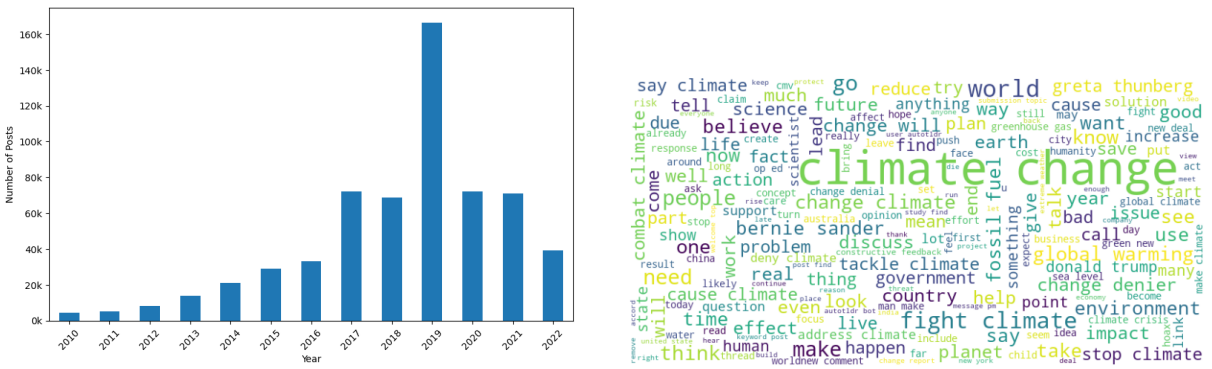
original meaning when isolated. Consequently, the data was tokenised with SpaCy. Lemmatization using SpaCy is then applied to the cleaned text to reduce words to their base forms, allowing for more standardized analysis by removing variations of the same word. After preprocessing, the corpus consists of 11 million tokens and a vocabulary of 9990 words.

2.3 Preliminary Data Analysis and Visualization

Before training the models and testing the hypothesis, a short description of the data is given[5]. Although the number of Reddit users has steadily grown over time (Richter, 2018), the histogram of posts over the years in Figure 1a shows that the prominence of the topic ”climate change” has varied throughout the past decade. There was a noticeable rise in interest starting in 2010 with a large peak in 2019. However, there is sharp drop in popularity from 2020 on. Without data on posts not mentioning climate change, any conclusions about that change are speculative. However, it’s plausible that the COVID-19 pandemic and the Russian war of aggression have supplanted climate change as pressing concerns in people’s minds.

The analysis of word frequency in Figure 1b and Appendix A.3 provides deeper insights into the overarching topic. In addition to the keywords ”climate” and ”change” words such as ”people”, ”world”, and ”global” frequently appear in the dataset, reflecting the global significance and underlying urgency conveyed in the texts. Various dimensions of climate change are highlighted through this frequency analysis: For instance, the prominence of words like ”political”,”government”, ”policy” and ”Trump” suggests a focus on the role of policymakers and their influence in addressing climate change. Conversely, words like ”science” and ”scientist” underscore the importance of research, while terms like ”care”, ”together” and ”believe” emphasise the need for a global effort without leaving anyone behind.

As climate change and possible solutions are threatening peoples’ health, jobs and futures (in)directly, discussions can get emotional. Thus, we analysed and classified emotions over time, which clearly showed a trends towards negative emotions over time[6].



(a) Number of Posts per Year (b) Most frequent (lemmatised) words

Figure 1: Descriptive Analysis of Subreddit Activity: Annual Post Volume and Most Frequent Words

[5]More plots can be found in Appendix A.1.

[6]Figures can be found in Appendix A.4 due to space limits.

3 Deep Learning Model

3.1 Model Architectures and Options

Both a continuous-bag-of-words (CBOW) and skipgram model were trained, but the CBOW model was ultimately chosen for its performance and computational efficiency. The CBOW model is a bag-of-words model[7] that predicts the current word based on the *n* surrounding words by averaging their embeddings and passing them through a single-layer neural network with one non-linear hidden layer. The value(s) of this layer - whose dimensionality is user defined - for each word are the word embeddings. The vocabulary was constrained to words that occurred at least 40 times, resulting in a total of 9,990 words. The produced word embeddings are static, i.e. each word has only one embedding, independent of the multiple contexts a word can be used in (Mikolov, 2013). Since our analysis compares word meanings across time and subreddits, the word embeddings need to vary by time and subreddit Thus, one model was trained and tuned[8] on the whole corpora. The final parameters were chosen based on minimum loss. The accuracy and F-score of the final model on the validation set was 0.18 on the validation and test set[9]. Subsequently, various models were fine-tuned based on the main model using yearly and subreddit-specific data to produce the respective embeddings, e.g. for "*r/askscience*".

3.2 Model Evaluation and Choice

Ultimately, the CBOW model was chosen for its faster training times and superior performance, which was particularly important given the limited computational resources on Google Colab. Additionally, the CBOW model produced higher-quality word embeddings. To assess embedding quality, two metrics were used: first, 20 analogies with words from the model's vocabulary were created, and the model was tasked with predicting the last word. Both skipgram and CBOW failed this task completely, predicting no analogy correct, not being close most of the times. Second, quality was also assessed on a continuous scale by calculating the correlation with human word similarity scores from the crowd-sourced WordSim353 dataset (Finkelstein et al., 2001). Here, CBOW performed better (0.07) than the skipgram model (0.00). This low correlation is an indicator for low embedding quality, though, this is not surprising considering the limited computational resources available: The first CBOW model by Mikolov (2013) was trained on a 6 billion token corpora with a vocab size of one million. Training was conducted on an 11 million token corpus with a vocabulary size of 9,990. Crucially, the model still manages to capture real word relationships between words, as can be seen in section 3.4.

3.3 Benchmark Models from MTEB Leaderboard

To further evaluate the model's performance, a comparison was made with state-of-the-art pre-trained embedding models from the Massive Text Embedding Benchmark (MTEB) leaderboard. Two models were selected from the leaderboard to serve as benchmarks. Given the constraints on computational resources, it was crucial to choose models that were relatively small and compatible with the available infrastructure, still, we had to use a random sample of 100,000 posts. As a result, the first model chosen was a GloVe model, an extension of Word2Vec (Reimers and

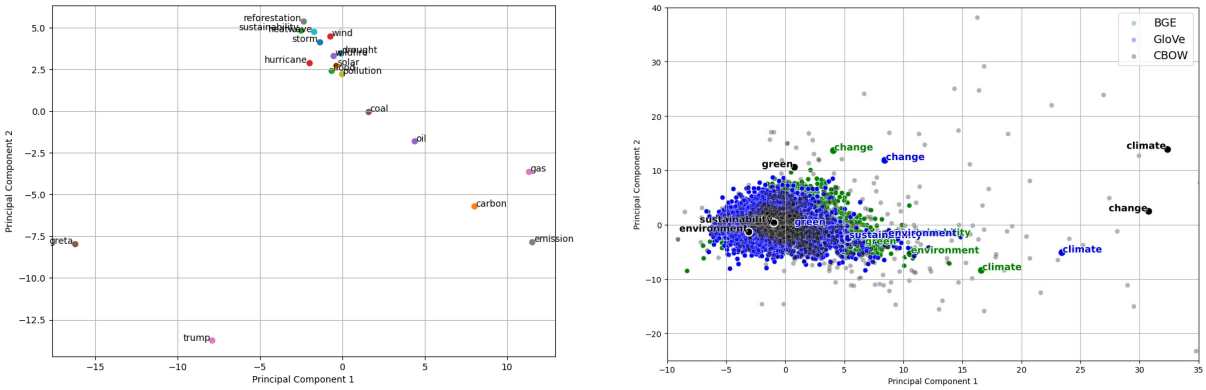
[7] Which means the word order of the context words does not matter.
[8] For the sampling space, see Table 2 in the Appendix B.1.
[9] The tuned skipgram model achieved an accuracy and F-score of 0.06 on validation and test data.

Gurevych, 2019). The second model selected was the smaller version of the BGE model. This choice was made because the larger version currently holds the top position on the MTEB leaderboard (Xiao et al., 2023).

3.4 Exploration of semantic similarity and word meaning

Word meaning is operationalized as word embeddings, which are *n*-dimensional vector representations of words that capture their meaning in an *n*-dimensional space and enable arithmetic operations (Mikolov, 2013). A basic definition of word meaning, aligned with the model's training process, is followed: "[t]wo words are considered to have the same meaning when they are used in similar contexts" (Feder et al., 2022, p. 81). Consequently, if words are used in different contexts, their meaning differs [10].

Principal Component Analysis (PCA) is used to analyze and visually interpret the word embeddings, which are normalized to ensure that each feature has the same impact. The PCA visualized in Figure 2a illustrates how different terms related to climate change discussions relate to each other in a two-dimensional space. The terms "reforestation", "sustainability", "drought" and "pollution" are situated in close proximity to one another at the origin of the plot, suggesting that the CBOW model has discerned their inherent substantive proximity. Interestingly, the negative effects of climate change like "droughts" are located close by possible solutions for climate change like "reforestation", suggesting that discussions of climate crises are constructive and discuss possible solutions. Distinctively away from these words are fossil-related terms such as "coal", "emission" and "gas", demonstrating that the model understands their distinctive substantive difference. Of note is the positioning of "Trump" and "Greta" far away from all other words, suggesting that people are positioned in a separate space. Interestingly, the distance between "Trump" and "Greta" is similar to the difference between "reforestation" and "coal", which reflects their distinct and disparate substantial positions in discussions.



(a) Words in 2-dim space (b) Word Embeddings across Models

Figure 2: PCA for Word Embeddings (CBOW and Comparison with BGE and GloVe)

3.5 Comparison with state-of-the-art-embeddings

The performance of GloVe and BGE was also assessed on the analogy task: GloVe correctly predicted 21.5% of all analogies, while BGE, which was expected to perform better, achieved only 5.26%. However, even the incorrect predictions from both models were usually very close. The

[10] see Appendix B.2.

continuous correlation with human word similarity scores from the crowd-sourced WordSim353 dataset (Finkelstein et al., 2001) was therefore also inspected. As expected, both models performed significantly better: GloVe achieved a correlation of 0.31, while the BGE model achieved 0.4. However, since these technical differences do not provide insights into the substantial differences in assessed word meanings, these were also compared by plotting the embeddings from different models in Figure 2b. The majority of words are tightly clustered together, indicating that most word embeddings are relatively similar, such as ”environment” ”global” and ”warming”. However, some CBOW embeddings (”climate”, ”change”) deviate significantly from the central cluster, while BGE and GloVe are quite close. Overall, the embeddings are relatively similar, with only some CBOW predictions differing notably.

4 Social Science Research Analysis

Climate change is one of the most pressing crisis humanity has faced so far. Thus, climate change and its solutions are intensively discussed, including on social media. However, after Covid-19 and multiple other crises around the globe, climate change seems to be less of a salient topic. For instance, while 59 % of people mentioned climate change as the most important problem in Germany in 2019, this number shrunk to 24 % in 2024 (Wahlen, 2024) . Similarly, green parties seem to be losing voters' favour, while private companies appear to be abandoning their sustainability goals as well, as seen in BlackRock's recent retreat from some of its ESG commitments (Masters and Bryan, 2024). Recent literature supports the observation that climate change is losing its foothold in public discourse. Studies indicate that while climate change was a dominant concern in the late 2010s, its perceived urgency has diminished as other crises, such as the Covid-19 pandemic and geopolitical conflicts, have taken center stage (Marlon et al., 2020) Researchers have found that media coverage of climate issues has also decreased, leading to a lower salience of the topic among the public (Spisak et al., 2022). Moreover, political scientists have noted a shift in voter priorities, with environmental issues being overshadowed by immediate economic and health concerns (Abou-Chadi and Kayser, 2017). Additionally, the discourse around climate change has become less scientifically rigorous over time, with more focus on political and economic aspects rather than scientific evidence (Lamb et al., 2020). This trend is further compounded by the retreat of major corporations from their sustainability pledges, as they face mounting pressures to focus on short-term profitability rather than long-term environmental goals (e.g., Sjåfjell (2016). While these papers offer interesting insights, most are outdated and focus on platforms, or time frames, leaving three key questions unanswered.

1. How is the understanding of climate change changing over time?
2. How are the topics in climate change discussions evolving?
3. How is the scientific rigor of climate change discussions evolving?

Building on these questions and the presented literature, three hypotheses are formulated, which will be tested in the following sections:

1. H1: The meaning of climate change related words changes over time.
2. H2: Climate change related topics vary in salience over time.
3. H3: Climate change related discussions grow less scientific over time.

4.1 How is the understanding of climate change evolving over time?

To analyze how climate change-related words evolve over time, PCA was used to reduce the word embeddings to their principal component, which was then plotted over time in Figure 3. To maintain focus, the analysis centers on the changing meanings of four key terms— 'climate' 'change' 'environment' and 'green'. The trajectories of "climate" and "change" show considerable variability, with notable peaks and troughs, indicating significant shifts in their semantic contexts over the 12-year period. In contrast, the first principal component values for "environment" and "green" remain closer to zero, suggesting these terms have exhibited greater stability in usage and meaning. The years 2012, 2018, and 2020 are marked by notable shifts, particularly for 'climate' and 'change' potentially reflecting changes in public discourse or major events influencing their contextual meanings. This could be explained by Downs' Issue-Attention Cycle Theory, which suggests that the meanings of words evolve as public attention shifts during crises (Downs, 2016).

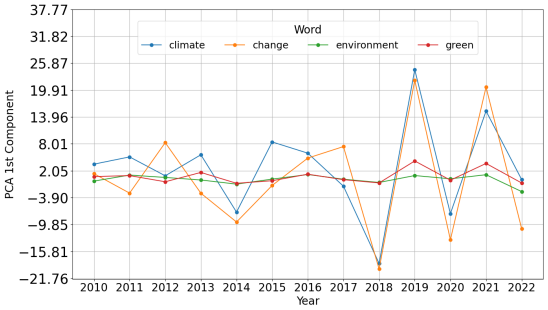


Figure 3: Change in Word Meaning Over Time

4.2 How are the topics in climate change discussions evolving?

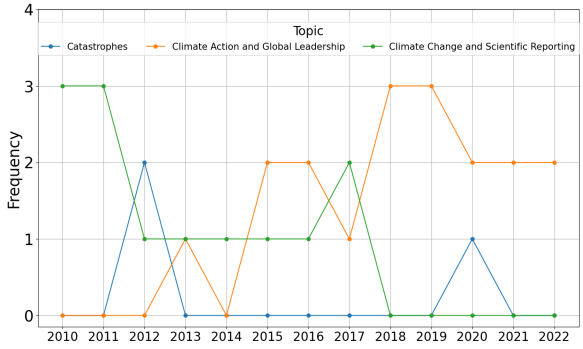
To gain a high-level overview of how discussions about climate change have changed over time, topic modeling was employed using k-means clustering with $k = 3$ to identify the main topics each year.[11]. The change of topics over time, plotted in Figure 4a provides evidence regarding H2: Climate change topics vary clearly over time. While catastrophes are relevant when they occur (Hurricane Sandy in 2012, Covid-19 in 2020), climate changes and scientific reports were more of a topic in the early 2010s and diminish in salience later, providing evidence in favour of H3. Around 2015, climate action and global leadership becomes more salient, mostly driven by US-politics as in-depth analysis shows, whose discussion is beyond the scope of this report. These findings can be interpreted in two ways: Either, reddit users moved on from consuming scientific information towards discussing solutions, or the focus of the discussions shifted from scientific knowledge towards people and day-to-day politics. The shift in climate change discussions, from scientific reports to political and action-oriented discourse, also aligns with Downs' Issue-Attention Cycle, where public attention peaks during crises and shifts towards more practical solutions and politics as the perceived urgency diminishes (Downs, 2016).

4.3 How is the scientific rigor of climate change discussions evolving?

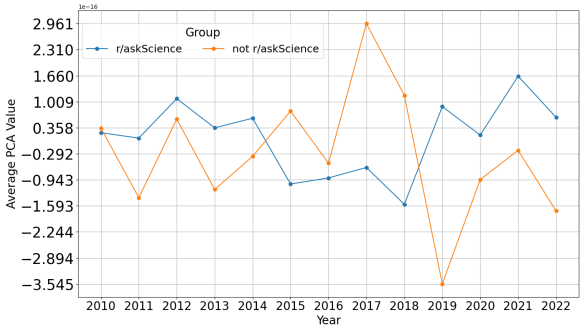
To analyze how scientific rigor has changed over time, a benchmark for scientific discussions was first established. The benchmark was defined using all posts from the *r/askscience* subreddit, a community where scientists pose and answer questions, as the authors believed that discussions in this subreddit are among the most scientific on Reddit. The CBOW model was then fine-tuned on *r/askscience* and all non-*r/askscience* posts. Specifically, the analysis compared how far

[11] After manual and iterative inspection, we identify the following four overarching topics in Table 4, which is found in the Appendix C.1.

the embeddings of *r/askscience* posts are from all remaining Reddit posts and how this distance changes over time. This analysis is illustrated in Figure 4b. The plot displays the average PCA value (first component) for *r/askscience* and non *r/askscience* subreddits. While interpreting each value in isolation can be challenging, comparing them reveals insights into their similarity or dissimilarity. Notably, in 2017 and 2019, the discrepancy between the average values is most pronounced. This indicates that discussions became less scientific on reddit in this time period. However, the distance decreased in 2020 again, i.e. discussions became more scientific again, and stayed relatively constant until 2022. We conjecture that this may be due to the weaker focus on climate change after 2019[12], which might have led to less discussions among non-scientists while scientists continued to discuss climate change. This pattern aligns with Agenda-Setting Theory, which suggests that media and public attention to topics like climate change can influence the frequency and nature of discussions within social forums (McCombs and Shaw, 1972).



(a) Frequency of Topics over Time



(b) Change in Scientificness over Time

Figure 4: Analysis of Discussions over Time

5 Conclusion and Critique

The findings of this study are relevant not only to policymakers, but to all people having an interest in mitigating climate change as discussions shape public opinions which, in turn, shape policies. The data shows that discussions are less scientific now compared to 2010, which is not a good development, as scientific discussions do not only inform people the most, but also prevent emotional discussions and polarisation - both of which are common in climate change discussions nowadays. Moreover, the analysis of topics showed that the focus of discussions shifted from scientific reports to climate action and political actors. Although the lower saliency of climate reports can be seen as a sign of a weaker role of science, the shift towards climate action topics can also be seen positively, as action is underway. More research is necessary to investigate whether discussions actually focus on constructive solutions or rather get caught up in daily (party) politics. However, several limitations of this report must be acknowledged: The discourse on climate change is inherently multifaceted (Willis, 2017), encompassing scientific, political, economic, and social dimensions, making it challenging to capture comprehensively through (simple) linguistic analysis alone. Additionally, the posts stemming only from reddit certainly introduces sampling bias while our limited computational resources led to suboptimal results. Moreover, we lacked data about the users for causal/correlational analysis. Despite these limitations, this report still serves as a stepping stone for further research.

[12] See the lower number of posts after 2019 in Appendix A.3

A Appendix: Data Collection and Preliminary Analysis

A.1 Data Overview

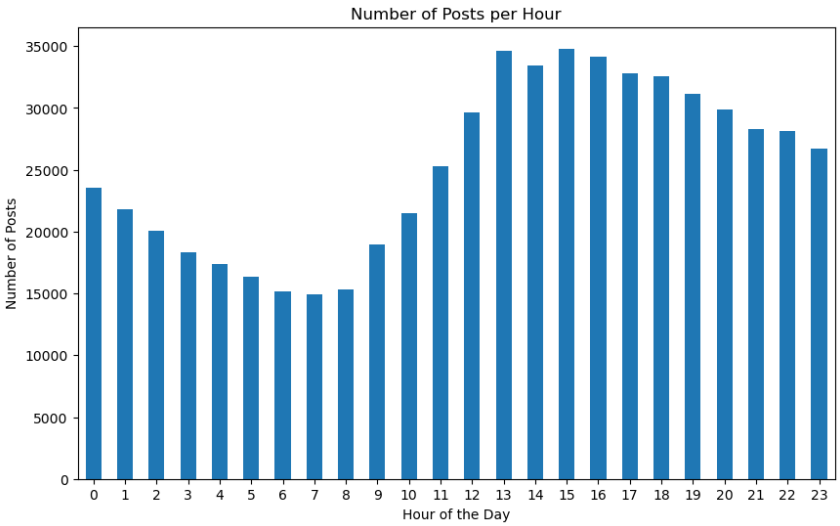


Figure 5: Posts per Hour

A.2 Comparison of Tokenisers

Tokenizer	Vocab Size
BPE	10000
Unigram	14797
WordPiece	10000
spaCy	14818

Table 1: Tokenizer Vocab Sizes for 2000 Posts

A.3 Frequency Analysis

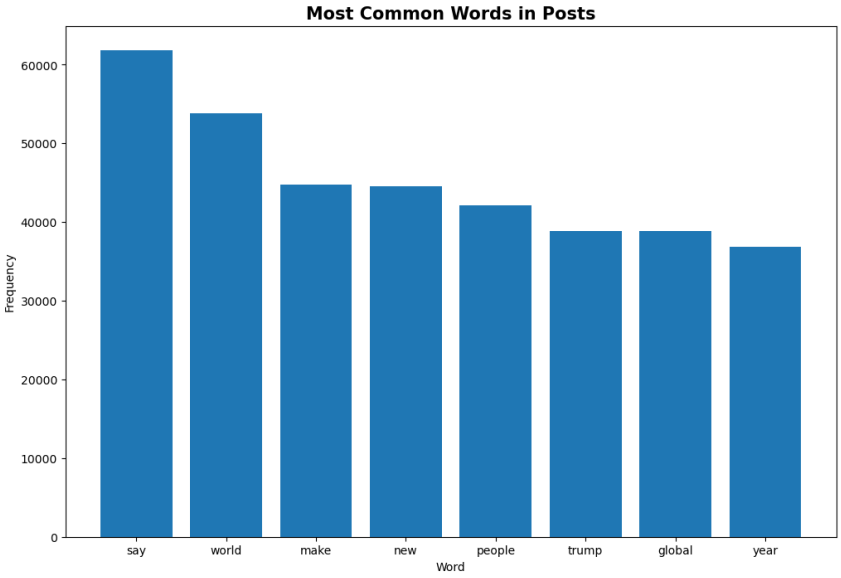


Figure 6: Most Frequent Words

A.4 Sentiment Analysis

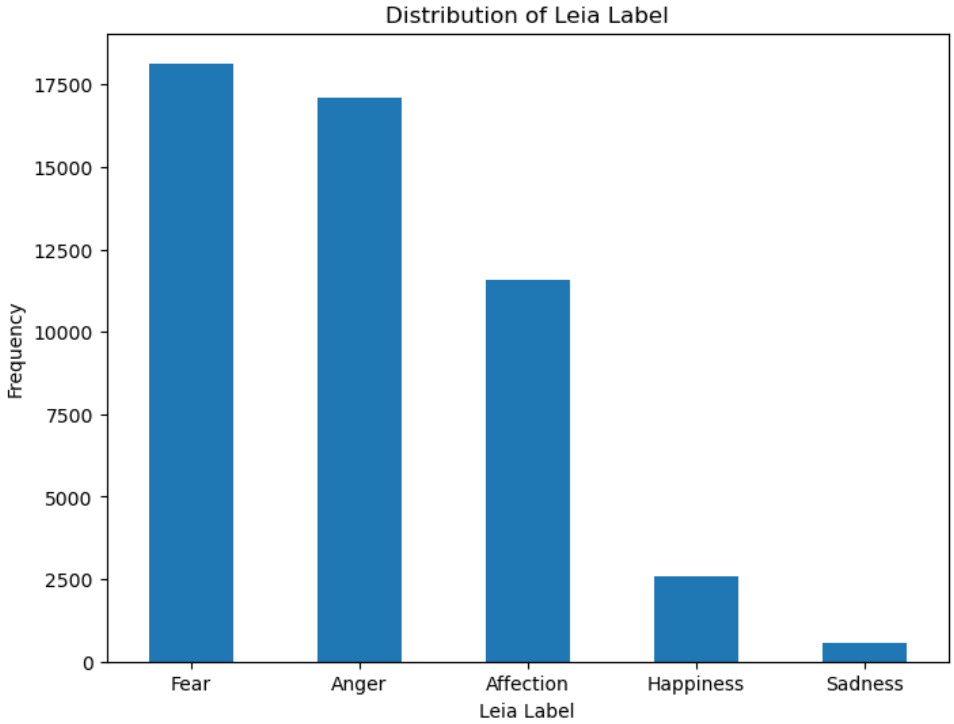


Figure 7: Distribution of LEIA Labels

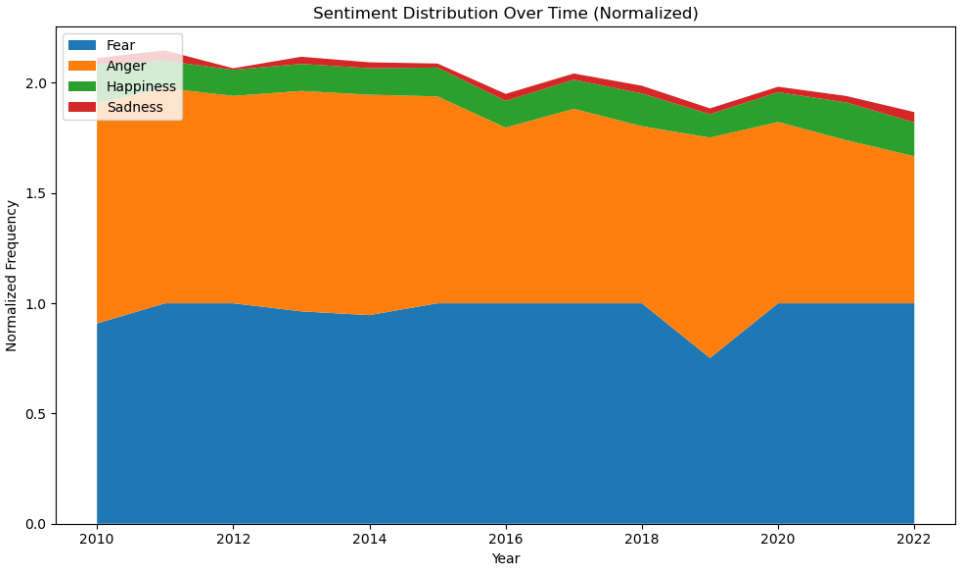


Figure 8: Normalized Emotion Sentiment

B Appendix: Deep Learning Model

B.1 Model Tuning

Parameter	Possible Values	Final Parameters
Context Size	2, 3, 4, 5	2
Dropout	0.1, 0.2, 0.3, 0.4, 0.5	0.2
Learning Rate (lr)	0.01 - 0.0001	0.0092
Batch Size	64, 128, 256, 512, 1024	512
Epochs	50, 60, 70, 80, 90, 100, 110, 120, 130, 140	120
Patience	5, 10, 15	10
Min Delta	0.01 - 0.0001	0.0006
Gamma	0.1, 0.25, 0.5, 0.75, 0.9	0.9
Step Size	5, 10, 20	5
Weight Decay	0.01 - 0.0001	0.0001
Embedding Dimension	100, 500, 1000	100

Table 2: Hyperparameter Space for Tuning and Final Parameters.

Figure 9: Loss Curves CBOW

B.2 Cosine Similarities

Subreddit	Term	Similar Word	Similarity
askscience	carbon	water	1.00
		heat	1.00
		fossil	1.00
	climate	it	0.76
		temperature	0.76
		scientific	0.75
	warming	cooling	1.00
		deforestation	1.00
		ecological	1.00
democrats	carbon	cut	1.00
		aggressive	1.00
		school	1.00
	climate	its	0.54
		coal	0.53
		green	0.52
	warming	american	1.00
		when	1.00
		both	1.00
republicans	carbon	policy	1.00
		because	1.00
		there	1.00
	climate	time	0.76
		fight	0.76
		expert	0.76
	warming	denier	1.00
		because	1.00
		serious	1.00
NOT_askscience	carbon	greenhouse	0.63
		gas	0.62
		cut	0.59
	climate	harry	0.38
		traditional	0.38
		youtube	0.37
	warming	temperature	0.59
		cooling	0.55
		fahrenheit	0.39

Table 3: Similarity scores of terms in different subreddits.

C Appendix: Social Science Research Analysis

C.1 Topic modeling

Topic Name	Top 5 Words
Climate Change and Scientific Reporting	global, warming, science, new, report
Climate Action and Global Leadership	action, leadership, global, fight, plan
Catastrophes	coronavirus, real, stop, global, hurricane
Global Energy and Environmental Science	energy, science, global, environmental, world

Table 4: Summary of Topics and Their Top 5 Words

References

Abou-Chadi, T. and Kayser, M. A. (2017). It's not easy being green: Why voters punish parties for environmental policies during economic downturns. *Electoral studies*, 45:201–207.

Downs, A. (2016). Up and down with ecology: The "issue-attention cycle". In *Agenda setting*, pages 27–33. Routledge.

ExplosionAI (2024). spacy 101: Everything you need to know. Accessed: 2024-08-27.

Feder, A., Keith, K. A., Manzoor, E., Pryzant, R., Sridhar, D., Wood-Doughty, Z., Eisenstein, J., Grimmer, J., Reichart, R., Roberts, M. E., et al. (2022). Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2001). Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414.

Haim, M., Karlsson, M., Ferrer-Conill, R., Kammer, A., Elgesem, D., and Sjøvaag, H. (2021). You should read this study! it investigates scandinavian social media logics □. *Digital Journalism*, 9(4):406–426.

Lamb, W. F., Mattioli, G., Levi, S., Roberts, J. T., Capstick, S., Creutzig, F., Minx, J. C., Müller-Hansen, F., Culhane, T., and Steinberger, J. K. (2020). Discourses of climate delay. *Global Sustainability*, 3:e17.

Marlon, J., Howe, P., Mildenberger, M., Leiserowitz, A., and Wang, X. (2020). Yale climate opinion maps 2020. *Yale program on climate change communication*, 2.

Masters, B. and Bryan, K. (2024). Blackrock's support for esg measures falls to new low. *Financial Times*. Accessed on August 27, 2024.

McCombs, M. E. and Shaw, D. L. (1972). The agenda-setting function of mass media. *Public opinion quarterly*, 36(2):176–187.

Mikolov, T. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Pavellexyr (2020). The reddit climate change dataset. Data set.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Richter, F. (2018). The number of subreddits continues to grow. Statista.

Sjåfjell, B. (2016). Achieving corporate sustainability: What is the role of the shareholder? *Hanne Birkmose, Shareholders' Duties in Europe (Kluwer Law International, 2017)*.

Spisak, B. R., State, B., van de Leemput, I., Scheffer, M., and Liu, Y. (2022). Large-scale decrease in the social salience of climate change during the covid-19 pandemic. *Plos one*, 17(1):e0256082.

Statista (2024). Länder mit dem höchsten anteil am traffic von reddit. Retrieved August 18, 2024.

Treen, K., Williams, H., O'Neill, S., and Coan, T. G. (2022). Discussion of climate change on reddit: Polarized discourse or deliberative debate? *Environmental Communication*, 16(5):680–698.

Wahlen, F. (2024). Was ist ihrer meinung nach gegenwärtig das wichtigste problem in deutschland? (zwei nennungen waren möglich; ausgewählte probleme). Accessed on August 27, 2024.

Willis, R. (2017). Taming the climate? corpus analysis of politicians' speech on climate change. *Environmental Politics*, 26:212 – 231.

Xiao, S., Liu, Z., Zhang, P., and Muennighoff, N. (2023). C-pack: Packaged resources to advance general chinese embedding.