communications earth & environment

Article

https://doi.org/10.1038/s43247-024-01573-7

Hierarchical machine learning models can identify stimuli of climate change misinformation on social media

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

Misinformation about climate change poses a substantial 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. In this study, we address this gap by developing a two-step hierarchical model. The Augmented Computer Assisted Recognition of Denial and Skepticism (CARDS) model is specifically designed for categorising climate claims on Twitter. 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 contrarian climate claims on Twitter involve attacks on climate actors. Spikes in climate contrarianism coincide with one of four stimuli: political events, natural events, contrarian influencers, or convinced influencers. Implications for automated responses to climate misinformation are discussed.

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 influential 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].

Climate change has long been a key target of misinformation on social media platforms. Analysis of tweets around COP climate summits found that contrarian tweets and polarization have substantially grown since 2009[10]. Geographically, hoax-themed tweets that question the reality of climate change are more prominent in conservative U.S. states, relative to liberal states or tweets from the UK, Canada, or Australia[11]. An analysis of tweets about a 2013 Intergovernmental Panel on Climate Change (IPCC) report found that Twitter users unsupportive of climate science or policies were most active in sending tweets about the IPCC, with uncivil tweets being the most viral contrarian tweets[12]. Similarly, tweets that are skeptical about climate change have been found to show tones of incivility[13], and climate change deniers on Twitter use aggressive language and negative sentiment[14].

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

[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]Monash Data Futures Institute, Monash University, Clayton, 3800 Victoria, Australia. [5]Exeter Q-Step Centre, University of Exeter, Exeter, UK. [6]Melbourne Centre for Behaviour Change, University of Melbourne, Parkville, Victoria, Australia. ✉e-mail: jocook@unimelb.edu.au; Yuanfang.Li@monash.edu

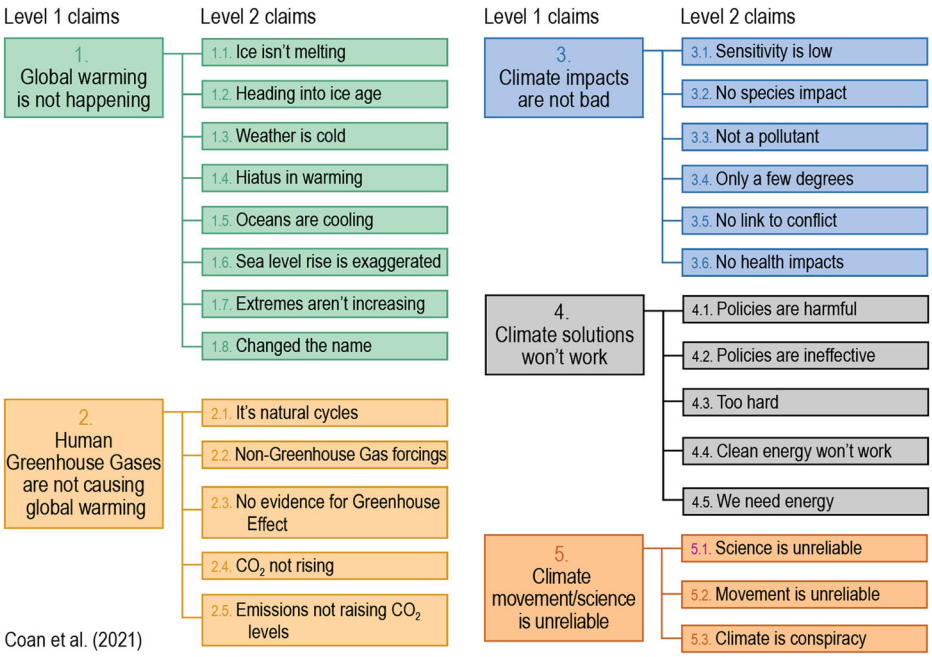

Fig. 1 | CARDS taxonomy of contrarian climate claims. This taxonomy provides a comprehensive overview of the frequently employed main claim and its corresponding subarguments utilized to bolster contrarian perspectives on climate change[26].

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[15]. Further, once misinformation has taken hold, it is difficult to dislodge—a phenomenon known as the continued influence effect[16]. 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"[17]. There have been efforts to automatically detect and fact-check misinformation across various domains[18,19]. 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[20], link corporate funding to polarizing climate text[21], and identify climate framings in newspaper articles[22]. There have also been efforts to detect logical fallacies in climate misinformation as well as across general topics[23–25].

The CARDS (Computer Assisted Recognition of Denial & Skepticism) model used supervised machine learning to detect and categorize contrarian claims about climate change[26]. The model has been shown to be effective in categorizing a wide range of contrarian claims about climate change. It was trained to classify contrarian claims based on the taxonomy illustrated in Fig. 1.

However, the CARDS model was only trained using text from contrarian blogs and conservative think-tank websites—prolific sources of climate skepticism—and its performance in classifying content from other datasets (e.g., from social media platforms) has yet to be assessed. This study evaluates and enhances the CARDS model's performance in classifying contrarian climate claims 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 contrarian peaks. We trained a model with a superior performance to classify the taxonomy in Fig. 1. Nevertheless, the content of contrarian claims on Twitter primarily consists 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) the climate movement/science is 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).

Methods

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

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. In Fig. 2, the general model architecture is illustrated. It consists of an initial layer trained on a binary detection task to differentiate between convinced and contrarian tweets, coupled with an additional layer trained on a multilabel task to classify the taxonomy illustrated in Fig. 1.

Both classifiers incorporate the DeBERTa language model, structured based on the auto-encoding transformer architecture introduced in BERT[27]. The innovation includes disentangled attention and a more extensive pre-training process[28,29]. Specifically, we utilized the large version of DeBERTa, consisting of 24 transformer blocks with a hidden size of 1024 and 16 attention heads. In addition, an extra dense layer was employed for the classification task, bringing the total number of parameters to ~355 million.

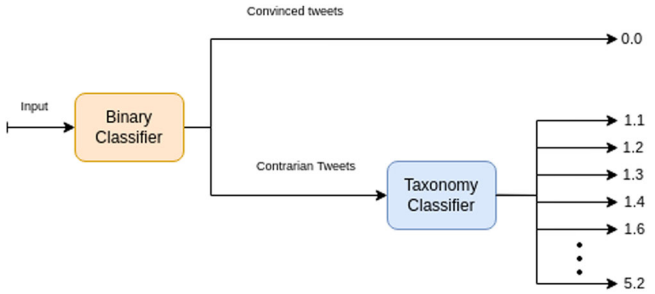

Fig. 2 | Model architecture. The two-step hierarchical model consists of a binary classifier that determines the main stance of a climate-related claim, followed by a taxonomy classifier that details the types of arguments used if the claim is identified as contrarian.

Table 1 | Assessment of F1-scores achieved, comparing the original CARDS model with the Augmented CARDS model

Task	Datasets	Models		Support
		CARDS	Augmented CARDS	
	CARDS	**90.3**	89.1	2931
	Twitter Climate Change	68.1	**88.2**	4395
Binary Detection	*Expert Annotated Climate Tweets*	67.8	**81.1**	2711
	CARDS	**72.7**	73.100	2904
Taxonomy classification	*Expert Annotated Climate Tweets*	43.7	**53.6**	2607

Bold values indicate which model showed higher performance.

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. In addition, 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[29], fine-tuning is necessary to capture the linguistic features specific to the platform.

Training details

To enhance the model's performance, we incorporated the Climate Change Twitter Dataset labelled by the University of Waterloo, featuring a 90/10 ratio of verified and misleading tweets (https://www.kaggle.com/datasets/edqian/twitter-climate-change-sentiment-dataset), 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[26] taxonomy. This dataset, denoted as Expert Annotated Climate Tweets in Table 1, was composed of 2607 tweets related to climate change, sampled from the platform in the second half of 2022.

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[30]. 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.

Results

Assessing the Augmented CARDS model

Table 1 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[26]. 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 ref. 26. In addition, 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.

Our results indicate that Twitter is a challenging task due to the remarkable disparities in language and writing style observed between the original sources and the platform. On the other hand, the Augmented CARDS model achieves a notable improvement in the F1-Score for both the Twitter Climate Change and the Expert Annotated Climate Tweets datasets for both tasks.

The technical contributions 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 1, 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 2.

The most prominent contrarian categories were 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 2 shows that our model exhibits the most substantial improvements within 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).

Applying Augmented CARDS to 2022 climate tweets

We applied the Augmented Cards model to over 5 million climate-related tweets in a 6-month period in 2022, providing insight into the proliferation of climate-contrarian claims on Twitter. This novel investigation, to the best of our knowledge, uncovered the triggers that related to 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 related to "climate change"[30]. Figure 3 illustrates the daily frequency of climate change-related tweets, revealing notable fluctuations, including a substantial peak in late July. On average, the dataset contains 27,464 tweets per day about climate change. However, during the peaks in late July and mid-November, the number of tweets surged to 65,196 and 43,647, respectively.

Table 2 | F1-scores per category obtained from the Augmented CARDS model on the Expert Annotated Climate Tweets dataset

Category	CARDS	Augmented CARDS	Support
0.0	70.9	**81.5**	1049
1.1	60.5	**70.4**	28
1.2	40	**44.4**	20
1.3	37	**48.6**	61
1.4	62.1	**65.6**	27
1.6	56.7	**59.7**	41
1.7	46.4	**52**	89
2.1	68.1	**69.4**	154
2.3	**36.7**	25	22
3.1	**38.5**	34.8	8
3.2	61	**74.6**	31
3.3	54.2	**65.4**	23
4.1	38.5	**49.4**	103
4.2	**37.6**	28.6	61
4.4	30.8	**54.5**	46
4.5	19.7	**39.4**	50
5.1	32.8	**38.2**	96
5.2	38.6	**53.5**	498
5.3	–	**62.9**	200
Macro Average	43.69	**53.57**	2407

Bold values indicate which model showed higher performance.

To further understand the observed peaks, we performed statistical analyses to identify words with major variations and establish correlations between these shifts and remarkable 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.

Between July 18 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[31].

The second-largest peak of overall climate tweets was associated with COP27, as indicated by the changes in word distribution. This event was associated with a doubling of the number of tweets between September 7th and 9th, 2022. The third highest peak of overall climate tweets in our dataset coincided with Hurricane Ian[32]. Tweets relating to the Hurricane became the major topic of discussion related to 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, Fig. 4 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 6-month period.

Overall, we identified four distinct categories of events that coincided with an upsurge in the publication of contrarian tweets, as outlined in Table 3. The triggering events can be broadly classified into three primary groups: Natural Events, Political Events, and Influencer Posts.

Natural Events, such as Hurricane IAN, and Political Events like COP27, were external occurrences originating outside the platform[33,34]. They coincided with 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 major peaks in the percentage of contrarian tweets. 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. The percentage of contrarian claims reached a peak of 24.7% around the time of President Biden's declaration.

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

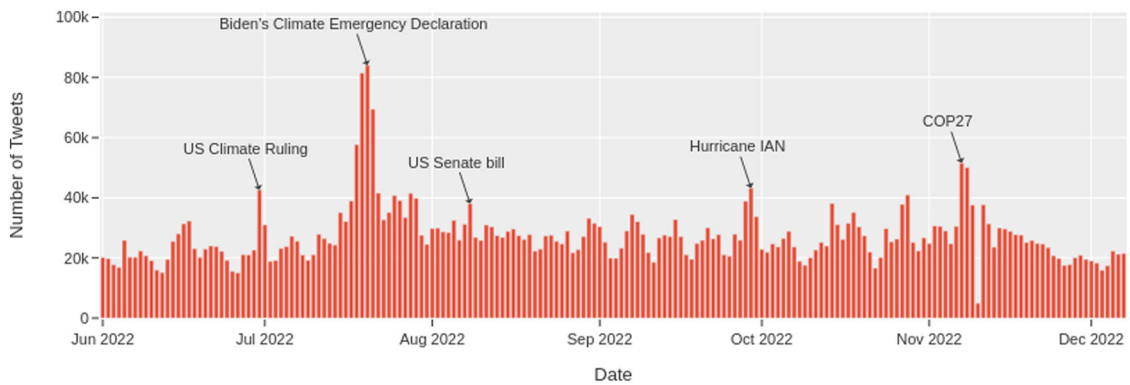

Fig. 3 | Climate change Tweeter trends in 2022. Number of tweets related to climate change published during the second semester of 2022.

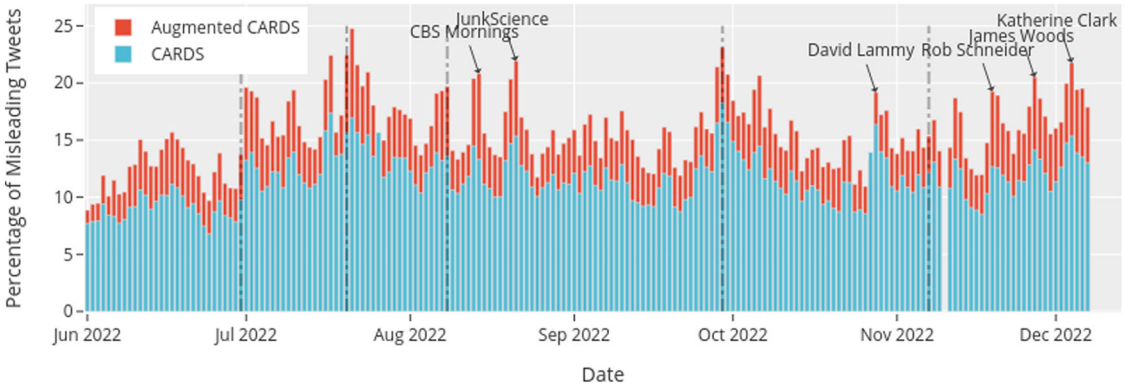

Fig. 4 | Detection of contrarian tweets in 2022. Percentage of contrarian tweets detected by the CARDS and Augmented CARDS models.

Table 3 | Categorization of events that induced spikes in climate contrarianism on Twitter

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

reached similarly high levels. Discussions were centred around the impact of climate change on extreme weather conditions. For instance, FoxNews tweeted one of their articles, titled "Democrats blaming climate change for Hurricane Ian at odds with science, experts say".

The COP27 event stands in contrast to the Biden emergency declaration and Hurricane Ian. While the latter events led to important increases in both the volume of tweets and the proportion of contrarian claims, COP27 resulted in only a 1% rise in the proportion of contrarian claims among climate-related tweets.

Moreover, aside from the increase in tweet volume associated with Political and Natural Events, other cases showed a rise linked to tweets published by influential accounts. In addition, contrarian claims increased in response to influencers regardless of whether they expressed a convinced or contrarian view. These influencers, including politicians, comedians, film directors, and media figures, all share the characteristic of being public figures with a substantial number of followers. It's important to note that our categorization of influencers is based on the positions adopted in their 2022 publications, not necessarily their current personal viewpoints.

Our final analysis is the categorisation of contrarian tweets by the typology of ref. 26 as inferred by the Augmented CARDS model. Figure 5 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 contrarianism 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 ~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 are not increasing) receives an important share of the distribution during the Hurricane Ian period.

Generating a time evolution of contrarian claims allows us to identify which categories dominate based on the different types of triggers. Table 4 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, increases in category 4.1 were associated with political events, which is to be expected given that the category involves criticisms of climate policies.

The fluctuations generated by influencers lean substantially 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 supported a contrarian viewpoint, there was a discernible increase in the prevalence of conspiracy theories. Conversely, in instances where the influencer expressed a convinced stance, the distribution shifted slightly more toward posts asserting that climate change is a natural cycle, accompanied by a concurrent rise in conspiracy theories.

Finally, a considerable number of accounts exhibited an unusually high volume of contrarian tweets across all primary categories. On average, each user shared between 1 and 2 contrarian tweets. Notably, category 5.3 had the highest average of 1.9 tweets per user, closely followed by category 5.2 with an average of 1.8. The remaining three categories had averages of 1.57 contrarian tweets or fewer per user. This suggests that discussions revolving around conspiracy theories were generating a meaningful amount of content from a relatively closed pool of users. However, users with hundreds of published tweets could be found across all principal categories. For example, the user with the maximum number of posts related to conspiracies (5.3) reached 921 tweets just in this category. Similar cases could be found in other categories as well: 5.2 (881), 4.1 (627), 2.1 (424), and 1.7 (108). Some of these outlier accounts were evidently automated. For instance, one user posted 1977 tweets within our data period, with only 206 of them featuring unique textual content, while the rest repetitively rephrased the same idea with minor grammatical changes. Other accounts were evidently managed by individuals engaging in discussions and expressing opinions. Overall, the proportion of unique content in contrarian detections is ~93.3%, with categories 5.3 and 5.2 containing around 91.3% and 94.1% unique content, respectively. In summary, the existence of accounts managed by automated systems and users generating a substantial volume of contrarian tweets is indisputable, with their content consistently reflecting the trends identified in our taxonomy analysis. In addition, we observed that ~6% of the analyzed tweets could be categorized as spam. However, content generated by modern AI models could potentially circumvent our current analytical approach and exacerbate this issue.

Discussion

Our study showed that a classifier trained only on climate contrarian 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 from 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.

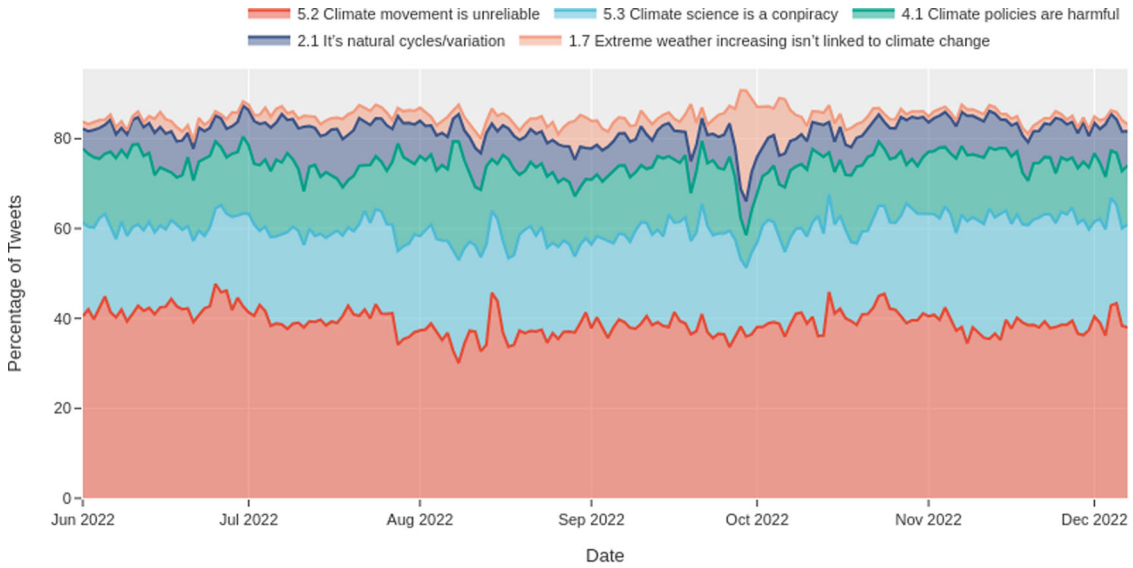

Fig. 5 | Distribution of the Top 5 contrarian arguments. Breakdown of the five most relevant categories detected by the Augmented CARDS model on Twitter in 2022. The percentages of these categories are relative to the total proportion of contrarian claims identified by the binary model.

Table 4 | Preferred arguments based on the type of source

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 Influencer	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

Percent changes in the distribution of arguments used by the four agents analysed in contrast with the distribution of the dataset. Bold values indicate the most frequently used category as an argument by each trigger.

Our analysis of climate tweets through a 6-month period in 2022 also revealed the dominant categories of contrarian claims and climate misinformation on social media relative to other information sources, such as contrarian blogs and CTT websites. While CTTs focused on policy claims and contrarian blogs focused on attacking climate science, more than half of the relevant climate tweets focused on either attacking climate actors or conspiracy theories. This result is consistent with other research finding that attacks on climate supporters are a dominant theme among tweets by climate contrarians, as are climate-themed conspiracy theories[35]. These findings are also consistent with recent scholarship on the personalisation of politics on social media[36,37] and the observation that personal attacks tend to drive engagement on platforms such as Facebook and Twitter/X[38], suggesting that the nature of misinformation—i.e., personalised versus non-personalised—is at least particularly dependent on the platform of interest. Overall, the dominance of personal attacks in climate misinformation on social media 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, associated with external events (political or natural) or influencer posts (contrarian or convinced). External events coincided with a spike in the total number of climate tweets. while influencer-associated peaks were associated with an increase in the proportion of misinformation tweets but not an increase in overall climate 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 showed a 680% increase in category 1.7 claims making the argument that weather events weren't linked to climate change. Political events coincided with 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.

Future research may also apply supervised learning models to a broader period of time to draw out long-term trends on Twitter and other platforms. For example, there has been much discussion surrounding the recent changes to Twitter's ownership[39] and the impact this has had on climate misinformation[40]. Our data does not span a large enough time period to make any concrete conclusions about these recent changes.

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.

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 considerable reductions in the "false positive problem" associated with training on contrarian actors alone.

These findings have practical implications. Adopting our model could help Twitter/X to augment and enhance ongoing manual fact-checking procedures by offering a computer-assisted procedure for finding the tweets most likely to contain climate misinformation. This adoption could make finding and responding to climate-related misinformation more efficient and help Twitter/X enforce policies to reduce false or misleading claims on the platform. Yet environmental groups have shown that Twitter/X ranks dead last among major social media platforms in its policies and procedures for responding to climate misinformation and there is little evidence that X will improve these procedures in the near term[41]. Alternatively, our model could provide the basis for an API that Twitter/X users could employ to assess climate-related claims they are seeing in their feeds. Overall, the potential practical applications of our model underscores the need for continued academic work to monitor misinformation on Twitter/X and raises important questions on the data needed to hold social media platforms accountable for the spread of false claims.

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"[17].

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 preemptively neutralize anticipated misinformation narratives.

Data availability

The experimental data used to train the models and perform the analysis in this study are available on Figshare, accessible through the following identifier: https://doi.org/10.6084/m9.figshare.25465036.

Code availability

The machine learning models used for the analysis conducted in this study can be found in the following repositories for reproducibility of our results: 1. https://huggingface.co/crarojasca/BinaryAugmentedCARDS. 2. https://huggingface.co/crarojasca/TaxonomyAugmentedCARDS.

Received: 8 December 2023; Accepted: 18 July 2024;
Published online: 16 August 2024

References

1. Ranney, M. A. & Clark, D. Climate change conceptual change: scientific information can transform attitudes. *Top. Cogn. Sci.* **8**, 49–75 (2016).
2. Van der Linden, S., Leiserowitz, A., Rosenthal, S. & Maibach, E. Inoculating the public against misinformation about climate change. *Glob. Chall.* **1**, 1600008 (2017).
3. Geiger, N. & Swim, J. K. Climate of silence: Pluralistic ignorance as a barrier to climate change discussion. *J. Environ. Psychol.* **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**, e0175799 (2017).
5. Ross Arguedas, A. A. et al. "it's a battle you are never going to win": perspectives from journalists in four countries on how digital media platforms undermine trust in news. *Journal. Stud.* **23**, 1821–1840 (2022).
6. Allcott, H. & Gentzkow, M. Social media and fake news in the 2016 election. *J. Econ. Perspect.* **31**, 211–236 (2017).
7. Martens, B., Aguiar, L., Gomez-Herrera, E. & Mueller-Langer, F. in *The Digital Transformation of News Media and the Rise of Disinformation and Fake News* (European Commission, 2018).
8. Tsfati, Y. et al. Causes and consequences of mainstream media dissemination of fake news: literature review and synthesis. *Ann. Int. Commun. Assoc.* **44**, 157–173 (2020).
9. Hsu, T. & Thompson, S. Disinformation researchers raise alarms about A.I. chatbots. *The New York Times* (2023).
10. Falkenberg, M. et al. Growing polarization around climate change on social media. *Nat. Clim. Change* **12**, 1114–1121 (2022).
11. Jang, S. M. & Hart, P. S. Polarized frames on "climate change" and "global warming" across countries and states: evidence from Twitter big data. *Glob. Environ. Change* **32**, 11–17 (2015).
12. Pearce, W., Holmberg, K., Hellsten, I. & Nerlich, B. Climate change on Twitter: topics, communities and conversations about the 2013 IPCC Working Group 1 report. *PLoS ONE* **9**, e94785 (2014).
13. Anderson, A. A. & Huntington, H. E. Social media, science, and attack discourse: how Twitter discussions of climate change use sarcasm and incivility. *Sci. Commun.* **39**, 598–620 (2017).
14. Effrosynidis, D., Sylaios, G. & Arampatzis, A. Exploring climate change on Twitter using seven aspects: Stance, sentiment, aggressiveness, temperature, gender, topics, and disasters. *PLoS ONE* **17**, e0274213 (2022).
15. Vosoughi, S., Roy, D. & Aral, S. The spread of true and false news online. *Science* **359**, 1146–1151 (2018).
16. Ecker, U. K., Lewandowsky, S. & Tang, D. T. Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Mem. Cognition* **38**, 1087–1100 (2010).
17. Hassan, N. et al. The quest to automate fact-checking. *In Proceedings of the 2015 Computation + journalism symposium* (Citeseer, 2015).
18. Andersen, J. & Søe, S. O. Communicative actions we live by: the problem with fact-checking, tagging or flagging fake news–the case of Facebook. *Eur. J. Commun.* **35**, 126–139 (2020).
19. Guo, Z., Schlichtkrull, M. & Vlachos, A. A survey on automated fact-checking. *Trans. Assoc. Comput. Linguist.* **10**, 178–206 (2022).
20. Boussalis, C. & Coan, T. G. Text-mining the signals of climate change doubt. *Glob. Environ. Change* **36**, 89–100 (2016).
21. Farrell, J. Corporate funding and ideological polarization about climate change. *Proc. Natl Acad. Sci. USA* **113**, 92–97 (2016).
22. Stecula, D. A. & Merkley, E. Framing climate change: economics, ideology, and uncertainty in American news media content from 1988 to 2014. *Front. Commun.* **4**, 6 (2019).
23. Alhindi, T., Chakrabarty, T., Musi, E. & Muresan, S. Multitask instruction-based prompting for fallacy recognition. In *Proc. 2022 Conference on Empirical Methods in Natural Language Processing*, 8172–8187 (2023).
24. Jin, Z. et al. Logical fallacy detection. *Findings of the Association for Computational Linguistics: EMNLP 2022*, 7180–7198 (2022).
25. Zanartu, F., Cook, J., Wagner, M. & Gallego, J. G. Automatic detection of fallacies in climate change misinformation (2023).
26. Coan, T., Boussalis, C., Cook, J. & Nanko, M. Computer-assisted detection and classification of misinformation about climate change. https://doi.org/10.31235/osf.io/crxfm (2021).
27. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *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)* (eds. Burstein, J., Doran, C. & Solorio, T.) 4171–4186 (Association for Computational Linguistics, 2019).
28. Liu, Y. et al. Roberta: a robustly optimized BERT pretraining approach. https://doi.org/10.48550/arXiv.1907.11692 (2019)
29. He, P., Liu, X., Gao, J. & Chen, W. Deberta: decoding-enhanced bert with disentangled attention. https://api.semanticscholar.org/CorpusID:219531210 (2020).
30. Brüggemann, M. & Sadikni, R. Online media monitor on climate change (omm): analysis of global tweets and online media coverage. https://ogy.de/OMM (2023).

31. Smith, S. Biden under pressure to declare climate emergency after Manchin torpedoes bill. *The Guardian*. https://www.theguardian.com/environment/2022/jul/19/joe-biden-climate-emergency (2022).
32. Luscombe, R. Hurricane Ian: more than 2m without power as Florida hit with 'catastrophic' wind and rain. *The Guardian*. https://www.theguardian.com/us-news/2022/sep/28/hurricane-ian-millions-florida-path-deadly-cyclone (2022).
33. Luscombe, R. Hurricane Ian: 'catastrophic' damage in Florida as storm heads to South Carolina. *The Guardian*. https://www.theguardian.com/us-news/2022/sep/29/florida-rescue-crews-search-residents-trapped-hurricane-ian-floods (2022).
34. Van der Zee, B. & Horton, H. Cop27 day one: Un chief warns world is 'on highway to climate hell'—as it happened. *The Guardian* (2022).
35. Xia, Y., Chen, T. H. Y. & Kivelä, M. Spread of tweets in climate discussions: a case study of the 2019 Nobel peace prize announcement. *Nord. J. Media Stud.* **3**, 96–117 (2021).
36. Bennett, W. L. The personalization of politics: political identity, social media, and changing patterns of participation. *ANNALS Am. Acad. Political Soc. Sci.* **644**, 20–39 (2012).
37. Kissas, A. Populist everyday politics in the (mediatized) age of social media: the case of Instagram celebrity advocacy. *N. Media Soc.* **0**, 14614448221092006 (2022).
38. Rathje, S., Bavel, J. J. V. & van der Linden, S. Out-group animosity drives engagement on social media. *Proc. Natl Acad. Sci. USA* **118**, e2024292118 (2021).
39. Barrie, C. Did the Musk takeover boost contentious actors on Twitter? *Harvard Kennedy School Misinformation Review* (2023).
40. Institute for Strategic Dialogue Deny (ISD). *Deceive, Delay: Exposing New Trends in Climate Mis- and Disinformation at COP27. Vol. 2.* https://www.isdglobal.org/wp-content/uploads/2023/01/Deny-Deceive-Delay-Vol.-2.pdf (2023).
41. Climate Action Against Disinformation (CAAD). Climate of misinformation: ranking big tech. *Climate Action Against Disinformation and Friends of the Earth and Greenpeace.* https://caad.info/wp-content/uploads/2023/09/Climate-of-Misinformation.pdf (2023).

Acknowledgements

We gratefully acknowledge the large dataset provided by the Online Media Monitor (OMM) from the University of Hamburg and the Monash Data Futures Institute for their support.

Author contributions

The paper's authorship contributions are as follows: John Cook and Yuan-Fang Li led the study's conceptualization and design; data collection was carried out by Frank Algra-Maschio, Cristian Rojas and Travis Coan; machine learning algorithm development and data analysis were led by Cristian Rojas and Yuan-Fang Li; Frank Algra-Maschio and Mark Andrejevic explored social correlations and interpreted results. All authors contributed to drafting the manuscript, critically reviewed the findings, and approved the final version.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43247-024-01573-7.

Correspondence and requests for materials should be addressed to John Cook or Yuan-Fang Li.

Peer review information *Communications Earth & Environment* thanks Sonny Rosenthal, Giuseppe Veltri and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: Adam Switzer and Carolina Ortiz Guerrero. A peer review file is available.