Analyzing the Dynamics of Climate Change Discourse on Twitter: A New Annotated Corpus and Multi-Aspect Classification

Shuvam Shiwakoti[1,†,*], Surendrabikram Thapa[2,†], Kritesh Rauniyar[1], Akshyat Shah[1], Aashish Bhandari[3], and Usman Naseem[4]
[1]Delhi Technological University, India [2]Department of Computer Science, Virginia Tech, USA
[3]RMIT University, Australia [4] Macquarie University, Australia
*ssiwakoti12@gmail.com

Abstract

The discourse surrounding climate change on social media platforms has emerged as a significant avenue for understanding public sentiments, perspectives, and engagement with this critical global issue. The unavailability of publicly available datasets, coupled with ignoring the multi-aspect analysis of climate discourse on social media platforms, has underscored the necessity for further advancement in this area. To address this gap, in this paper, we present an extensive exploration of the intricate realm of climate change discourse on Twitter, leveraging a meticulously annotated *ClimaConvo* dataset comprising 15,309 tweets. Our annotations encompass a rich spectrum, including aspects like relevance, stance, hate speech, the direction of hate, and humor, offering a nuanced understanding of the discourse dynamics. We address the challenges inherent in dissecting online climate discussions and detail our comprehensive annotation methodology. In addition to annotations, we conduct benchmarking assessments across various algorithms for six tasks: relevance detection, stance detection, hate speech identification, direction and target, and humor analysis. This assessment enhances our grasp of sentiment fluctuations and linguistic subtleties within the discourse. Our analysis extends to exploratory data examination, unveiling tweet distribution patterns, stance prevalence, and hate speech trends. Employing sophisticated topic modeling techniques uncovers underlying thematic clusters, providing insights into the diverse narrative threads woven within the discourse. The findings present a valuable resource for researchers, policymakers, and communicators seeking to navigate the intricacies of climate change discussions. The dataset and resources for this paper are available at https://github.com/shucoll/ClimaConvo.

Keywords: climate change, multi-aspect annotations, language resources

1. Introduction

Climate change is a formidable challenge confronting not only our species but all life forms on Earth. The severe and long-term effects of climate change have drawn the attention of the public and leaders worldwide. This led to the first global climate accord in 2015, Paris Agreement (Dimitrov, 2016), which legally bound the 196 participating states to commit measures to limit the global temperature increase to 1.5 °C above the pre-industrial era. The 1.5 °C limit has now become a statement and is used by scientists, leaders, and the public alike to advocate their concerns on climate change. This also led to numerous activists like the FridayForFuture (FFF) movement (Wallis and Loy, 2021) started by Greta Thunberg to put moral pressure on policymakers to adhere to their promise of contributing to the 1.5 °C goal. Young people worldwide took to the streets protesting to draw their leaders' attention toward climate change. While movements like this on climate change progressed, online social media also witnessed a spike in a discussion of climate change (Segerberg and Bennett, 2011).

Social media, particularly Twitter, has evolved into a diverse landscape for expressing opinions and emotions regarding climate change (Fownes et al., 2018). This digital space has become a platform where individuals champion strategies to mitigate climate change's impact while others vehemently deny its existence and effects (Ross and Rivers, 2019). This multifaceted discourse presents an imperative that necessitates comprehensive examination. Thus, it provides a prospect for exploring the socio-linguistic intricacies inherent in a discourse concerning climate change within the informal social media domain (Chen et al., 2019).

Within the discourse landscape concerning climate change on the Twitter platform, a multitude of intricate elements converge, each necessitating a methodical and thorough exploration (Kirilenko et al., 2015). These elements encompass various dimensions, including stance, hate speech, the direction of hate speech, targets of hate speech, and even humor. All aspects hold equal significance in comprehending the nuances of the climate change discourse. For example, evaluating stance within discourse serves as a crucial analytical lens as it provides insights into the range of viewpoints individuals express toward climate change. It identifies supporters, skeptics, and deniers and gauges the intensity and commitment behind each perspec-

[†]These authors contributed equally to this work. The names are listed in alphabetical order.

LREC-COLING 2024, pages 984–994

tive. Understanding different stances within the discourse is essential for gauging public sentiment, capturing shifts in opinions, and identifying trends that might shape future discussions and policies. Furthermore, assessing stance can be a potent tool for raising awareness and fostering well-informed public engagement. With the recent addition of the feature to add references to tweets on Twitter, understanding individuals' stances on climate change becomes pivotal in identifying tweets that would benefit from supplementary references. This proactive approach ensures that tweets align with informed perspectives, contributing to a more substantiated discourse within the evolving landscape of climate change discussions.

Similarly, exploring the realm of hate speech within climate change discourse is equally important. Hate speech within this specific context can intensify the polarization of opinions, impeding productive discussions and the collaborative process of consensus-building. Therefore, a crucial step toward comprehending the multifaceted nature of climate change discourse involves understanding the underlying drivers of hate speech and addressing its divisive elements. Furthermore, an essential aspect of this exploration involves investigating the direction of hate speech, particularly identifying its targets within the climate change discourse. This dimension sheds light on the specific groups or entities that become the focus of this hostility. This knowledge holds significant value for devising interventions tailored to the nuances of climate-related hate speech, advocating for those adversely affected, and fostering an atmosphere of discourse that is both respectful and inclusive. By addressing hate speech at its source and concentrating efforts on affected areas, the discourse environment can be effectively nurtured to encourage healthier and more constructive exchanges of ideas.

Additionally, humor, often embedded in the language, can serve as a mechanism to relay complex issues, facilitate engagement, and convey perspectives that might otherwise be challenging to express. Examining the humor aspect within climate change discourse uncovers nuances of communication strategies, enriching our comprehension by introducing an additional layer of insight into the diverse emotions and viewpoints populating the social media landscape.

To comprehensively delve into the intricacies of aspects such as stance, hate speech, humor, and relevance within the climate change discourse, applying advanced language models (LMs) is a potent avenue (Min et al., 2021). However, to effectively harness the potential of language models for these tasks, the availability of well-annotated datasets becomes a foundational prerequisite. Nonetheless, a significant scarcity of comprehensive works and datasets addressing climate change discourse remains. This scarcity underscores the unique and relatively underexplored nature of this domain within the broader field of NLP. To address this pressing gap, we present **ClimaConvo**, a meticulously curated dataset comprising 15,309 English tweets relevant to climate change and activist movements such as Fridays For Future (FFF). This dataset is enriched with annotations for six distinct tasks: **Task A** involves "relevance assessment", **Task B** pertains to "stance classification" (support, denial, neutral), **Task C** focuses on "hate speech identification", **Task D** centers on the "determination of the direction of hate speech" (directed vs. undirected), **Task E** deals with the "identification of hate speech targets" (individuals, organizations, communities), and **Task F** encompasses "humor detection". Our main contributions are:

- We present a comprehensive and multi-aspect large-scale annotated dataset encompassing 15,309 tweets focused on climate change and associated activist movements.
- Our analysis delves into discourse dynamics via topic modeling and sentiment analysis, adding a layer of nuanced understanding.
- We establish benchmark performances using state-of-the-art algorithms across diverse tasks. These benchmarks identify areas for potential improvement and elicit scholarly attention within this domain.

Our contribution, encompassing a wide array of tasks, holds substantial appeal for the field of Natural Language Processing (NLP) and serves as a valuable response to the pressing need to address climate change-related issues. Moreover, the introduction of an analysis of discourse concerning climate change aligns effectively with several United Nations Sustainable Development Goals (SDGs), including SDG13: Climate Action. This analysis represents a step forward in understanding and addressing the challenges of climate change, aligning with the global objectives outlined in the SDGs.

2. Related Works

2.1. Climate Change and NLP

Advancements in Natural Language Processing (NLP) owe much to pre-trained language models (LMs). To understand the discourse related to climate, Webersinke et al. (2021) introduced *ClimateBERT*, a domain-specific LM trained on a dataset of 2,046,523 paragraphs sourced from climate-related news, abstracts, and reports. This targeted approach enhances accuracy in addressing climate-related challenges.

Works	Data Source	Size	Context	Aspects/ Tasks
Stammbach et al. (2022)	Twitter	3,000	Climate Discourse	Stance
Gautam et al. (2020)	Twitter	9,973	MeToo movement	Stance, Relevance, Hate speech, Dialogue acts, and Sarcasm
Salawu et al. (2021)	Twitter	62,587	Cyberbullying	Insult, Bullying, Profanity, Sarcasm, Threat, Exclusion, Porn, and Spam
Mollas et al. (2022)	YouTube and Reddit	998	General Discourse	Violence, Directed vs Generalised, Gender, Race, National Origin, Disability, Religion, and Sexual Orientation
Ousidhoum et al. (2019)	Twitter	13,014	General Discourse	Directness, Hostility, Target, Group, and Annotator
Zampieri et al. (2019)	Twitter	14,100	Social Media	Offensive, Targeted: Individual, Group, and Other
ClimaConvo (Ours)	**Twitter**	**15,309**	**Climate Discourse**	**Relevance, Stance, Hate speech, Direction of Hate Speech, Targets of Hate Speech, Humor**

Table 1: Summary of different related works and datasets

Moreover, as climate change is predominantly driven by human activities like greenhouse gas emissions and deforestation, research efforts have been directed toward understanding environmental claims, particularly within the business sector. Stammbach et al. (2022) curated an environmental-driven dataset encompassing claims made by businesses, often within the finance sector. Their data collection included text from annual reports, conference calls, and sustainability statements. Notably, the choice to focus on the financial sector stems from its critical role in mitigating climate change. This curated collection, consisting of 3,000 binary datasets labeled as *Environmental claim* and *Negative example*, forms the basis of their study. It was noted that the various transformer models employed in the experiments consistently outperform the non-neural models.

However, despite these valuable contributions, there exists a notable gap in the study of climate change discourse, specifically within the multi-aspect analysis of climate discourse on social media platforms, warranting further exploration and investigation.

2.2. Multi-Aspect Annotated Data

Many studies have delved into analyzing multi-aspects within discourses across diverse contexts (Rauniyar et al., 2023). For instance, Gautam et al. (2020) compiled a dataset of 9,973 tweets centered around the MeToo movement. Through meticulous manual annotation, five linguistic dimensions were scrutinized: *stance, relevance, hate speech, dialogue acts,* and *sarcasm*. In a parallel context, Salawu et al. (2021) focused on cyberbullying and crafted an extensive multi-aspect dataset comprising 62,587 tweets. The dataset encompassed eight annotation aspects, including *Insult, Bullying, Profanity, Sarcasm, Threat, Exclusion, Porn,* and *Spam*. Remarkably, *Profanity* dominated with 51,014 occurrences, while *Exclusion* had the lowest count at 10. Similarly, Mollas et al. (2022) present the ETHOS dataset, encompassing two distinct versions: a balanced binary hate dataset and a multi-labeled dataset. The multi-labeled dataset contained 433 instances of hate speech annotated for eight distinct dimensions: *Violence, Directed vs. Generalised, Gender, Race, National Origin, Disability, Religion,* and *Sexual Orientation.* Such multi-aspect annotations helped to understand the intricacies of hate comments in detail.

While studying hate speech, it is also important to study the targets. Zampieri et al. (2019) presented a dataset of 14,100 tweets, categorizing offensive language and its targets in general social media discourse. Their analysis employed three baseline models to determine tweet nature and intended targets. The dataset utilized a three-tiered annotation scheme, encompassing binary labels for *Offensive* and *Targeted_insult* at the first and second levels, and three sub-labels (*Individual, Group,* and *Other*) at the third level. Similarly, Ousidhoum et al. (2019) undertook a multi-lingual and multi-aspect study using Twitter data encompassing three languages (French, English, and Arabic) and focusing on five aspects of annotation. Notably, discussions around highly contentious subjects, such as general feminism, English discourse on *illegal immigrants*, French discourse on *Islamo-gauchisme* (Islamic leftism), and Arabic discourse on *Iran*, often ignited heated debates. These exchanges frequently contained harmful remarks and insulting patterns.

Similarly, in climate change discourse, diverse perspectives lead to varied aspects of expression on social media platforms. This multifaceted nature underscores the need to identify and comprehend different dimensions of speech to foster a respectful online space. Our work adopts a multi-aspect approach to annotate the dataset, focusing on climate change and related activist movements, contribut-

ing to a nuanced understanding of climate change discourse.

3. Dataset

In this section, we describe our data collection process and annotation schema.

3.1. Data Collection

The data collection process encompassed tweets posted between January 1, 2022, and December 30, 2022. The selection criteria involved hashtags such as #climatecrisis, #climatechange, #ClimateEmergency, #ClimateTalk, #globalwarming, as well as activist-oriented hashtags like #fridaysforfuture, #actonclimate, #climatestrike, #extinctionrebellion, #ClimateAlliance, #climatejustice, #climateaction, etc. To collect this dataset, the Twitter API[1] was effectively employed, enabling the retrieval of tweets that matched the designated criteria within the specified timeframe. For tweet filtering, we only considered tweets composed in the English language. Finally, we annotated 15,309 tweets using the comprehensive annotation guidelines mentioned below.

3.2. Annotation Process

The effectiveness of datasets relies on accurate annotations; in the absence of meticulous annotations, model performance on downstream tasks can be significantly compromised, leading to distorted outcomes (Thapa et al., 2023; Assimakopoulos et al., 2020). To ensure high-quality annotations, a team of experienced annotators, comprising four members, was engaged. They were provided with comprehensive annotation guidelines encompassing specific tasks, associated labels, and illustrative examples. These annotators possessed a broad understanding of the climate-related matter, facilitating their comprehension of the guidelines and ensuring unbiased annotation. An iterative approach was adopted, incorporating annotator feedback into the guidelines to enhance the accuracy and consistency of the annotation process.

The annotation process was done for six specific tasks, each designed to identify aspects related to relevance, stance, hate speech, the direction of hate speech, targets of hate speech, and humor within the dataset. To address any inconsistencies or inaccuracies, a structured three-phase annotation schema was used. Furthermore, the data underwent thorough cross-checking to enhance its clarity and consistency. This methodical approach enhances the reliability and comprehensiveness of our work. The three-phase annotation schema consists of an initial dry run, an instruction revision phase, and a consensus-building and resolution phase.

- **Initial Dry Run:** We initiated the annotation process with an initial dry run involving the annotation of 50 sample tweets. This phase was crucial in gauging the comprehensibility and effectiveness of the annotation instructions and guidelines. Given the intricate nature of climate change discourse, annotators were introduced to the contextual intricacies, equipping them to navigate potential challenges. Initially, annotators faced confusion, particularly in identifying the humor and hate speech.
- **Instruction Revision Phase:** Building upon insights from the dry run, the annotation process entered a second phase where 200 additional tweets were annotated. During this phase, annotators were provided with refined instructions, which were adjusted based on the feedback from the initial dry run. This step aimed to enhance the clarity and precision of annotations, particularly in identifying hate speech and its targets.
- **Conflict Resolution:** In the final stage, annotators engaged in a collaborative discussion to address discrepancies that arose while annotating 200 tweets after the revision of instructions. This consensus-building process allowed for a thorough review of annotations and a shared understanding of the final guidelines. The resolution of occasional ambiguities was achieved through regular meetings and consultations with experts in annotation, including professors. The resolution of ambiguities ensured consistency and accuracy of annotations, enhancing the overall quality of the dataset.

3.3. Annotation Guidelines

To develop a pattern in the dataset and bring about consistency in our work, we devised detailed annotation guidelines to help the annotators. Given a tweet, it was annotated for various aspects.

A. Relevance: This annotation task aims to identify relevant tweets in the context of climate change discourse. It is important to note that the presence of climate change-related hashtags does not automatically indicate relevance to climate change. Annotators were guided to distinguish tweets directly relevant to the climate change discourse. Tweets that exploit climate change hashtags for spam or unrelated content were annotated as irrelevant to the topic.

[1] https://developer.twitter.com/en/docs/twitter-api

Task	Class	Examples
Task A: Relevance	Relevant	most important youth movement worldwide is the movement against global warming. we want the states to commit. #climatechange #globalwarming
	Non-relevant	Great question! Maybe fear doesn't grow while we sleep. For dreams to #evolve fear needs take a back seat #leonardodavinci #Artist #truth #quote #Jobs #fridaysforfuture
Task B: Stance Detection	Support	I am joining the Global Climate Strike as we demand policymakers and world leaders to prioritize #PeopleNotProfit! #FridaysForFuture
	Neutral	Heat wave in India and Pakistan is so frightening because the period above average is significantly long and seems like the next months will be a real challenge. #climatecrisis
	Oppose	why the wokeness as well as numerous astroturfing like #Unteilbar #FridaysForFuture #LetzteGeneration, which (are supposed to) divert society's attention to sideshows, are financed by the families of the huge fortunes.
Task C: Hate Speech Detection	Hate Speech	[username] Liars and haters spread these #fake-pictures. I would want everyone to know that you are a liar and should be shamed publicly. #climatechange #scammers
	No Hate Speech	We can't leave it all to [username] We all have a responsibility to move away from our fossil fuel addiction and make the right political choices. #climatestrike #canpoli
Task D: Direction of Hate Speech	Directed	Two days ago, [username] was found with [username]. No wonder they both are scamming us together in the name of climate change. #ClimateChange #FridaysForFuture
	Undirected	[username] It is all of us who are to be blamed for this! God will make us burn in hell! I hate everyone! #LossAndDamage #globalwarming
Task E: Targets of Hate Speech	Individual	Greta is brainwashing people on a problem that is non-existent. Brainwashing innocent teenagers isn't cool. #FridaysForFuture #Greta
	Organization	JICA and Sumitomo are trying to build a coal-fired power plant in Bangladesh. Let's kick them out of the nation! #jicanotwelcome #co2emission #ClimateJustice
	Community	climate change is the only problem of white people with certain hairstyles! #notmyproblem #ClimateChange #myth
Task F: Humor Detection	Humor	Trying to explain climate change to my cat: 'You see, Fluffy, the planet is getting warmer because humans are driving around in big metal boxes emitting invisible stuff called greenhouse gases.' #ClimateChange #greenhouse #ClimateConfusion
	No Humor	As global temperatures continue to rise due to human activity, urgent action is needed to mitigate the impacts of climate change on our planet. #ClimateChange #ClimateAction

Table 2: Examples of tweets for each class label across all the tasks.

B. Stance: The stance annotation task involves categorizing tweets into one of three groups: support, denial, or neutral, based on their context within climate change discourse.

- **Support**: Tweets falling under this category demonstrate alignment with climate change objectives and related activist movements. These tweets express agreement with efforts to combat climate change, advocate for sustainable practices, and address the consequences of climate change. Annotators were instructed to look for positive language endorsing climate change initiatives, enthusiasm for activist activities, and suggestions for dealing with climate-related challenges.
- **Denial**: Tweets in this category exhibit disagreement with climate change and related activist movements. The annotators were made to identify the tweets that deny the existence of climate change, oppose measures to reduce carbon emissions, or question the purpose of activist movements. The annotators were instructed to look for the negative language used to describe the movement, expressions of disagreement with its objectives, and outright denial of climate change's effects.
- **Neutral**: Neutral tweets do not express a definitive stance toward climate change or related activist movements. Annotators identified such tweets as those sharing factual information or news articles related to climate change without offering a personal opinion. Annotators looked for tweets that avoided sentiments or language indicating explicit support or denial and labeled them neutral.

C. Hate speech Identification: Annotators were tasked with identifying hate speech labels for tweets deemed relevant to climate change discourse. The focus was on recognizing instances of language that conveyed hateful sentiments or offensive content. Importantly, annotators were instructed to differentiate between tweets expressing strong disagreement without resorting to offensive language and those genuinely exhibiting hate speech elements. This distinction aimed to ensure that tweets containing genuinely hateful language were labeled as such.

D. Direction of Hate Speech: In this task, annotators were required to determine the direction of hate speech in relevant tweets. Specifically, they needed to ascertain whether the hate speech was directed towards specific individuals, groups, or entities or if it was of a more generalized nature. Annotators were instructed to categorize hate speech as "directed" when it targeted specific entities and "undirected" when it exhibited generalized hateful language without pinpointing specific targets.

E. Targets of Hate Speech: Within this aspect, annotators were tasked to identify the precise targets of directed hate speech within relevant tweets. These targets were systematically divided into three observable categories: “Individual”, “Organization”, and “Community”. The hate tweets were labeled as “individual” targets when they targeted an individual. These individuals could include stakeholders in climate change, political leaders, or activists. Similarly, the label "Organization" was assigned when the hate speech was directed toward corporate entities, institutions, or similar organizations. Finally, the “Community” label was assigned for instances where the hate speech was directed toward broader social, ethnic, or cultural groups. In cases where the intended targets of directed hate speech exhibited ambiguity or were not explicitly mentioned, annotators were instructed to assign the label that most confidently represented their assessment. Additionally, when faced with potential overlaps between “Community” and “Organization” categories, a collaborative resolution process was undertaken to ensure accurate categorization.

F. Humor: In this aspect of annotation, annotators were tasked with identifying tweets that exhibit humor, entertainment, or satire within the context of climate change or activist movements. Annotators looked for tweets that contained jokes, playful language, witty commentary, exaggeration, irony, or wordplay that aimed to evoke laughter or amusement. Annotators were guided to distinguish the presence of humor or satire in tweets, even if they involve a lighthearted take on serious topics. The objective was to capture the nuanced use of language for comedic or satirical effects, contributing to a more comprehensive understanding of the diverse expressions within the climate change discourse.

The comprehensive guideline and iterative annotation schema helped the annotators annotate for various sub-tasks with minimal errors. This helped us create a strong foundation for a proper dataset structure. Table 2 provides examples of tweets for each class label across all the tasks.

Task	Label	#Samples	%
Relevance	Relevant	10,407	67.9%
	Non-relevant	4,902	32.1%
Stance	Support	6,146	40.1%
	Denial	994	6.4%
	Neutral	3,267	21.3%
Hate Speech	Hate	1,277	8.3%
	Non-Hate	9,130	59.6%
Hate Direction	Directed	999	6.5%
	Undirected	278	1.8%
Hate Targets	Individual	804	5.2%
	Organization	151	0.9%
	Community	44	0.2%
Humor	Humor	163	1.06%
	Non-Humor	10,244	66.9%

Table 3: Distribution of class labels for all tasks

4. Dataset Analysis

This section offers analytical insights into the dataset. Table 3 and Figure 1 show the statistics and wordcloud of *ClimaConvo* dataset respectively.

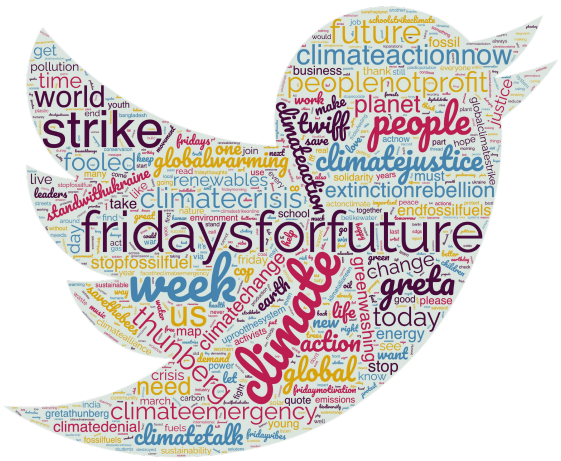

Figure 1: Wordcloud of words in ClimaConvo dataset. The font size of the words in the wordcloud are inline with how prominent the words are in the dataset as per their TF-IDF scores

4.1. Inter-annotator Agreement

Assessing the consistency of annotations is crucial to ensure the reliability of the results (Bhandari et al., 2023; Akhtar et al., 2019; Ljubešić et al., 2023). Thus, we utilized Fleiss' Kappa (κ) as a statistical measure (Falotico and Quatto, 2015) to evaluate the inter-annotator agreement across all six tasks. Our analysis using Fleiss' Kappa reveals

Annotation Phase	Annotators	$\kappa_{\text{task A}}$	$\kappa_{\text{task B}}$	$\kappa_{\text{task C}}$	$\kappa_{\text{task D}}$	$\kappa_{\text{task E}}$	$\kappa_{\text{task F}}$
Pilot Phase	α_1 and α_2	0.82	0.65	0.53	0.60	0.42	0.59
	α_1 and α_3	0.89	0.68	0.59	0.53	0.52	0.63
	α_1 and α_4	0.84	0.61	0.60	0.65	0.55	0.57
	α_2 and α_3	0.87	0.72	0.49	0.62	0.51	0.66
	α_2 and α_4	0.88	0.69	0.64	0.60	0.60	0.68
	α_3 and α_4	0.85	0.74	0.62	0.59	0.60	0.60
Final Phase	α_1 and α_2	0.89	0.73	0.67	0.69	0.63	0.77
	α_1 and α_3	0.95	0.79	0.74	0.76	0.71	0.73
	α_1 and α_4	0.94	0.79	0.75	0.73	0.69	0.81
	α_2 and α_3	0.94	0.81	0.69	0.79	0.73	0.70
	α_2 and α_4	0.95	0.80	0.71	0.69	0.63	0.80
	α_3 and α_4	0.96	0.84	0.72	0.78	0.74	0.84

Table 4: Cohen's Kappa (κ) for annotation during different phases by four annotators

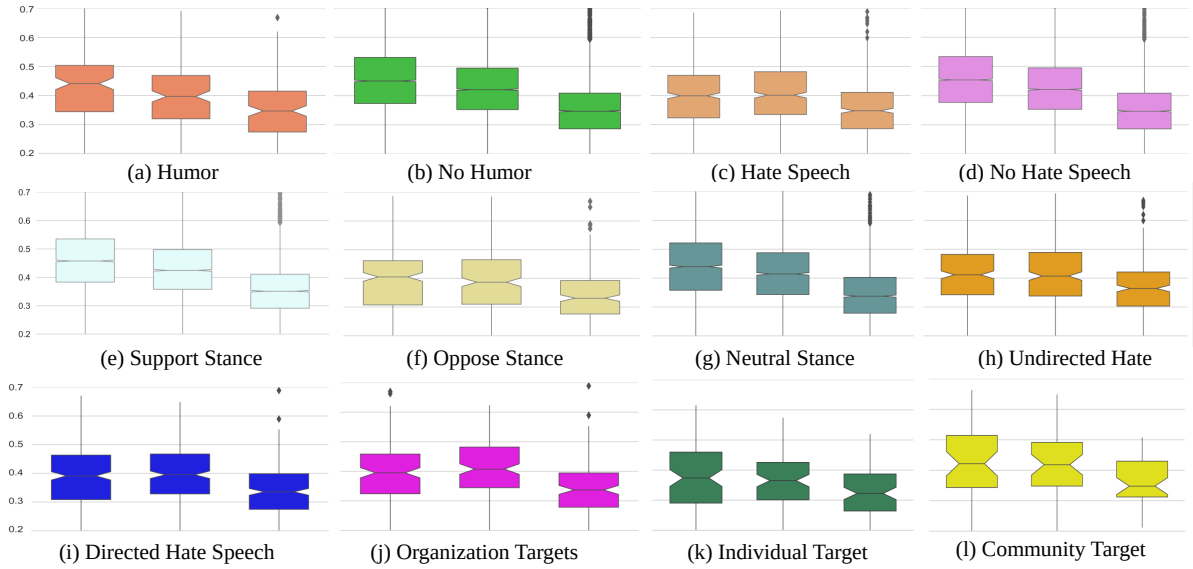

Figure 2: Valence, Dominance, and Arousal scores for various class labels based on NRC VAD lexicon. The first box denotes Valence, the second is Dominance, and the third is Arousal.

the following inter-annotator agreements: κ_{taskA} (2-class annotation of 'Relevant' or 'Non-Relevant') = 0.93, κ_{taskB} (3-class annotation of 'Support', 'Denial', or 'Neutral') = 0.78, κ_{taskC} (2-class annotation of 'Hate' or 'Non-Hate') = 0.70, κ_{taskD} (2-class annotation of 'Directed' or 'Undirected') = 0.74, κ_{taskE} (3-class annotation of 'Individual', 'Organization', or 'Community') = 0.67, and κ_{taskF} (2-class annotation of 'Humor' or 'Non-Humor') = 0.76. Further diving into different values of Cohen's Kappa (Blackman and Koval, 2000) scores at different phases of annotation shows that the 3-step annotation schema was helpful in getting more accurate and conflict-less annotations (Table 4).

4.2. Sentiment Analysis

We analyzed Valence (V), Arousal (A), and Dominance (D) across all classes utilizing the NRC VAD lexicon (Mohammad, 2018). We calculated the V, A, and D scores of each tweet within each class label by taking averages of the V, A, and D scores of words in the tweet. Then, label box plots were created for each class using the scores obtained for individual tweets. The outcomes have been consolidated in Figure 2. Among all classes, Individual Hate, Community Hate, and Oppose have the widest range of valence scores, likely attributed to the highly polarized opinions conveyed in those tweets. The lower levels of Arousal and higher levels of Dominance across all the class labels denote that the discourse involving climate change on Twitter is generally calm but authoritative in nature.

4.3. Keywords and Topic Modeling

We used a topic modeling method, SAGE (Eisenstein et al., 2011), to recognize significant words across various class labels in the dataset. SAGE (Sparse Additive Generative Models of Text) can

No Hate	SAGE	Hate Speech	SAGE
strike	0.70	puppet	1.91
today	0.65	denier	1.76
action	0.65	fake	1.71
week	0.63	gonna	1.69
climate	0.61	expect	1.68
Undirected Hate	**SAGE**	**Directed Hate**	**SAGE**
fake	2.32	puppet	2.38
fool	2.17	denier	2.30
gonna	2.09	white	2.27
expect	1.80	finland	2.08
drive	1.79	absolutely	2.00
Organization Hate	**SAGE**	**Community Hate**	**SAGE**
sumitomo	2.59	white	3.51
root	2.58	destructive	3.10
corporation	2.53	german	2.89
jica	2.47	bad	2.88
logging	2.42	run	2.84
Individual Hate	**SAGE**	**Stance: Neutral**	**SAGE**
puppet	3.09	china	1.01
denier	3.00	armung	0.94
thunberg	2.09	bau	0.91
etc	2.91	temperature	0.88
seriously	2.83	ppm	0.85
Stance: Support	**SAGE**	**Stance: Denial**	**SAGE**
join	0.34	white	2.52
prioritize	0.32	german	2.26
awesome	0.31	gonna	2.23
speak	0.29	freedom	2.15
reparation	0.25	etc	2.01
Humor	**SAGE**	**No Humor**	**SAGE**
stupid	3.51	climate	0.64
rest	2.41	strike	0.61
fool	2.40	week	0.57
shame	2.33	action	0.54
ice	2.29	today	0.52

Table 5: Top 5 words identified by SAGE topic modeling for various class labels

recognize words that separate specific corpus segments. For keyword assessment, we created a version of the dataset where hashtags and stopwords were excluded, which ensured that only relevant and meaningful words from the corpus were considered. Findings in Table 5 showcase the most significant keywords for each class identified by SAGE and their corresponding salience scores. In the Support label, words like *join* and *prioritize* seem relevant as they motivate people to join the FFF movement. Some similarities between denial and some hate labels can be seen, as denial tweets often hint at hate. Words like *corporation* and corporation names like *sumitoma* and *jica* seem relevant for the organization hate label. Similarly, words addressing a community like *white* and *German* seem relevant for the community hate label.

5. Baselines and Analysis

We performed baseline classification across all 6 tasks using advanced transformer models.

Task	Model	Acc ↑	F1 ↑	MMAE ↓
Task A: Relevance	BERT	0.811	0.785	0.206
	DistillBERT	0.802	0.782	0.220
	RoBERTa	0.813	0.795	0.209
	ClimateBERT	**0.825**	**0.812**	**0.193**
Task B: Stace Detection	BERT	0.586	0.466	0.633
	DistillBERT	0.610	0.527	0.583
	RoBERTa	0.648	0.542	0.595
	ClimateBERT	**0.651**	**0.545**	**0.583**
Task C: Hate Speech Detection	BERT	**0.901**	**0.708**	**0.322**
	DistillBERT	0.896	0.664	0.355
	RoBERTa	0.842	0.662	0.368
	ClimateBERT	0.884	0.704	0.338
Task D: Direction of Hate Speech	BERT	0.695	0.633	0.294
	DistillBERT	0.728	0.713	0.287
	RoBERTa	**0.750**	**0.747**	**0.251**
	ClimateBERT	0.630	0.627	0.362
Task E: Targets of Hate Speech	BERT	0.641	**0.554**	0.650
	DistillBERT	0.603	0.550	0.664
	RoBERTa	**0.716**	0.501	0.682
	ClimateBERT	0.604	0.549	**0.623**
Task F: Humor Detection	BERT	**0.921**	**0.565**	**0.451**
	DistillBERT	0.850	0.530	0.463
	RoBERTa	0.805	0.519	0.462
	ClimateBERT	0.818	0.519	0.464

Table 6: Performance of different algorithms on our ClimaConvo Dataset for different tasks

5.1. Baseline Models

We used four transformer-based models - BERT (Devlin et al., 2018), DistillBert (Sanh et al., 2019), RoBERTa (Liu et al., 2019), and ClimateBert (Webersinke et al., 2021) for performing benchmarks on our dataset. The first three models are transformer models trained on generic data from various domains. Whereas, ClimateBert is a BERT-based model trained on 2 million paragraphs of climate-related texts. We utilized the hugging face library to import the pre-trained models. Accuracy, Macro F1-score, and MMAE (Macro-averaged Mean Absolute Error) were the metrics we used to assess the performance of the models across all tasks. All models were run on a batch size of 16 for 3 epochs with a learning rate (LR) of 10^{-3} except DistillBERT which had an LR of 10^{-5}.

5.2. Performance Analysis and Insights

Table 6 shows how various algorithms perform across different tasks. BERT excels in Task C and Task F, achieving F-1 scores of 0.708 and 0.565, respectively. With an F-1 score of 0.747, RoBERTa performs the best in Task D. ClimateBERT, being further pre-trained on BERT, performs the best across two tasks: Task A and Task B. BERT has the highest F1-score for Task C. It is also essential to highlight that the smallest model, DistilBERT, does not come out on top in any task. Future works can focus on developing new methods to improve the performance on this task.

6. Conclusion

In this paper, we have presented *ClimaConvo*, a comprehensive multi-aspect annotated dataset comprising 15,309 tweets focused on climate change discourse and related activist movements. Our dataset provides a valuable resource for researchers in NLP and serves as a foundation for addressing pressing challenges within the climate change domain. The annotations capture the nuanced aspects of discourse, enabling researchers to develop and enhance models to understand better and engage with the complex discussions surrounding climate change. Moreover, our research endeavors have yielded several noteworthy findings. The employment of the SAGE model has facilitated topic modeling, enabling us to discern key themes and discussions within the dataset. Additionally, the analysis of valence, arousal, and dominance scores has shed light on the emotional tone of the tweets, contributing to a deeper understanding of sentiment dynamics. The sentiment analysis and benchmarking with state-of-the-art algorithms have provided valuable insights into the sentiment distribution and the performance of various models across different tasks. As climate change discourse continues to evolve, building a solid foundation for understanding the various dimensions of discussions is imperative, and our dataset contributes significantly in this direction. We anticipate this dataset will drive innovation, encourage scholarly cooperation, and ultimately contribute to a more informed and constructive discourse surrounding climate change and related activist movements.

7. Limitations

In this work, we introduce an extensive dataset for the identification of multiple aspects of speech like stance, humor, hate, and its targets in discourse involving climate change. We also present the analysis of our dataset using intuitive topic modeling methods, sentiment analysis and benchmarks across 6 tasks. However, it is vital to recognize the several constraints inherent in our study. Initially, our dataset originated solely from a single microblogging platform, Twitter, during a specific period of time when the FridayForFuture movement gained its pace and online climate discourse was greatly influenced by it. This might not holistically mirror the speech dynamics across more diverse contexts. Secondly, for tasks like hate target identification, the categorization schema we employed relies on overarching classification (Individual, Organizations, and Communities), potentially omitting more granular or intricate target designations. Moreover, the process of annotation is inherently subjective, with annotators potentially diverging on whether a particular tweet may qualify for a specific label. Thirdly, the benchmarks we establish rest on a limited assortment of characteristics, thereby leaving room for the possibility that alternative features or structures could yield enhanced performance. Lastly, it is crucial to note that the nature of this work which includes tasks like identification of hate speech and its targets, stance classification, and sarcasm detection may raise ethical questions, including potential bias. These ethical considerations warrant attention and resolution during the implementation of such technological solutions.

8. Ethical Considerations

While utilizing publicly available tweets removed the necessity for explicit informed consent from individual users, ensuring the privacy of users remained a pivotal ethical consideration. Throughout this study, rigorous measures were taken to anonymize all usernames and identifiable user information, for example, safeguarding their identities.Furthermore, we publicly release the dataset with only tweet IDs. This will allow users to have full privacy over their tweets. Others cannot access the data in case the user deletes the tweets or makes the profile private. For the annotations, we hired four experienced annotators and paid them a living wage as per the local rate. Considering the possible inclusion of sensitive and offensive language in the dataset, annotators were duly informed about the nature of the content they would encounter. The annotators were also made aware that they had access to mental health personnel in the institutions where the annotations were carried out. The supervising authors were also available by phone if the annotators had any concerns. The annotators did not waive any rights to withdraw from the annotation task which allowed the annotators to leave annotations anytime. It is important to acknowledge that, like with any annotated dataset, unintended bias might be present. However, our high kappa score demonstrates a high level of agreement among annotators, supporting the validity of the annotations.

We actively urge fellow researchers to factor in the environmental repercussions of their undertakings and to implement strategies that limit their carbon footprint. As a suggestion, we propose the utilization of carbon footprint assessment tools like the one introduced by Lacoste et al. (2019) to gauge the environmental consequences of their research initiatives.

9. Reproducibility Statement

In this paper, we provide hyperparameter information for easy reproducibility. The code and dataset are available in our GitHub repository: https://github.com/shucoll/ClimaConvo.

10. References

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2019. A new measure of polarization in the annotation of hate speech. In *AI* IA 2019–Advances in Artificial Intelligence: XVIIIth International Conference of the Italian Association for Artificial Intelligence, Rende, Italy, November 19–22, 2019, Proceedings 18*, pages 588–603. Springer.

Stavros Assimakopoulos, Rebecca Vella Muskat, Lonneke van der Plas, and Albert Gatt. 2020. Annotating for hate speech: The MaNeCo corpus and some input from critical discourse analysis. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5088–5097, Marseille, France. European Language Resources Association.

Aashish Bhandari, Siddhant B Shah, Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. Crisishatemm: Multimodal analysis of directed and undirected hate speech in text-embedded images from russia-ukraine conflict. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1993–2002.

Nicole J-M Blackman and John J Koval. 2000. Interval estimation for cohen's kappa as a measure of agreement. *Statistics in medicine*, 19(5):723–741.

Xingyu Chen, Lei Zou, and Bo Zhao. 2019. Detecting climate change deniers on twitter using a deep neural network. In *Proceedings of the 2019 11th international conference on machine learning and computing*, pages 204–210.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Radoslav S Dimitrov. 2016. The paris agreement on climate change: Behind closed doors. *Global environmental politics*, 16(3):1–11.

Jacob Eisenstein, Amr Ahmed, and Eric P Xing. 2011. Sparse additive generative models of text. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 1041–1048.

Rosa Falotico and Piero Quatto. 2015. Fleiss' kappa statistic without paradoxes. *Quality & Quantity*, 49:463–470.

Jennifer R Fownes, Chao Yu, and Drew B Margolin. 2018. Twitter and climate change. *Sociology Compass*, 12(6):e12587.

Akash Gautam, Puneet Mathur, Rakesh Gosangi, Debanjan Mahata, Ramit Sawhney, and Rajiv Ratn Shah. 2020. # metooma: Multi-aspect annotations of tweets related to the metoo movement. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 209–216.

Andrei P Kirilenko, Tatiana Molodtsova, and Svetlana O Stepchenkova. 2015. People as sensors: Mass media and local temperature influence climate change discussion on twitter. *Global Environmental Change*, 30:92–100.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Nikola Ljubešić, Igor Mozetič, and Petra Kralj Novak. 2023. Quantifying the impact of context on the quality of manual hate speech annotation. *Natural Language Engineering*, 29(6):1481–1494.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2021. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 174–184.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. Ethos: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*, 8(6):4663–4678.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. *arXiv preprint arXiv:1908.11049*.

Kritesh Rauniyar, Sweta Poudel, Shuvam Shiwakoti, Surendrabikram Thapa, Junaid Rashid, Jungeun Kim, Muhammad Imran, and Usman Naseem. 2023. Multi-aspect annotation and analysis of nepali tweets on anti-establishment election discourse. *IEEE Access*.

Andrew S Ross and Damian J Rivers. 2019. Internet memes, media frames, and the conflicting logics of climate change discourse. *Environmental communication*, 13(7):975–994.

Semiu Salawu, Jo Lumsden, and Yulan He. 2021. A large-scale english multi-label twitter dataset for cyberbullying and online abuse detection. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Alexandra Segerberg and W Lance Bennett. 2011. Social media and the organization of collective action: Using twitter to explore the ecologies of two climate change protests. *The Communication Review*, 14(3):197–215.

Dominik Stammbach, Nicolas Webersinke, Julia Anna Bingler, Mathias Kraus, and Markus Leippold. 2022. A dataset for detecting real-world environmental claims. *arXiv preprint arXiv:2209.00507*.

Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. From humans to machines: can chatgpt-like llms effectively replace human annotators in nlp tasks. In *Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media*.

Hannah Wallis and Laura S Loy. 2021. What drives pro-environmental activism of young people? a survey study on the fridays for future movement. *Journal of Environmental Psychology*, 74:101581.

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2021. Climatebert: A pretrained language model for climate-related text. *arXiv preprint arXiv:2110.12010*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. 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)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.