Augmented CARDS: A machine learning approach to identifying triggers of climate change misinformation on Twitter

Cristian Rojas[1], Frank Algra-Maschio[2], Mark Andrejevic[3,6], Travis Coan[4], John Cook[5*], Yuan-Fang Li[1,6*]

[1]Department of Data Science & AI, Monash University, Clayton, 3800, Victoria, Australia.
[2]School of Social and Political Sciences, Monash University, Clayton, 3800, Victoria, Australia.
[3]School of Media, Film, and Journalism, Monash University, Clayton, 3800, Victoria, Australia.
[4]Exeter Q-Step Centre, University of Exeter, Exeter, UK.
[5]Melbourne Centre for Behaviour Change, University of Melbourne, Parkville, Victoria, Australia.
[6]Monash Data Futures Institute, Monash University, Clayton, 3800, Victoria, Australia.

*Corresponding author(s). E-mail(s): jocook@unimelb.edu.au; Yuanfang.Li@monash.edu;
Contributing authors: cristian.rojascardenas@monash.edu; Frank.Algra-Maschio@monash.edu; Mark.Andrejevic@monash.edu; T.Coan@exeter.ac.uk;

Abstract

Misinformation about climate change poses a significant threat to societal well-being, prompting the urgent need for effective mitigation strategies. However, the rapid proliferation of online misinformation on social media platforms outpaces the ability of fact-checkers to debunk false claims. Automated detection of climate change misinformation offers a promising solution. A representative approach is the CARDS model, capable of identifying and classifying a wide range of contrarian claims related to climate change. However, the CARDS model was trained on misinformation from blogs and conservative think-tanks, leaving

its performance on social media platforms untested. In this study, we address this gap by developing a two-step hierarchical model—the Augmented CARDS model—specifically designed for detecting climate misinformation on Twitter. The Augmented CARDS architecture modularizes the classification tasks to enhance the pipeline performance while addressing the imbalanced data distribution. In addition, it implements a better model with disentangled attention. Moreover, we introduce additional data and categories relevant to the Twitter context. Furthermore, we apply the Augmented CARDS model to five million climate-themed tweets over a six-month period in 2022. We find that over half of climate misinformation on Twitter involves attacks on climate actors or conspiracy theories. Spikes in climate misinformation are triggered by one of four stimuli: political events, natural events, contrarian influencers, or convinced influencers. Implications for automatic responses countering climate misinformation are discussed.

Keywords: climate change, misinformation, machine learning

1 Introduction

Misinformation about climate change causes a number of negative impacts. It reduces public support for mitigation policies [1] and thwarts efforts to communicate accurate information [2]. Misconceptions about the prevalence of contrarian views have a self-silencing effect [3]. While misinformation has an overall impact of reducing climate literacy [1], this effect varies across the political spectrum, resulting in exacerbated polarisation [4].

Social media platforms have become an active site for the spread of misinformation on a wide range of topics and have received increased scrutiny for their role in undermining trust in scientific and journalistic expertise [5]. At the same time, these platforms are becoming an increasingly significant source of news and information that have an important role in shaping public awareness and discussion of issues of social importance [6]. The decentralized and networked character of the internet lowers the barriers to posting and sharing misinformation, which ends up being further amplified by engagement-maximizing commercial algorithms [7]. Regulatory regimes that protect social media platforms from editorial responsibility contribute to the "wild west" information environment in which contrarian claims circulate alongside and often more widely than traditional forms of journalistic and scientific consensus [6]. Social media also serves as a conduit for mainstreaming contrarian claims when their prevalence online results in their being taken up by news outlets and political actors [8]. The problems caused by the spread of misinformation online are likely to be exacerbated by recent advances in generative artificial intelligence. As an executive of a company that tracks misinformation online put it, "Crafting a new false narrative can now be done at dramatic scale, and much more frequently — it's like having A.I. agents contributing to disinformation" [9].

The climate change discussion has long been a key target of misinformation. As automated systems contribute to the generation and circulation of contrarian claims,

there will be an increased need for automated detection, tracking, and response. One result will be increased pressure on journalists, platforms, watchdogs, and regulators to find ways of keeping pace with the spread of such claims. For the purposes of addressing the challenges posed by contrarian information, it is useful to be able to determine the nature of false claims. Doing so makes it possible to provide a response that addresses the substance of the claim. The ability to identify and categorize claims also makes it possible to determine the prevalence of different types of misinformation in order to shape "pre-bunking" strategies for inoculating the public against particular categories of false claims [2].

It is imperative that interventions are developed and deployed to counter these negative impacts. However, this is made challenging by the fact that misinformation spreads through social media faster than factual information [10]. Further, once misinformation has taken hold, it is difficult to dislodge—a phenomenon known as the continued influence effect [11]. Consequently, solutions that can detect and respond to misinformation in a rapid fashion are required.

However, automatic detection and correction of misinformation are technically challenging, earning the label "the holy grail of fact-checking" [12]. There have been efforts to automatically detect and fact-check misinformation across various domains [13, 14]. On climate misinformation, there have been few efforts to detect misinformation. Unsupervised topic analysis has been employed to identify the major themes in conservative think-tank (CTT) texts [15], link corporate funding to polarizing climate text [16], and identify climate framings in newspaper articles [17]. There have also been efforts to detect logical fallacies in climate misinformation as well as across general topics [18–20].

The CARDS (Computer Assisted Recognition of Denial & Skepticism) model used supervised machine learning to detect and categorize contrarian claims about climate change [21]. The model has been shown to be effective in categorising a wide range of contrarian claims about climate change. The model was based on a taxonomy of contrarian claims consisting of five main categories: 1) global warming isn't happening, 2) humans aren't causing global warming, 3) climate impacts aren't bad, 4) climate solutions won't work, and 5) climate movement/science are unreliable). At the second level of this taxonomy are sub-categories of contrarian claims such as 5.2 (climate actors are unreliable) and 5.3 (conspiracy theories).

However, the CARDS model was only trained using text from contrarian blogs and conservative think-tank websites—prolific sources of climate misinformation—and its performance in classifying climate misinformation from other datasets (e.g., from social media platforms) has yet to be assessed. This study assesses and augments the CARDS model's performance in classifying climate misinformation in Twitter data. We apply the Augmented CARDS model to a dataset of climate tweets, in order to examine the various arguments that are characteristic of different types of misinformation peaks.

2 Methods

The original CARDS model was trained using a dataset comprising paragraphs extracted from sources known for their wealth of climate contrarian content, such

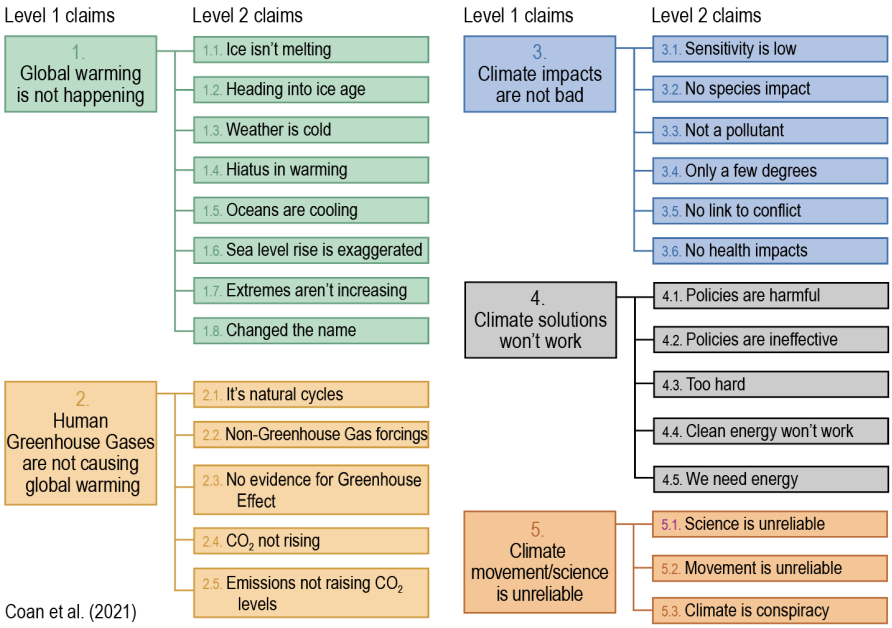


Fig. 1: CARDS taxonomy of contrarian climate claims [21].

as conservative think-tank articles and contrarian blog posts. This training approach showed strong performance when tested on similar content sources. Nevertheless, the model's ability to effectively differentiate between contrarian and convinced text (reflecting the scientific consensus on climate change) within the context of Twitter remained uncertain. To mitigate this uncertainty, we present an enhanced CARDS model introducing an initial binary classifier. This classifier's primary function is to distinguish between convinced and contrarian claims, aided by the inclusion of supplementary Twitter data. Subsequently, we include an additional layer responsible for classifying contrarian claims into their respective typology.

2.1 Model Architecture

Augmented CARDS enhances the performance of the original CARDS model on Twitter by utilizing additional data from the platform and rectifying category imbalances through a two-stage hierarchical architecture. Figure 2 illustrates the general model architecture, consisting of an initial layer trained in a binary detection task to differentiate between convinced and contrarian tweets, coupled with an additional layer trained in a multilabel task to classify the taxonomy.

Both classifiers incorporate the DeBERTa language model, structured based on the auto-encoding transformer architecture introduced in BERT [22]. The innovation includes disentangled attention and a more extensive pretraining process [23, 24]. Specifically, we utilized the large version of DeBERTa, consisting of 24 transformer blocks with a hidden size of 1024 and 16 attention heads. Additionally, an extra dense

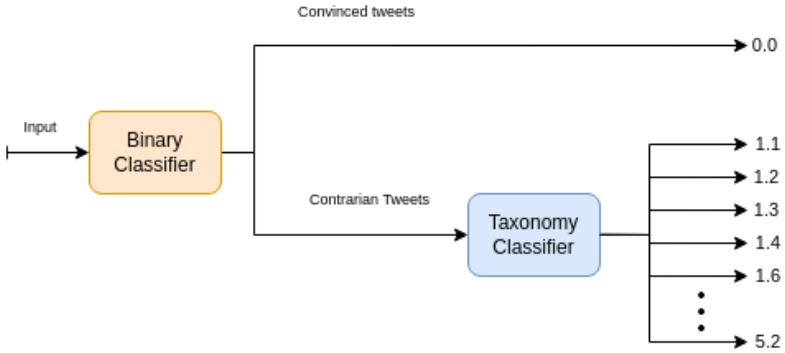


Fig. 2: Model Architecture.

layer was employed for the classification task, bringing the total number of parameters to approximately 355 million.

We aim to specialize the classifiers in their respective tasks, undergoing training tailored to their specific contexts. The implementation of a hierarchical architecture responds to the necessity of modularizing both tasks to effectively handle the fine-tuning process of the pipeline and improve its performance. Additionally, it mitigates the issue of unbalanced data distribution. Given that the datasets are overly dominated by convinced claims, the challenge lies in effectively detecting the remaining 17 classes of the taxonomy.

Moreover, DeBERTa's transfer learning capabilities are mainly attributed to its pretraining on web-sourced texts. Nevertheless, since Twitter was not incorporated into the pretraining procedure, fine-tuning is necessary to capture the linguistic features specific to the platform.

2.2 Training Details

To enhance the model's performance, we incorporated the Climate Change Twitter Dataset labeled by the University of Waterloo, featuring a 90/10 ratio of verified and misleading tweets, [25] to the binary classifier training set. Furthermore, the taxonomy classifier underwent training using the CARDS dataset, incorporating the 5.3 category ("climate change is a conspiracy theory"), which differed from original CARDS which merged category 5.3 with category 5.2. Separating these two categories was deemed appropriate due to the substantial prevalence of conspiracy theories in climate change tweets.

The models were fine-tuned over 3 epochs with a learning rate of 1e-5 in a v100 GPU with a batch size of 6. The input was constrained to sequences of 256 tokens with a padding method. These parameters, along with the seed were kept constant for comparison with the original CARDS method.

To assess the model's capabilities, climate change experts labelled a testing set of tweets following the [21] taxonomy. This dataset, denoted as *"Expert Annotated Climate Tweets"* in Table 1a, was composed of 2607 tweets related to climate change, sampled from the platform in the second half of 2022.

2.3 Data Analysis

The analysis was carried out on a large dataset of climate change-related tweets. This dataset was compiled between July and December of 2022 by the Online Media Monitor (OMM) at the University of Hamburg. We examined the temporal frequency of the data and identified intervals of interest based on distinct patterns discerned by the model. Within these intervals, a word frequency analysis was conducted and compared against the overall word frequency of the entire dataset. This comparison enabled us to highlight specific shifts in word usage during those periods and establish a connection between this information and relevant events that took place.

The word frequency was calculated by comparing the log-fold change and the p-value derived from the distribution differences for various words. Subsequently, a filter was applied to keep only those words with a significance level greater than 0.05, and they were ranked based on their log-fold change in descending order. Finally, the top 10 most relevant words were used to characterize the event (see Table B).

3 Results

3.1 Assessing the Augmented CARDS model

Table 1a compares the performance of the original CARDS and Augmented CARDS models in identifying contrarian claims in the original CARDS testing set (comprised of contrarian blogs and CTTs) and in our new Twitter dataset. We subdivide this task into two stages: binary detection (distinguishing between contrarian and convinced claims) and taxonomy detection (identifying claims from the CARDS taxonomy).

The original CARDS model performed exceptionally well in datasets sharing linguistic features with its original training data, including CTT articles and contrarian blog posts. This is demonstrated by the F1-score achieved in CARDS for binary detection (89.9), slightly outperforming Augmented CARDS.

However, in the taxonomy detection task, the original model showed a 5% performance decrease relative to the CARDS metrics [21]. This decline is attributed to the inclusion of the 5.3 category (contrarian claims involving conspiracy theories) in our analysis. This category is highly relevant in the Twitter context but was excluded from the original model in [21]. In this scenario, the Augmented CARDS architecture demonstrated better adaptability, achieving a 76.6 F1-score with additional data from Twitter, where climate change conspiracy arguments hold more significance among contrarians.

Nonetheless, our results indicate that Twitter is a challenging task due to the significant disparities in language and writing style observed between the original sources and the platform. On the other hand, the Augmented CARDS model achieves

Task	Datasets	Models		Support
		CARDS	Augmented CARDS	
Binary Detection	*CARDS*	**89.92**	89.05	4395
	Twitter Climate Change	68.13	**87.26**	2904
	Expert Annotated Climate Tweets	70.38	**81.63**	2711
Taxonomy Detection	*CARDS*	**77.42**	74.63	2904
	Expert Annotated Climate Tweets	47.04	**58.92**	2607

(a)

Category	CARDS	Augmented CARDS	Support
0.0	66.64	**79.56**	1049
1.1	68.75	**75.86**	28
1.2	**51.06**	36.92	20
1.3	**55.56**	54.41	61
1.4	**60.38**	50.00	27
1.6	**63.01**	58.97	41
1.7	58.23	**63.51**	89
2.1	67.62	**68.14**	154
2.3	**28.00**	23.53	22
3.1	25.00	**29.27**	8
3.2	**61.11**	58.82	31
3.3	**51.06**	45.90	23
4.1	33.57	**51.61**	103
4.2	13.70	**46.67**	61
4.4	47.37	**51.76**	46
4.5	27.27	**53.33**	50
5.1	33.87	**44.87**	96
5.2	27.39	**60.30**	498
5.3	48.63	**61.19**	200
Macro Average	46.64	**53.40**	2407

(b)

Table 1: Model Results. (a). Assessment of F1-scores achieved, comparing the original CARDS model with the Augmented CARDS Model. (b) F1-scores per category obtained from the Augmented CARDS model on the *"Expert Annotated Climate Tweets"* dataset.

a significant improvement in the F1-Score for both the *"Twitter Climate Change"* and the *"Expert Annotated Climate Tweets"* datasets for both tasks.

The technical advantages of Augmented CARDS included leveraging additional data from the Twitter context and addressing category imbalances through a hierarchical architecture. Based on these two factors, as shown in Table 1a, the Augmented CARDS model demonstrated a relative 16% performance improvement for binary detection and 14.3% for taxonomy detection on our *"Expert Annotated Climate Tweets"* dataset. This translates to an F1-score of 81.6 for binary detection and 53.4 for taxonomy detection, while maintaining a similar level of performance in the original

domain. Although there is still room for improvement, especially in taxonomy detection, it would require collecting a larger Twitter-based dataset for the less common categories in this context. Most of the categories with low F1-scores are infrequent on Twitter as illustrated in Table 1b.

In contrast, on Twitter, the most prominent contrarian categories are 5.2 (climate actors are unreliable), 5.3 (conspiracy theories), 4.1 (policies are harmful), 2.1 (global warming is naturally caused), and 1.7 (extreme weather isn't linked to climate change). Table 1b shows that our model exhibits the most substantial improvements with these categories. The F1-scores achieved by Augmented CARDS demonstrate an overall enhancement across most categories, with major improvement in the more relevant ones. Compared to blogs and CTT articles, the distribution of contrarian arguments on Twitter shows a different emphasis, with ad hominem fallacies (category 5.2) directed at climate actors being the most common type of argument. The second most common type of contrarian argument is conspiracy theories about climate change (category 5.3).

3.2 Applying Augmented CARDS to 2022 climate tweets

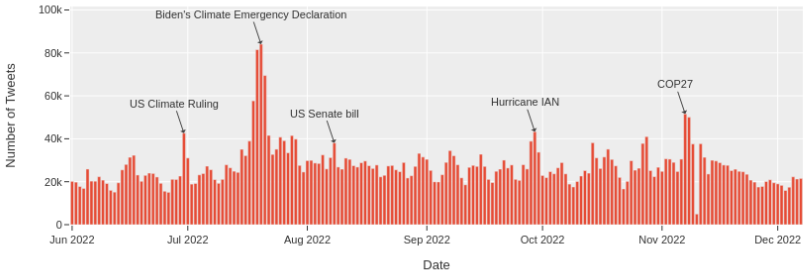
We applied the Augmented Cards model to over 5 million climate-related tweets in a six-month period in 2022, providing insight into the proliferation of climate-contrarian claims on Twitter. This novel investigation enabled an analysis of the triggers that caused an upsurge in contrarian claims on the platform and the most common types of contrarian claims.

The tweets used in our analyses were collected by the University of Hamburg by filtering for terms similar to "climate change" (see Appendix A). Figure 3a illustrates the daily frequency of tweets related to climate change, showing notable fluctuations in the frequency of climate tweets, such as the significant peak in late July. On average, 27,464 tweets per day are related to climate change in this data set, with the significant peaks in late July and mid-November resulting in 65,196 and 43,647 of tweets respectively.

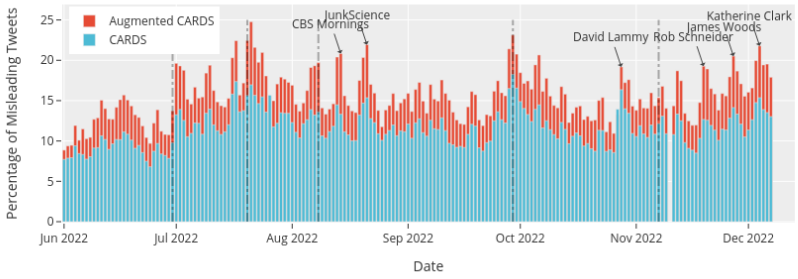
To investigate the causes of these peaks, we performed statistical analyses to identify words with major variations and establish correlations between these shifts and significant events that occurred during the corresponding periods. The word frequency analysis involved comparing changes in word distributions during specific periods in relation to the entire dataset. We computed the log fold change and p-value to assess differences in these distributions (see Appendix B for more details).

Between July 19 and 21, marking the period with the largest peak in climate tweets, the terms "climate emergency" and "Biden" showed the greatest shifts. Based on news reports from that time, these discussions occurred when it became apparent that President Joe Biden was considering declaring a climate emergency in response to the heatwave affecting both the United States and Europe [26].

The second-largest peak of overall climate tweets was associated with COP27, as indicated by the changes in word distribution illustrated in Table B2c. This event led to a doubling of the number of tweets between September 7th and 9th, 2022. The third highest peak of overall climate tweets in our dataset relates to Hurricane Ian [27]. Tweets relating to the Hurricane became the major topic of discussion related to



(a)



(b)

Fig. 3: (a) Number of tweets related to climate change topics by date in the 2022 period. (b) Percentage of misinformative tweets detected by the CARDS and Augmented CARDS models.

climate change between September 28 and October 1, although they generated only half the number of tweets compared to Biden's declaration event.

Turning now to the analysis of contrarian tweets, Figure 3b displays the percentage per day of contrarian tweets detected by the Augmented CARDS model through a binary inference process. This analysis indicates that the average proportion of contrarian tweets per day is 15.5%, yet there is clearly a number of peaks of contrarian tweets throughout the six-month period.

Overall, we identified four distinct categories of events that led to an upsurge in the publication of contrarian tweets, as outlined in Table 2. These events were categorized based on the nature of the event that triggered them. The triggers can be broadly classified into three primary groups: Natural Events, Political Events, and Influencer Posts.

Nature of trigger	Events
Natural Event	- Hurricane IAN
Political Event	- US Climate Ruling - Biden's Climate Emergency Declaration - US Senate bill
Contrarian Influencer	- Steve Milloy - Rob Schneider - James Woods
Convinced Influener	- Dan Rather - CBS Mornings - David Lammy - Katherine Clark

Table 2: List of occurrences that induced spikes in climate contrarism on Twitter.

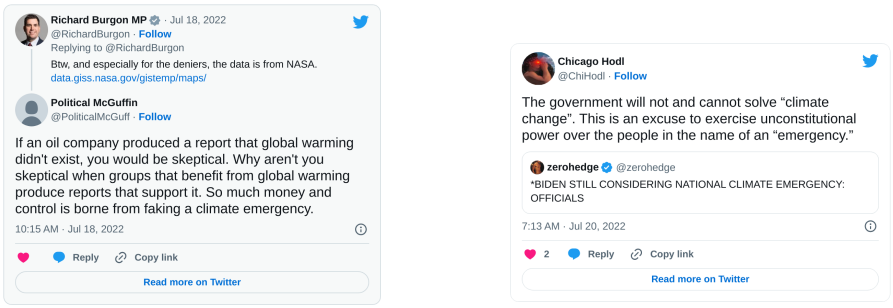


Fig. 4: Tweets sampled from the trend peak related to climate change observed between July 18 and July 21, 2022.

Natural Events, such as Hurricane IAN, and Political Events like COP27, were external occurrences originating outside the platform [28, 29]. They resulted in a general increase in public discourse surrounding the climate change topic and occasionally prompted shifts in contrarian positions.

For example, the Biden declaration was seized upon by climate change contrarians, triggering significant peaks in the percentage of contrarian tweets. In Figure 4, we present several examples illustrating some of the contrarian opinions. The primary concern revolved around the possibility that climate warming might be used as a political pretext to declare an emergency, potentially granting expanded powers to President Biden, which could disrupt the existing state equilibrium. This event caused the percentage of misinformation to surpass 20%, reaching a peak of 24.7%.

Similarly, the Natural Event of Hurricane Ian triggered an increase in all tweets related to climate change and the percentage of contrarian claims. Despite generating only half the number of tweets compared to Biden's declaration, the proportion of contrarian tweets related to Hurricane Ian reached similarly high levels. Discussions were centred around the impact of climate change on extreme weather conditions. Some examples of tweets are illustrated in Figure 5. For instance, FoxNews tweeted

one of their articles, titled "Democrats blaming climate change for Hurricane Ian at odds with science, experts say".

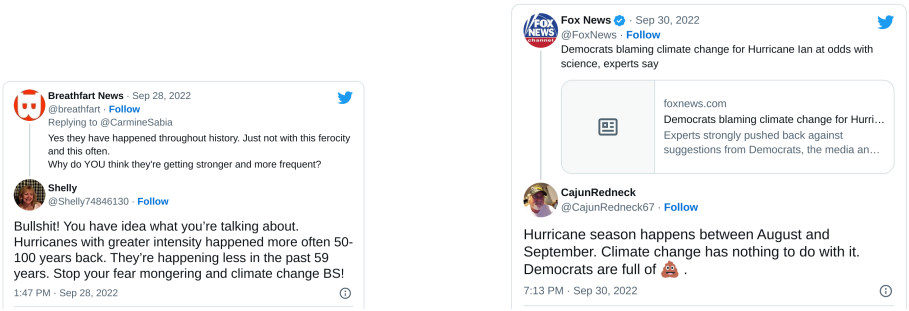


Fig. 5: Tweets that deny the link between extreme weather and climate change.

The COP27 event highlights a contrast between the emergency declaration and Hurricane Ian. While both of these latter events led to a rise in the volume of tweets and contrarian claims, COP27 triggered only a 1 per cent increase in contrarian claims.

While Political and Natural Events are likely to increase the volume of tweets related to climate change and, in some cases the percentage of contrarian claims, Influencer Posts triggered contrarian responses despite having no significant impact on the overall volume of climate change-related discussions. Further, contrarian claims increased in response to influencers regardless of whether they expressed a convinced or contrarian view. Peaks in the percentage of climate misinformation in Figure 3b don't necessarily correspond with a change in overall frequency of climate tweets in Figure 3a. These peaks are mainly induced by influencers on both sides, whether they are proponents or contrarians of climate change. These influencers could be politicians, comedians, film directors, or media figures. Nonetheless, they all share the characteristic of being public figures with a substantial number of followers, which is sufficient to trigger these fluctuations. It's important to note that our categorization of influencers is based on the positions adopted by their publications during 2022, not necessarily their current personal viewpoints.

Our final analysis is the categorisation of contrarian tweets by the typology of [21] as inferred by the Augmented CARDS model. Figure 6 represents the distribution of the tweets by the most common categories identified in the climate-related tweets. The distribution of contrarian categories remains relatively stable even on dates with significant deviations. The most common form of climate misinformation involves criticisms of climate actors such as climate scientists and environmentalists (category 5.2), comprising 40% of the total number of misleading tweets. This is followed by category 5.3, which includes tweets categorizing climate change as a conspiracy, making up approximately 20% of the segment. Categories 4.1 (climate policies are harmful) and 2.1 (natural cycles are causing global warming, not humans) make up the next two most relevant categories. The fifth most common category, 1.7 (extreme events

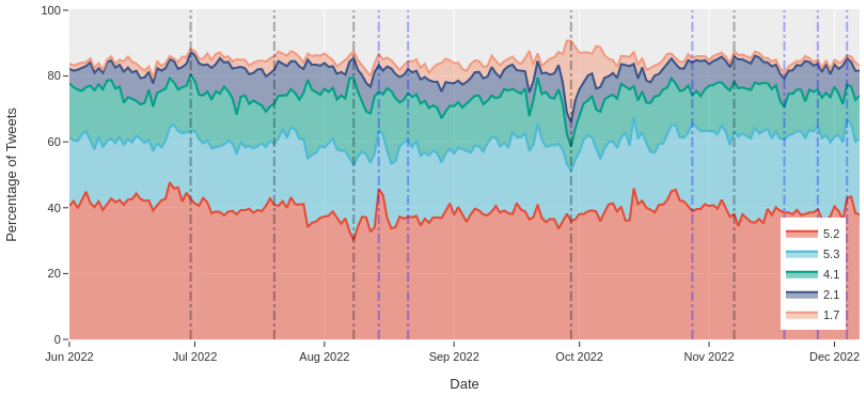


Fig. 6: Breakdown of the five most relevant categories predicted by the CARDS model in the Hamburg dataset.

Nature of trigger	**5.2**	**5.3**	**4.1**	**2.1**	**1.7**	**others**
Contrarian Influencer	-2	**12.84**	-17.19	5.82	-41.19	9.29
Convinced Influener	3.07	9.1	-15.37	**11.05**	-35.96	-4.13
Natural Event	-8.24	-26.76	-48.77	0.37	**680.15**	-38.22
Political Event	-3.24	4.23	**25.2**	2.15	-31.55	-15.62

Table 3: Percentage changes in the distribution of contrarian tweets based on the nature of the trigger.

are not increasing) receives a significant share of the distribution during the Hurricane Ian period.

Generating a time evolution of climate misinformation allows us to identify which categories dominate based on the different types of triggers. Table 3 shows that Natural Events and Political Events shifted the distribution towards topics related to categories 1.7 and 4.1, respectively. This is expected given that 1.7 relates specifically to extreme events and increased during the Hurricane Ian period. Moreover, the increase associated with 4.1 was in response to political events, which is to be expected given that the category involves criticisms of climate policies.

The fluctuations generated by influencers lean significantly towards categories 5.3 (conspiracy theories) and 2.1 (natural cycles/variation), irrespective of whether the influencer holds a contrarian or convinced stance. Notably, when the influencer supports a contrarian viewpoint, there is a discernible increase in the prevalence of conspiracy theories. Conversely, in instances where the influencer is convinced, the distribution leans slightly more toward posts stating that climate change as a natural cycle, accompanied by a concurrent rise in conspiracy theories.

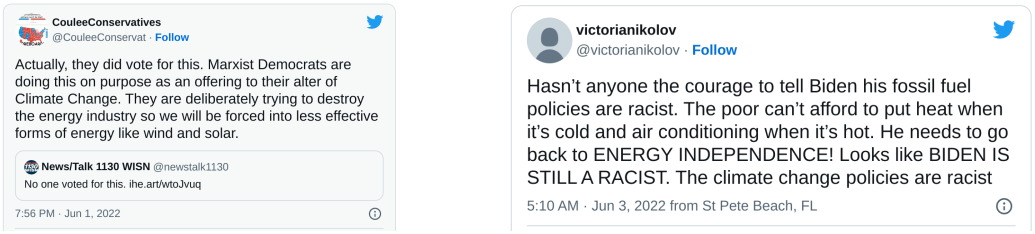


Fig. 7: Tweets against climate change policies.

4 Discussion

Our study shows that a classifier model trained only on misinformation text (e.g., the original CARDS model) struggles at the binary classification task of distinguishing between convinced and contrarian text. We found that adding training data that includes annotations of both convinced and contrarian examples improved performance in binary classification. This addressed a limitation of the original CARDS model, which performed well with known misinformation sources but struggled with general climate text that could have originated from both convinced and unconvinced sources.

Our analysis of climate tweets through a six-month period in 2022 also revealed the dominant categories of climate misinformation on social media relative to other information sources, such as contrarian blogs and CTT websites. While CTTs focused on policy misinformation and contrarian blogs focused on attacking climate science, more than half of climate misinformation tweets focused on either attacking climate actors or conspiracy theories. This underscores the importance of better understanding the impact of climate misinformation in the form of ad hominem attacks and conspiracy theories, as well as exploring the efficacy of interventions that neutralise their negative impact.

We also identified the different types of misinformation peaks on Twitter: external events (political or natural) and influencer posts (contrarian or convinced). External events tended to cause a spike in the total number of climate tweets while influencer-caused peaks tended not to increase overall climate tweets but raised the proportion of misinformation tweets.

There were predictable patterns in the types of arguments in response to different peak types. The clearest signal was in response to natural events, which led to a 680% increase in category 1.7 claims, arguing that weather events weren't linked to climate change. Political events were followed by category 4.1 claims, arguing that climate policy was harmful.

A limitation of this study is that its scope was restricted to climate misinformation on Twitter. It is yet to be seen whether the Augmented CARDS model performs at similar levels on other data sources. Future research could focus on a wider range of information sources, such as other social media platforms, congressional testimonies, public speeches, online video transcripts, and newspaper articles. Such an analysis could also yield which misinformation categories are dominant across these different information sources. Within a single information source, cross-country analysis could

also interrogate different emphases in climate misinformation across different cultures. Similarly, analysis of output from different mainstream media publishers could identify the relative proportion of climate misinformation among different outlets.

Another limitation of the CARDS model is that, to date, it has been trained on English text only. Future research could apply our methodology with training sets of non-English text, to facilitate detection of climate misinformation in other languages across different countries.

5 Conclusion

This study has taken a step closer to the goal of automatically detecting and correcting climate misinformation in real-time. We have shown an improvement in classifying misinformation in climate tweets, with significant reductions in the "false positive problem".

However, there are still numerous hurdles to overcome before the goal of automated debunking is achieved. An effective debunking requires both explanation of the relevant facts and exposing the misleading fallacies employed by the misinformation. Contrarian climate claims can contain a range of different fallacies, so automatic detection of logical fallacies is another necessary task that, used in concert with the CARDS model, could bring us closer to the "holy grail of fact-checking".

Regardless, this research has already provided greater understanding of climate misinformation on social media, identifying four types of misinformation spikes. Knowing the types of arguments that are likely to be posted on social media in response to external events such as climate legislation or natural events can inform interventions that seek to pre-emptively neutralize anticipated misinformation narratives.

References

[1] Ranney, M.A., Clark, D.: Climate change conceptual change: Scientific information can transform attitudes. Topics in cognitive science **8**(1), 49–75 (2016)

[2] Linden, S., Leiserowitz, A., Rosenthal, S., Maibach, E.: Inoculating the public against misinformation about climate change. Global challenges **1**(2), 1600008 (2017)

[3] Geiger, N., Swim, J.K.: Climate of silence: Pluralistic ignorance as a barrier to climate change discussion. Journal of Environmental Psychology **47**, 79–90 (2016)

[4] Cook, J., Lewandowsky, S., Ecker, U.K.: Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. PloS one **12**(5), 0175799 (2017)

[5] Ross Arguedas, A.A., Badrinathan, S., Mont'Alverne, C., Toff, B., Fletcher, R., Nielsen, R.K.: "it'sa battle you are never going to win": Perspectives from journalists in four countries on how digital media platforms undermine trust in news. Journalism Studies **23**(14), 1821–1840 (2022)

[6] Allcott, H., Gentzkow, M.: Social media and fake news in the 2016 election. Journal of economic perspectives **31**(2), 211–236 (2017)

[7] Martens, B., Aguiar, L., Gomez-Herrera, E., Mueller-Langer, F.: The digital transformation of news media and the rise of disinformation and fake news (2018)

[8] Tsfati, Y., Boomgaarden, H.G., Strömbäck, J., Vliegenthart, R., Damstra, A., Lindgren, E.: Causes and consequences of mainstream media dissemination of fake news: literature review and synthesis. Annals of the International Communication Association **44**(2), 157–173 (2020)

[9] Hsu, T., Thompson, S.: Disinformation researchers raise alarms about a.i. chatbots. The New York Times (2023). Accessed 2023-02-08

[10] Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. science **359**(6380), 1146–1151 (2018)

[11] Ecker, U.K., Lewandowsky, S., Tang, D.T.: Explicit warnings reduce but do not eliminate the continued influence of misinformation. Memory & cognition **38**, 1087–1100 (2010)

[12] Hassan, N., Adair, B., Hamilton, J.T., Li, C., Tremayne, M., Yang, J., Yu, C.: The quest to automate fact-checking. In: Proceedings of the 2015 Computation+ Journalism Symposium (2015). Citeseer

[13] Andersen, J., Søe, S.O.: Communicative actions we live by: The problem with fact-checking, tagging or flagging fake news–the case of facebook. European Journal of Communication **35**(2), 126–139 (2020)

[14] Guo, Z., Schlichtkrull, M., Vlachos, A.: A survey on automated fact-checking. Transactions of the Association for Computational Linguistics **10**, 178–206 (2022)

[15] Boussalis, C., Coan, T.G.: Text-mining the signals of climate change doubt. Global Environmental Change **36**, 89–100 (2016)

[16] Farrell, J.: Corporate funding and ideological polarization about climate change. Proceedings of the National Academy of Sciences **113**(1), 92–97 (2016)

[17] Stecula, D.A., Merkley, E.: Framing climate change: Economics, ideology, and uncertainty in american news media content from 1988 to 2014. Frontiers in Communication **4**, 6 (2019)

[18] Alhindi, T., Chakrabarty, T., Musi, E., Muresan, S.: Multitask Instruction-based Prompting for Fallacy Recognition (2023)

[19] Jin, Z., Lalwani, A., Vaidhya, T., Shen, X., Ding, Y., Lyu, Z., Sachan, M., Mihalcea, R., Schölkopf, B.: Logical Fallacy Detection (2022)

[20] Zanartu, F., Cook, J., Wagner, M., Gallego, J.G.: Automatic detection of fallacies in climate change misinformation (2023)

[21] Coan, T., Boussalis, C., Cook, J., Nanko, M.: Computer-assisted detection and classification of misinformation about climate change (2021)

[22] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1423 . https://aclanthology.org/N19-1423

[23] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)

[24] He, P., Liu, X., Gao, J., Chen, W.: Deberta: Decoding-enhanced bert with disentangled attention. ArXiv **abs/2006.03654** (2020)

[25] Chris Bauch, University of Waterloo. https://www.kaggle.com/datasets/edqian/twitter-climate-change-sentiment-dataset

[26] Smith, S.: Biden under pressure to declare climate emergency after manchin torpedoes bill. The Guardian. Accessed 2023-11-10

[27] Luscombe, R.: Hurricane ian: more than 2m without power as florida hit with 'catastrophic' wind and rain. The Guardian. Accessed 2023-06-02

[28] Luscombe, R.: Hurricane ian: 'catastrophic' damage in florida as storm heads to south carolina. The guardian (2022)

[29] Zee, B., Horton, H.: Cop27 day one: Un chief warns world is 'on highway to climate hell' – as it happened. The guardian (2022)

[30] Cook, J.: Deconstructing climate science denial. Research handbook on communicating climate change, 62–78 (2020)

[31] Twitter Climate Change Sentiment Dataset. https://www.kaggle.com/datasets/edqian/twitter-climate-change-sentiment-dataset. Accessed: 2023-05-31

[32] Online Media Monitor - Twitter Analysis. https://www.cen.uni-hamburg.de/en/icdc/data/society/omm-twitter.html. Accessed: 2023-05-31

Appendix A Data

A.1 CARDS

The CARDS dataset encompassed approximately 29,000 claims concerning climate change. Within this dataset, roughly 30% of the claims were determined to be misinformation, using the taxonomy devised by Coan et al. The detailed taxonomy can be found in Table A1. A group of coders well-versed in climate-related matters labelled these claims by analyzing 87,178 paragraphs extracted from communications originating from conservative think tanks (CTTs) and central contrarian blogs [21][30].

Code	Claim label
0	No claim
1	**Global warming is not happening**
1.1	Ice/permafrost/snow cover isn't melting
1.2	We're heading into an ice age/global cooling
1.3	Weather is cold/snowing
1.4	Climate hasn't warmed/changed over the last (few) decade(s)
1.6	Sea level rise is exaggerated/not accelerating
1.7	Extreme weather isn't increasing/has happened before/isn't linked to climate change
2	**Human greenhouse gases are not causing climate change**
2.1	It's natural cycles/variation
2.3	There's no evidence for greenhouse effect/carbon dioxide driving climate change
3	**Climate impacts/global warming is beneficial/not bad**
3.1	Climate sensitivity is low/negative feedbacks reduce warming
3.2	Species/plants/reefs aren't showing climate impacts/are benefiting from climate change
3.3	CO2 is beneficial/not a pollutant
4	**Climate solutions won't work**
4.1	Climate policies (mitigation or adaptation) are harmful
4.2	Climate policies are ineffective/flawed
4.4	Clean energy technology/biofuels won't work
4.5	People need energy (e.g. from fossil fuels/nuclear)
5	**Climate movement/science is unreliable**
5.1	Climate-related science is unreliable/uncertain/unsound (data, methods & models)
5.2	Climate movement is unreliable/alarmist/corrupt

Table A1: Taxonomy employed to categorize the misinformation claims within the CARDS dataset. It consists of two hierarchical levels, encompassing a total of 18 categories.

A.2 Waterloo

The dataset compiled by the University of Waterloo consists of labelled tweets pertaining to climate change, covering the time period from April 27, 2015, to February 21, 2018. In total, 43,943 tweets were annotated. Each tweet underwent individual labelling by three reviewers, and only those tweets that received unanimous agreement from all reviewers were included [31].

A.3 Hamburg

The Online Media Monitor (OMM) from the University of Hamburg contributed with a dataset of 5,236,660 unlabeled tweets gathered from June 21, 2022, to December 8, 2022. The data was filtered from the platform based on keywords or phrases that included: #climatechange, climate change, "global warming," climate crisis, or climate emergency [32].

Appendix B Anomalies: Word Analysis

Token	Log Fold Change	P value
heatwave	1.474412	1.13E-18
declare	1.432399	7.48E-16
fires	1.006194	4.70E-17
hot	0.958189	2.69E-45
biden	0.906806	4.52E-115
heat	0.897032	8.50E-112
@potus	0.854548	1.18E-50
summer	0.784148	1.27E-22
climate emergency	0.735294	1.44E-113
uk	0.574371	2.18E-17

(a)

Token	Log Fold Change	P value
ian	2.713324	6.19E-103
hurricane	2.584126	0.00E+00
florida	2.161538	1.11E-88
@danrather	2.107857	8.97E-23
storm	1.935989	3.65E-25
climate change	-0.089166	1.43E-233
global warming	-0.315675	2.28E-88
us	-0.355969	9.99E-28
global	-0.36468	1.96E-181
climate crisis	-0.405621	2.24E-45

(b)

Token	Log Fold Change	P value
egypt	1.827352	4.13E-20
cop27	1.707518	5.53E-53
conference	1.186116	1.62E-13
nations	0.955196	9.09E-11
leaders	0.871997	8.68E-11
countries	0.720918	1.46E-26
climate crisis	-0.1728	2.26E-11
people	-0.270252	3.04E-33
climate change	-0.29039	0.00E+00

(c)

Table B2: Top 10 words ranked by their Log Fold change during the evaluated period, in comparison to the overall timeline. (a). July 18 and July 21, 2022. (b). September 29, 2022. (c). November 7 and November 8, 2022.