Understanding Disagreement: An Annotation Study of Sentiment and Emotional Language in Environmental Communication

Christina S. Barz
Faculty of Social Sciences
Darmstadt University
of Applied Sciences
Schöfferstraße 3
64295 Darmstadt
Germany
christina.barz@h-da.de

Melanie Siegel
Faculty of Computer Science
Darmstadt University
of Applied Sciences
Schöfferstraße 3
64295 Darmstadt
Germany
melanie.siegel@h-da.de

Daniel Hanss
Faculty of Social Sciences
Darmstadt University
of Applied Sciences
Schöfferstraße 3
64295 Darmstadt
Germany
daniel.hanss@h-da.de

Michael Wiegand
Digital Philology
Faculty of Philological and
Cultural Studies
University of Vienna
AT-1010 Vienna, Austria
michael.wiegand@univie.ac.at

Abstract

Emotional language is central to how environmental issues are communicated and received by the public. To better understand how such language is interpreted, we conducted an annotation study on sentiment and emotional language in texts from the environmental activist group Extinction Rebellion. The annotation process revealed substantial disagreement among annotators, highlighting the complexity and subjectivity involved in interpreting emotional language. In this paper, we analyze the sources of these disagreements, offering insights into how individual perspectives shape annotation outcomes. Our work contributes to ongoing discussions on perspectivism in NLP and emphasizes the importance of human-centered approaches and citizen science in analyzing environmental communication.

1 Introduction

Addressing the escalating environmental crises requires coordinated global action (IPCC, 2022; Fritsche and Masson, 2021). Emotions play a key role in motivating such action, shaping a range of behaviors from policy support to civil disobedience (Brosch, 2025; Schneider et al., 2021; Van Valkengoed and Steg, 2019).

Although there has been limited interdisciplinary research on the role of emotional language in environmental communication, existing studies suggest that such language can play a key role in mobilizing individuals for collective action (Salas Reyes et al., 2021; Kaushal et al., 2022; Zaremba et al., 2024). In this context, we define *emotional language* as the use of words or expressions that convey affective states. Importantly, we use the term *emotional language* - rather than emotion - to emphasize that our focus is on the strategic use of emotion-related expressions in group communication, rather than on measuring the actual felt emotions of individual speakers or writers. This distinction is particularly relevant when analyzing collective actors such as environmental groups, whose language is often shaped by strategic communication goals. However, the outcome of using emotional language in different socio-political contexts - especially in the discourse of groups with different ideologies, identities and thematic priorities - is still poorly researched and not well understood (Salas Reyes et al., 2021; Zaremba et al., 2024; Lehrer et al., 2023; Berger et al., 2019).

This paper is part of a broader project examining emotional language in environmental communication by highly visible and polarizing activist groups, and analyzing the emotional reactions such language provokes among the public (Barz et al., 2025). While the larger dataset includes multiple organizations, this study focuses on tweets from **Extinction Rebellion** (XR), a global activist group using nonviolent civil disobedience to demand urgent climate action. Our overarching goal is to develop a comprehensive, annotated dataset tai-

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1–20
July 5–10, 2020. ©2020 Association for Computational Linguistics

lored to environment-related communication, with applications in both environmental communication research and Natural Language Processing (NLP).

For this paper, we annotated sentiment and emotional language in XR's X (formerly Twitter) discourse, revealing substantial annotator disagreement. We analyze the factors driving this disagreement and explore how these insights can refine future annotation efforts in NLP and environmental communication research. Our findings highlight challenges in creating reliable annotated datasets and contribute to the broader debate on **perspectivism** in NLP, which recognizes that multiple valid interpretations of a text can coexist due to annotators' diverse backgrounds, experiences, and perspectives—challenging the notion of a single *ground truth* (Frenda et al., 2024; Uma et al., 2021; Rodríguez-Barroso et al., 2024).

To guide our investigation of these challenges and the implications of annotator subjectivity, our current work is structured around the following **research questions**:

RQ1 What factors may contribute to variation and disagreement in annotator labeling behavior?

RQ2 What insights can be gained from the observed disagreement, and how can they inform future annotation efforts?

The **main contributions** of this paper are as follows:

- We provide the first annotated and publicly available dataset of emotional language in XR's X discourse, contributing to the study of environmental communication.
- We perform analyses to systematically examine annotator disagreement, providing methodological insights into the influence of perspective in text annotation.
- We highlight the implications of perspectivism in annotation, demonstrating its relevance for both NLP applications and environmental communication research.

2 Related Work

This section reviews relevant literature on environmental communication as well as sentiment and emotion analysis.

2.1 Environmental Communication Studies

Environmental communication examines how humans perceive, discuss, and respond to environmental issues, with increasing attention to climate change communication (Carvalho and Peterson, 2024).

The study of environmental communication has gained prominence, particularly with social media's role in discourse and mobilization (Carvalho and Peterson, 2024; Schäfer, 2024; Lee et al., 2024; Amangeldi et al., 2024). Recent studies increasingly use computational methods, focusing on automated framing, discourse analysis, and translation studies (Hirsbrunner, 2024; Schäfer and Hase, 2023; Bird et al., 2024; Yasmin et al., 2024). However, NLP approaches beyond framing—such as sentiment, and emotion analysis—remain underexplored, despite emotional language's well-documented role in motivating collective action (Kaushal et al., 2022; Zaremba et al., 2024).

Research in this area has also predominantly analyzed news media (Anderson, 2024; Lahsen, 2022), prompting calls for broader investigations into the communication strategies of environmental groups and activist movements (Anderson, 2024).

2.2 Sentiment and Emotion Analysis, and Available Datasets

Emotion analysis is rarely applied to environmental communication, leading to a shortage of dedicated models and human-labeled datasets. Existing climate-related datasets primarily address sentiment, climate change denial, misinformation, or public opinion rather than emotional language (Stede and Patz, 2021). For instance, the *ClimaConvo* dataset includes 15,309 tweets from 2022 labeled for sentiment, climate change denial, hate speech, and humor (Shiwakoti et al., 2024). Similarly, the *Twitter Climate Change Sentiment Dataset* (Qian, 2021) comprises 43,943 tweets (2015–2018) labeled as news, pro (supporting anthropogenic climate change), neutral, or anti (rejecting anthropogenic climate change). A few datasets include emotional language, such as a collection of speeches by environmental activists, including Greta Thunberg, which focuses on anger (Ponton and Raimo, 2024). The *Emotional Climate Change Stories* (ECCS) dataset explores climate change storytelling and readers' emotional reactions, containing 180 short stories designed to evoke five emotions—anger, fear, com-

Climate Change and Sentiment Categories

Category	Example
Climate Detection	
About Climate Change	*Climate change is one of the greatest threats of our time.*
Climate Sentiment	
Positive/Opportunity	*Switching to renewable energy helps fight the climate crisis and creates new jobs.*
Negative/Risk	*Rising sea levels are threatening coastal cities around the world as average temperatures rise.*

Emotion Categories

Category	Example
Anger	*It's infuriating to see politicians ignore climate science!*
Concern	*Today we are disappointed and worried: The Supreme Court of Norway has chosen to back oil over our rights to a liveable future.*
Fear	*The alarming state of nature in the UK is a matter that should concern everyone.*
Hope	*Every tree planted is a step towards a healthier planet.*
Joy	*We're celebrating today as more cities commit to 100% renewable energy!*
Pride	*Proud of our community for coming together to reduce plastic waste!*
Sadness	*It's heartbreaking to witness the destruction of the Amazon rainforest.*
Solidarity	*In unity with our brothers and sisters across the globe, let's stand united for climate justice.*

Table 1: Annotation categories for multi-label document-level annotations and example tweets.

passion, guilt, and hope—as well as neutral stories (Zaremba et al., 2024).

To our knowledge, no dataset or study exclusively analyzes environmental organizations' or activist groups' communication. Most datasets capture individual opinions or personal expressions of sentiment and emotion within broader discourse (Dahal et al., 2019; El Barachi et al., 2021).

A key challenge in sentiment and emotion analysis is the inherent subjectivity of emotion recognition, especially in social media, where tone, context, and audience interpretation vary widely (Pozzi et al., 2016; Almeida et al., 2018). To address this, researchers have employed multi-label annotation approaches to allow overlapping emotional categories and dataset creation methods beyond majority voting to incorporate diverse perspectives (Mostafazadeh Davani et al., 2022; Alhuzali and Ananiadou, 2021).

3 Data and Annotation

This section outlines the dataset and annotation process used in our study.

3.1 Data

The dataset used in this study consists of 2,199 English-language tweets from the international activist group *Extinction Rebellion*, extracted in September 2024. The tweets were published between 2022 and 2024. The dataset includes the following metadata: group name, timestamp, retweet count, reply count, like count, and tweet ID. The complete dataset, including annotations, is provided in the supplementary materials and is publicly available to the research community at Hugging Face Datasets.

3.2 Annotation Process and Annotators

Our project employs **multi-label annotation**, where each tweet can be assigned multiple labels simultaneously from a predefined set of categories, reflecting the complex emotions and sentiments expressed. The annotations are made at the **document level**, meaning labels are applied to the entire tweet rather than single segments or sentences. This approach provides a compact and interpretable representation of each tweet. The dataset of 2,199 tweets was independently annotated by three experienced annotators. None of the annotators were involved in the authorship of this paper. To ensure consistency and clarity, we developed comprehensive annotation guidelines that provided clear definitions for each category, along with illustrative examples. The full guidelines are available in the supplementary material.

The annotation process was organized as follows: Initially, annotators labeled a small set of 10 tweets to familiarize themselves with the data format and task. Following this, each annotator participated in individual feedback sessions to address ambiguities and ensure alignment on labeling criteria. These sessions were conducted by one of the co-authors, who provided detailed guidance and clarification as needed. Periodic feedback sessions were held after every 500 tweets, allowing annotators to ask questions and resolve any issues that arose. While these sessions were conducted individually, all annotators received the same clar-

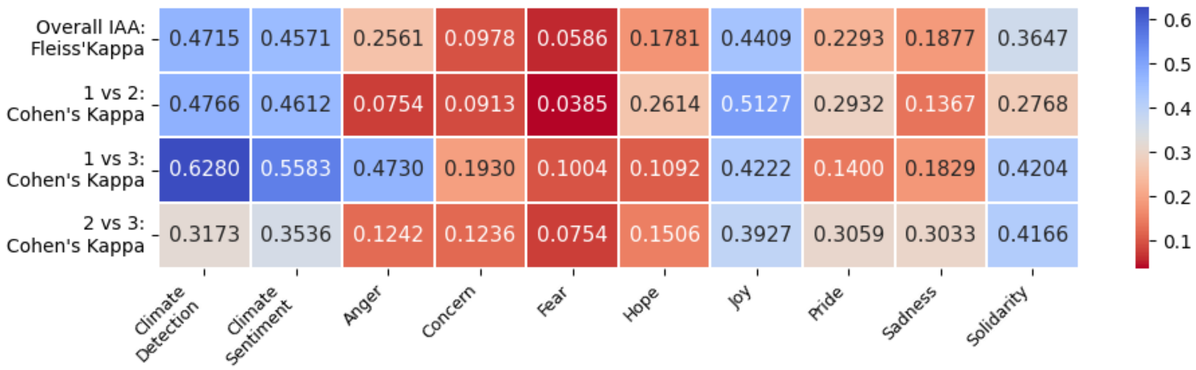

Figure 1: Heatmap displaying Fleiss' Kappa (Fleiss, 1971) and pairwise Cohen's Kappa coefficients (Cohen, 1960) to evaluate **overall and pairwise IAA** across all annotation categories.

ifications to maintain consistency across annotations. Any uncertainty raised by one annotator was systematically addressed with the others.

The annotators consisted of three paid research assistants, all proficient in English, female, and residing in Germany. Their academic backgrounds were as follows: Annotator 1 (A1) and Annotator 3 (A3) were students in *Business Psychology*, while Annotator 2 (A2) was a student in *Expanded Media*. Annotators were instructed to label tweets based on several categories: CLIMATE DETECTION (indicating whether a tweet relates to climate change), CLIMATE SENTIMENT (categorized as *risk*, *opportunity*, or *neutral*), and a set of emotion labels including ANGER, CONCERN, FEAR, HOPE, JOY, PRIDE, SADNESS, and SOLIDARITY, as outlined in Table 1. The climate detection and sentiment categories were adapted from prior annotation tasks and language models (Webersinke et al., 2021; Shiwakoti et al., 2024), while the emotional categories were refined through an in-depth qualitative analysis of a random sample from the larger dataset of several activist organizations in our project, identifying the most relevant emotions for the context. Annotators were instructed to assess **sentiment and emotion from the writer's perspective**.

Our dataset retains all annotations provided by the three annotators. This approach allows for the preservation of individual annotations, as they are central to our research focus.

4 Understanding Annotator Disagreement

To better understand the sources and implications of annotator disagreement in our dataset, we address our two research questions in two parts. First, we conduct a set of quantitative and qualitative analyses to identify factors that may contribute to variation in labeling behavior. Then, we reflect on the insights gained from these observations and how they can guide future annotation practices and research design.

4.1 Data Analysis

To address the factors that contribute to variation and disagreement in annotator labeling behavior (**RQ1**), we perform a number of analyses. In this section, we describe the approaches we use and the results we obtain for each of these analyses to answer **RQ1**.

Category	**Annotator**		
	1	**2**	**3**
CLIMATE DETECTION			
About Climate Change	647	461	805
CLIMATE SENTIMENT			
Risk	447	353	614
Opportunity	71	8	31
Emotions			
ANGER	269	55	184
CONCERN	566	54	151
FEAR	125	8	17
HOPE	150	74	33
JOY	32	22	33
PRIDE	38	9	4
SADNESS	61	9	30
SOLIDARITY	97	21	45

Table 2: **Absolute frequency distribution** per annotator for 2,199 tweets.

Label Distribution. We first examine individual annotation tendencies by counting the absolute frequencies of assigned labels. This allows us to identify differences in the annotators' labeling

ANGER			CONCERN			HOPE		
A1	**A2**	**A3**	**A1**	**A2**	**A3**	**A1**	**A2**	**A3**
murdering	tree	murdering	massively	corruption	warned	equitable	comments	**hope**
allow	hundred	**angry**	ongoing	threatening	massively	gather	expiration	touch
protested	immediate	denounce	escalating	reached	widely	preserve	helping	bit
false	helping	sleepwalking	allow	problems	horrific	joined	allowing	reasonable
lobbyists	training	address	twice	changing	suffer	motorway	degree	planning
sentence	lethal	hands	cultural	trust	deal	threats	faster	conference
murderous	claims	murderous	describes	result	ignore	achieve	linked	civilization
sleepwalking	politician	escalating	poorest	trees	positive	expiration	date	greed
polluting	camp	failure	tool	develop	propaganda	positive	ourselves	firm
exposing	release	behind	horrific	produce	further	voice	prevent	glass

Table 3: 10 words with the **highest PMI values** (listed from highest to lowest) for each annotator (A1, A2, A3) and the most frequent emotions, i.e., ANGER, CONCERN, and HOPE.

patterns and to assess the overall prevalence of categories in the dataset. Analysis of the label distributions across the three annotators (Table 2) reveals considerable variation in annotation choices. In particular, A2 assigns the fewest labels, indicating a more conservative approach, except for the category HOPE. In contrast, A1 and A3 tend to assign more labels, with A1 generally assigning the highest frequency. In addition, the categories PRIDE and JOY are the least frequently assigned across the dataset. The variation in the distribution of labels suggests that annotators may use different thresholds for identifying sentiment and emotional content.

Inter-Annotator Agreement. To assess the degree of agreement across categories, we compute both overall and pairwise IAA. The computed **Fleiss' Kappa** (Fleiss, 1971) values for all three annotators range from moderate agreement (0.4715 for CLIMATE DETECTION) to slight agreement (0.0586 for FEAR), with higher agreement observed for CLIMATE DETECTION, CLIMATE SENTIMENT, and JOY, as shown in Figure 1.
Low prevalence of categories generally results in lower IAA scores, as rare categories increase the likelihood of discrepancies between annotators (Artstein and Poesio, 2008). However, in our case, JOY-despite being one of the least frequently labeled emotions-has relatively high agreement. This suggests that while annotators identify JOY less frequently, when they do, they are more consistent in their judgments compared to other emotions. Notably, we do not find a clear relationship between category prevalence and IAA across the dataset.

To explore whether disagreement is linked to specific annotator pairs, we calculate **pairwise Cohen's Kappa** scores (Cohen, 1960), as shown in Figure 1. The results indicate that disagreement is not systematic, as no two annotators consistently exhibit a higher level of agreement while the third annotator deviates as an outlier across all categories. However, disagreement varies across pairs and categories; for example, A1 and A3 agree on ANGER with a score of 0.4730, while A1 and A2's agreement is only 0.0754. This variability suggests that subjectivity influences annotation, with more subjective categories showing lower agreement, and more objective categories like CLIMATE DETECTION and CLIMATE SENTIMENT showing higher agreement.

Pointwise Mutual Information. To address potential *lexical biases*—where certain words may lead annotators to consistently assign specific labels—we conducted a Pointwise Mutual Information (PMI) analysis for the most prevalent emotion categories (HOPE, ANGER, and CONCERN). PMI quantifies the strength of association between a word and a category by comparing their co-occurrence probability to what would be expected under independence, with higher PMI values indicating a stronger, non-random relationship (Church and Hanks, 1990). However, it is not appropriate for categories that are not frequently labeled. For infrequently labeled categories, the statistical reliability of the PMI is reduced because the occurrences of these categories are too sparse to yield meaningful associations.

Through our analysis, it became clear that A3 showed a lexical bias, paying close attention to words explicitly mentioning emotions, such as *hope* for HOPE and *angry* for ANGER (see Table 3). Our PMI analysis generally shows that annotations are not random, reflecting diverse associations for specific emotions. For example, A3 often assigns labels based on explicit emotional terms, while A1 links more indirect words such as *equitable* or

Topic ID	Topic Size	Topic Name
0	470	Global Fossil Fuel Protests
1	234	Extreme Weather and Climate Change
2	144	XR Decentralized Climate Advocacy
3	184	Climate Crisis and Health Responses
4	104	Climate Activism and Donations
5	88	Extreme Global Heat Events
6	89	Nonviolent Civil Disobedience in Movements
7	104	Climate Action and Sustainability
8	87	Plant-Based Diet and Agriculture
9	130	Peaceful Protest and Arrests
10	83	Citizens' Assemblies for Climate Action
11	94	Environmental Policy and Advocacy
12	99	Climate Change and Fascism Concerns
13	110	Climate and Resource Conflict in Congo
14	91	Critique of Economic Growth Models
15	45	Connecting with Local XR Groups
16	43	Environmental Pollution and Resource Extraction

Table 4: Topic modeling results from BERTopic including names generated by ChatGPT-4o and number of tweets categorized with this topic (OpenAI et al., 2024; Grootendorst, 2022).

achieve with HOPE, and *murdering* or *sentence* with ANGER. A2, in contrast, associates words like *comments* and *expiration* with HOPE, or *tree* and *hundred* with ANGER, indicating a stronger focus on context over specific words. For instance, A2 labeled the following tweet as expressing HOPE:

> That's an understandable doubt, Donald. However, the science isn't telling us a better world isn't possible. Surpassing 1.5C is a blow to everything we've been working towards, but there is no expiration on climate action. Every fraction of a degree saved counts.

Overall, the PMI analysis highlights distinct emotional associations and annotation strategies among annotators, as shown in Table 3.

Clustering-Based Topic Modeling. We applied **BERTopic** (Grootendorst, 2022) to examine potential *topic biases* in labeling the most prevalent emotion categories (i.e., HOPE, ANGER, and CONCERN). This clustering method leverages semantic embeddings and hierarchical density-based clustering (**HDBSCAN**) to automatically determine the number of clusters based on parameters such as *min_cluster_size*. To enhance interpretability, we used **ChatGPT-4o** to generate cluster names based on representative words (OpenAI et al., 2024). Our full parameter settings are provided in Table 5 in Appendix B. We clustered the dataset into 17 distinct topics (see Figure 2 for the resulted topics). Subsequently, we analyzed the most prevalent topics within tweets labeled with specific emotions for each annotator. The results indicate that annotators associated emotions with different topics, particularly in the case of HOPE (see Figure 2). In contrast, the emotions ANGER and CONCERN show greater overlap in their most frequently assigned topics; these results are included for completeness in Figures 5 and 6 in Appendix B.

Additionally, we computed **pairwise Cohen's Kappa scores** (Cohen, 1960) for each topic, revealing substantial variation in agreement across topics. This suggests that annotator disagreement is topic-dependent rather than systematic (see Figures 7, 8, and 9, Appendix B).

Temporal Analysis. We conducted a temporal analysis by calculating the mean labels for every set of 100 annotated tweets per annotator to track shifts in annotation patterns over time. The trends show that A1 assigned more emotion labels at the beginning of the annotation process compared to later stages, and also more than the other annotators (see Figure 4 in Appendix A). This could be due to the familiarization process, where annotators typically experience fluctuations at the start of the task, potentially influenced by feedback discussions during the initial phase. Other factors, such as annotators' daily moods or emotional states, and external influences like media exposure to environmental issues, could also have biased annotation patterns (Gautam and Srinath, 2024; Bodenhausen et al., 2000; Englich and Soder, 2009; Vrselja et al., 2024).

Spearman Correlations. To assess co-labeling

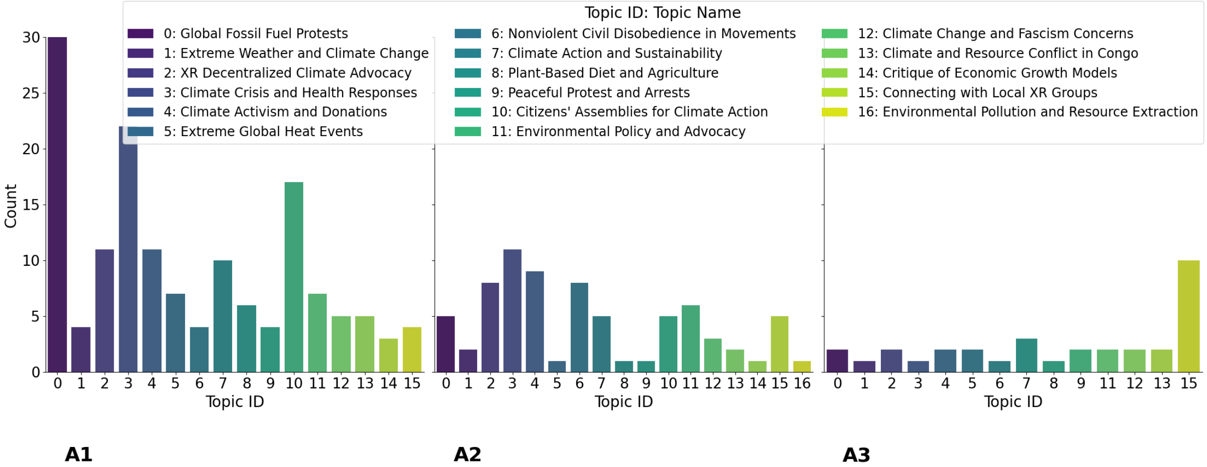

Figure 2: Plots showing the **count of tweets by topic** labeled with the emotion HOPE per annotator (A1, A2, A3).

frequency and potential difficulties in distinguishing categories, we calculated Spearman correlations (Spearman, 1904) for all label pairs separately for each annotator. With correlations of up to 0.33 between most positive emotions, we observe that A1 and A2 have higher correlations in some cases, reflecting a higher number of co-labels (see Figure 3 for the correlation patterns associated with A1). Conversely, correlations for A3 labels are predominantly near to zero. This suggests varying interpretations of emotions, particularly in their differentiation. For A1 and A2, positive emotions appear to be more closely related than for A3. Additionally, a topic bias was clearly observed, as A1 showed a correlation of 0.28 between CLIMATE DETECTION and CONCERN, indicating that tweets on climate change were more often labeled with CONCERN. Correlation matrices for all annotators are included in Figures 10, and 11 in Appendix C for completeness and detailed reference.

Qualitative Interviews. To explore sources of disagreement, we conducted qualitative interviews with all three annotators. These aimed at understanding individual perspectives rather than drawing statistical inferences.

All annotators reported following the same procedure that had been instructed, feeling confident in their understanding of the task, and recognizing that they should label emotions from the writer's perspective. However, they differed in their **emotional responses to environmental crises**. A1 primarily experiences *concern*, while also labeling CONCERN the most. A2's response is dominated by *anger*, which is also their most frequently assigned negative emotion. A3, despite reporting *fear* as their dominant reaction, labeled it the least. These differences may hint at subtle personal tendencies, as A1 and A2 more frequently assigned emotion labels that align with their own reported emotional reactions. We also explored annotators' **mental imagery or immediate associations with environmental groups**. A1 mentioned groups such as *Extinction Rebellion* and *Last Generation* and labeled more emotions overall, which might suggest a perceived link between radical activism and emotional expressiveness (Ostarek et al., 2024). In contrast, A2 and A3 associated environmental groups with *Fridays for Future* and *Greenpeace* and labeled fewer emotions, possibly reflecting differences in how they perceive the emotional tone of these groups.

Another key factor was **personal affectedness**. A1 did not consider themselves personally affected, while A2 described their perceived affectedness in their home country of Nigeria and A3 reported an indirect sense of affectedness, emphasizing empathy for strongly affected populations worldwide. Notably, A1, despite feeling the least affected, labeled the highest number of emotions.

External factors may have also played a significant role. A3 **engaged with climate news** daily, A1 consumed little, and A2 had difficulty engaging with environmental news due to emotional reactions, often avoiding such content. However, no clear link emerged between news consumption and annotation behavior. Procedural influences, such as annotation guidelines and feedback discussions, may have shaped interpretations, along with

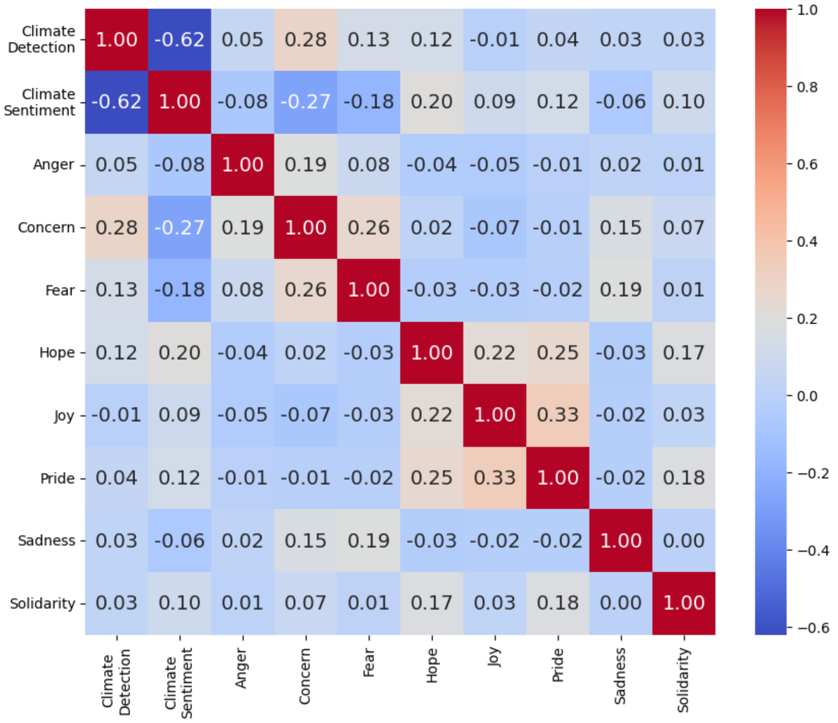

Figure 3: Spearman correlation (Spearman, 1904) matrix of the categories labeled by A1, showing the common occurrences of the labels.

differences in prior knowledge and familiarity with environmental discourse.

Final Considerations. Previous research has shown that distinguishing between annotation errors and perspectivism can be challenging (Weber-Genzel et al., 2024). However, given our research focus on understanding how individuals interpret environmental communication, we argue that variation in annotation tendencies is meaningful rather than problematic. Our study assumes that reading and interpreting environmental texts is inherently subjective, with recipient perspectives playing a crucial role in annotation outcomes. While factors such as annotation guidelines, feedback discussions, and annotator expertise may influence annotation subjectivity, they do not invalidate the presence of diverse and valuable perspectives in the data. This assumption aligns with prior research showing that emotion labeling is inherently subjective (Buechel and Hahn, 2022; Du et al., 2023), a tendency that is likely amplified in highly visible and polarized topics such as environmental activism (Ostarek et al., 2024).

4.2 Insights gained from Analysis

In this section, we discuss the valuable insights that can be gained from the observed disagreement in our annotations and how these insights can help inform future annotation efforts, addressing **RQ2**. While our analyses provide an initial understanding of the variability in annotation outcomes, the conclusions drawn are specific to our dataset and annotation context, and may not be easily generalized beyond this study.

The diversity in perspectives reflected in our annotations may be influenced by both internal and external factors. To improve the quality and reliability of future annotation efforts, it is crucial to systematically account for these influences. We acknowledge that high-quality annotations, as well as our proposed strategies to enhance them, come with increased resource demands, which are constrained by available research funding. Nevertheless, we aim to propose best practices that can be adapted based on available resources.

One potential approach is to collect **annotator-specific metadata** prior to annotation, including sociodemographic variables, domain expertise, prior engagement with the topic, personal stance, and emotional disposition toward the subject matter.

Additionally, intra-annotator variability should be considered by incorporating **daily self-reports** on factors such as recent exposure to the topic through media consumption, current emotional states, and subjective attitudes on the day of annotation. Furthermore, **external contextual variables**, such as ongoing political events or environmental incidents (e.g., natural disasters), should be tracked on a daily or weekly basis. Controlling for these factors would enable a more nuanced understanding of annotator subjectivity and facilitate structured dataset curation, allowing for more interpretable and representative NLP models. This approach aligns with the principles of **human-centered NLP**, which advocate for the explicit modeling of annotator subjectivity and diversity to enhance the interpretability and fairness of computational models (Soni et al., 2024; Kotnis et al., 2022).

Ideally, annotations should either be **representative of diverse perspectives or fully stratified into distinct target audience segments**. A potential implementation of this perspective-aware annotation strategy could involve weak perspectivism, where separate datasets are curated for different audience segments, with majority voting applied within each segment to create internally consistent annotations (Cabitza et al., 2023; Holovenko, 2024). Given that our research focuses on environmental communication, integrating author perspectives into the annotation process—akin to **citizen science**—could be highly beneficial when feasible (Paramonov and Poletaev, 2024; Bono et al., 2023; Klie et al., 2023). For instance, members of XR could annotate texts to better capture the writer's perspective, while non-members could provide annotations reflecting the reader's perspective. Alternatively, Large Language Models (LLMs) could be leveraged to infer writer intentions based on linguistic cues, while reader perceptions could be analyzed separately through annotations segmented by audience groups.

5 Conclusions and Future Work

This study examines disagreement in environmental communication annotation, particularly within activist group discourse. Our findings highlight the impact of internal factors, such as sociodemographic backgrounds and emotions, and external factors like the annotation process. These challenges hinder achieving high IAA in subjective language assessment, especially in emotionally charged topics like environmental activism. Our results align with previous research questioning the idea of a single ground truth in annotation tasks (Cabitza et al., 2023; Uma et al., 2021; Rodríguez-Barroso et al., 2024; Valette, 2024). Perspectivism in NLP tasks, such as hate speech detection and emotion recognition, underscores the role of individual annotators' perspectives on labeling outcomes (Abercrombie et al., 2024; Larimore et al., 2021; Frenda et al., 2024; Fleisig et al., 2023; Xu et al., 2024; Abercrombie et al., 2023; Du et al., 2023). This subjectivity is critical in environmental communication, where diverse reactions provide valuable insights into audience perceptions. Importantly, disagreements among annotators reveal the varied emotional engagement with environmental issues (Cabitza et al., 2023; Zaremba et al., 2024).

Future research should improve annotation methods to better address subjectivity. Adopting perspectivist frameworks, using pre-annotation surveys to capture annotators' backgrounds, and integrating LLMs to complement human labeling are promising approaches. Expanding our dataset to include more environmental groups and studying the temporal aspects of annotation subjectivity, such as emotions or external events, could offer further insights. Ultimately, applying these findings to tailor environmental communication strategies for diverse audiences will be crucial in bridging NLP and environmental communication.

Limitations

While our study provides valuable insights, it is imperative to acknowledge its limitations. First, the analysis is based on a relatively small group of annotators (n=3), all of whom are female students residing in Germany. While this approach is useful for an in-depth exploration of subjectivity, it limits the generalizability of our findings. Despite these limitations, our study is a first attempt to understand perspectivism in environmental communication. To enhance the range of perspectives that can be captured, future studies should aim to recruit a more diverse and larger pool of annotators. Second, the dataset consists solely of tweets from XR, a highly visible and polarizing activist group. While this allows for a focused analysis, it does not account for the full diversity of environmental communication used by different organizations. While we assume a higher likelihood that this group employs more radical and emotionally charged language, other groups may exhibit significantly less emotional language in their communication. Expanding the dataset to include posts from a wider range of environmental groups would enhance the robustness of the findings.

Third, part of our study relies on qualitative interviews conducted after the annotation process to infer annotator subjectivity. While these interviews provide valuable self-reported insights, they do not allow for real-time tracking of changes in annotation tendencies over time. Furthermore, it is not clear whether the results of the interviews depend on the previous annotations. For example, an annotator may have reported more concern about environmental crises simply because they labeled it more frequently in the tweets.

Additionally, we did not check reliability by giving our annotators the same tweets a second time. Implementing daily or real-time self-assessments during the annotation process would provide a more precise and accurate measurement of fluctuating annotator subjectivity.

Ethical Considerations

The annotation process involved reading environmental and climate-related texts, some of which addressed extreme weather events or broader environmental crises. Such content may evoke strong emotional responses, including feelings of eco- or climate anxiety, which can impact annotators' well-being. All annotators were financially compensated for their work, which involved engaging with potentially repetitive and emotionally challenging content.

To address these concerns, we took steps to protect the annotators' mental well-being. Annotators were informed that they could pause or discontinue the task at any time without providing a reason. We regularly checked in with them about their well-being during the annotation process and provided contact information for support services in case of psychological distress. Additionally, the annotators were fully informed about the purpose of their work, including the creation of a dataset for research purposes.

We also treated annotators' personal information with care. All sociodemographic data and mentions of individual annotators included in this paper were disclosed with their explicit consent.

Regarding the dataset, the collection and planned publication of tweet IDs were reviewed and approved in consultation with the university's data protection officer. The dataset does not contain personal data, as we only worked with group-level content (i.e., tweets published by the environmental activist group *Extinction Rebellion*). All usernames appearing in the dataset were anonymized, except for public figures such as politicians, in accordance with established ethical guidelines for working with social media data.

Acknowledgments

We would like to thank our student assistants, who preferred to remain anonymous, for their valuable contributions to the annotation work. This research was funded by the Office of Equal Opportunities at Darmstadt University of Applied Sciences, as well as by additional internal university funding.

References

Gavin Abercrombie, Dirk Hovy, and Vinodkumar Prabhakaran. 2023. Temporal and second language influence on intra-annotator agreement and stability in hate speech labelling. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 96–103, Toronto, Canada. Association for Computational Linguistics.

Gavin Abercrombie, Nikolas Vitsakis, Aiqi Jiang, and Ioannis Konstas. 2024. Revisiting annotation of online gender-based violence. In *Joint International Conference on Computational Linguistics, Language Resources and Evaluation 2024*, pages 31–41. ELRA Language Resources Association.

Hassan Alhuzali and Sophia Ananiadou. 2021. SpanEmo: Casting multi-label emotion classification as span-prediction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1573–1584, Online. Association for Computational Linguistics.

Alex MG Almeida, Ricardo Cerri, Emerson Cabrera Paraiso, Rafael Gomes Mantovani, and Sylvio Barbon Junior. 2018. Applying multi-label techniques in emotion identification of short texts. *Neurocomputing*, 320:35–46.

Daniyar Amangeldi, Aida Usmanova, and Pakizar Shamoi. 2024. Understanding environmental posts: Sentiment and emotion analysis of social media data. *IEEE Access*.

Alison Anderson. 2024. *Advancing the environmental communication field: A research agenda*, pages 47–68. De Gruyter Mouton, Berlin, Boston.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Christina Barz, Melanie Siegel, and Daniel Hanss. 2025. Analyzing the online communication of environmental movement organizations: NLP approaches to topics, sentiment, and emotions. In *1st Workshop on Ecology, Environment, and Natural Language Processing*.

Natalie Berger, Ann-Kathrin Lindemann, and Gaby-Fleur Böl. 2019. Wahrnehmung des Klimawandels durch die Bevölkerung und Konsequenzen für die Risikokommunikation. *Bundesgesundheitsblatt-Gesundheitsforschung-Gesundheitsschutz*, 62(5):612–619.

Steven Bird, Angelina Aquino, and Ian Gumbula. 2024. Envisioning NLP for intercultural climate communication. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 111–122, Bangkok, Thailand. Association for Computational Linguistics.

Galen V Bodenhausen, Shira Gabriel, and Megan Lineberger. 2000. Sadness and susceptibility to judgmental bias: The case of anchoring. *Psychological Science*, 11(4):320–323.

Carlo Bono, Mehmet Oğuz Mülâyim, Cinzia Cappiello, Mark James Carman, Jesus Cerquides, Jose Luis Fernandez-Marquez, Maria Rosa Mondardini, Edoardo Ramalli, and Barbara Pernici. 2023. A citizen science approach for analyzing social media with crowdsourcing. *IEEE Access*, 11:15329–15347.

Tobias Brosch. 2025. From individual to collective climate emotions and actions: A review. *Current Opinion in Behavioral Sciences*, 61:101466.

Sven Buechel and Udo Hahn. 2022. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. *arXiv preprint arXiv:2205.01996*.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6860–6868.

Anabela Carvalho and Tarla Rai Peterson. 2024. *Rethinking environmental communication scholarship*, pages 3–6. De Gruyter Mouton, Berlin, Boston.

Kenneth W. Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Biraj Dahal, Sathish AP Kumar, and Zhenlong Li. 2019. Topic modeling and sentiment analysis of global climate change tweets. *Social Network Analysis and Mining*, 9:1–20.

Quanqi Du, Sofie Labat, Thomas Demeester, and Veronique Hoste. 2023. Unimodalities count as perspectives in multimodal emotion annotation. In *2nd Workshop on Perspectivist Approaches to NLP (NLPerspectives 2023), co-located with the 26th European Conference on Artificial Intelligence (ECAI 2023)*, volume 3494. CEUR-WS. org.

May El Barachi, Manar AlKhatib, Sujith Mathew, and Farhad Oroumchian. 2021. A novel sentiment analysis framework for monitoring the evolving public opinion in real-time: Case study on climate change. *Journal of Cleaner Production*, 312:127820.

Birte Englich and Kirsten Soder. 2009. Moody experts—how mood and expertise influence judgmental anchoring. *Judgment and Decision Making*, 4(1):41–50.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. *arXiv preprint arXiv:2305.06626*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378.

Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2024. Perspectivist approaches to natural language processing: A survey. *Language Resources and Evaluation*, pages 1–28.

Immo Fritsche and Torsten Masson. 2021. Collective climate action: When do people turn into collective environmental agents? *Current Opinion in Psychology*, 42:114–119.

Sanjana Gautam and Mukund Srinath. 2024. Blind spots and biases: Exploring the role of annotator cognitive biases in NLP. *arXiv preprint arXiv:2404.19071*.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Simon David Hirsbrunner. 2024. Computational methods for climate change frame analysis: Techniques, critiques, and cautious ways forward. *Wiley Interdisciplinary Reviews: Climate Change*, 15(5):e902.

Anastasia Holovenko. 2024. What are your triggers? Context-dependent detection of emotional triggers in influence campaigns. Master's thesis, Ukrainian Catholic University.

IPCC. 2022. Climate change 2022: Impacts, adaptation, and vulnerability. *Intergovernmental Panel on Climate Change*.

Sanjay Kaushal, Sarvsureshth Dhammi, and Anamita Guha. 2022. Climate crisis and language–a constructivist ecolinguistic approach. *Materials Today: Proceedings*, 49:3581–3584.

Jan-Christoph Klie, Ji-Ung Lee, Kevin Stowe, Gözde Gül Şahin, Nafise Sadat Moosavi, Luke Bates, Dominic Petrak, Richard Eckart De Castilho, and Iryna Gurevych. 2023. Lessons learned from a citizen science project for natural language processing. *arXiv preprint arXiv:2304.12836*.

Bhushan Kotnis, Kiril Gashteovski, Julia Gastinger, Giuseppe Serra, Francesco Alesiani, Timo Sztyler, Ammar Shaker, Na Gong, Carolin Lawrence, and Zhao Xu. 2022. Human-centric research for NLP: Towards a definition and guiding questions. *arXiv preprint arXiv:2207.04447*.

Myanna Lahsen. 2022. Evaluating the computational ("big data") turn in studies of media coverage of climate change. *Wiley Interdisciplinary Reviews: Climate Change*, 13(2):e752.

Savannah Larimore, Ian Kennedy, Breon Haskett, and Alina Arseniev-Koehler. 2021. Reconsidering annotator disagreement about racist language: Noise or signal? In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 81–90.

Bruce Y Lee, Brian Pavilonis, Danielle C John, Jessie Heneghan, Sarah M Bartsch, and Ilias Kavouras. 2024. The need to focus more on climate change communication and incorporate more systems approaches. *Journal of Health Communication*, 29(sup1):1–10.

Lena Lehrer, Lennart Hellmann, Hellen Temme, Leonie Otten, Johanna Hübenthal, Mattis Geiger, Mirjam A Jenny, and Cornelia Betsch. 2023. Communicating climate change and health to specific target groups. *Journal of Health Monitoring*, 8(Suppl 6):36.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Aaron OpenAI, Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Markus Ostarek, Brent Simpson, Cathy Rogers, and James Ozden. 2024. Radical climate protests linked to increases in public support for moderate organizations. *Nature Sustainability*, pages 1–7.

IV Paramonov and A Yu Poletaev. 2024. Annotation of text corpora by sentiment and irony in a project of citizen science. *Automatic Control and Computer Sciences*, 58(7):797–807.

Douglas Mark Ponton and Anna Raimo. 2024. Comparative discourse strategies in environmental advocacy: Analysing the rhetoric of Greta Thunberg and Chris Packham. *Languages*, 9(9):307.

Federico Alberto Pozzi, Elisabetta Fersini, Enza Messina, and Bing Liu. 2016. *Sentiment analysis in social networks*. Morgan Kaufmann.

Ed Qian. 2021. Twitter climate change sentiment dataset. Accessed: 2024-12-30.

Nuria Rodríguez-Barroso, Eugenio Martínez Cámara, Jose Camacho Collados, M Victoria Luzón, and Francisco Herrera. 2024. Federated learning for exploiting annotators' disagreements in natural language processing. *Transactions of the Association for Computational Linguistics*, 12:630–648.

Raúl Salas Reyes, Vivian M Nguyen, Stephan Schott, Valerie Berseth, Jenna Hutchen, Jennifer Taylor, and Nicole Klenk. 2021. A research agenda for affective dimensions in climate change risk perception and risk communication. *Frontiers in Climate*, 3:751310.

Mike S Schäfer. 2024. Social media in climate change communication: State of the field, new developments and the emergence of generative AI. *Dialogues on Climate Change*, page 29768659241300666.

Mike S Schäfer and Valerie Hase. 2023. Computational methods for the analysis of climate change communication: Towards an integrative and reflexive approach. *Wiley Interdisciplinary Reviews: Climate Change*, 14(2):e806.

Claudia R Schneider, Lisa Zaval, and Ezra M Markowitz. 2021. Positive emotions and climate change. *Current Opinion in Behavioral Sciences*, 42:114–120.

Shuvam Shiwakoti, Surendrabikram Thapa, Kritesh Rauniyar, Akshyat Shah, Aashish Bhandari, and Usman Naseem. 2024. Analyzing the dynamics of climate change discourse on twitter: A new annotated corpus and multi-aspect classification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 984–994.

Nikita Soni, H. Andrew Schwartz, João Sedoc, and Niranjan Balasubramanian. 2024. Large human language models: A need and the challenges. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8631–8646, Mexico City, Mexico. Association for Computational Linguistics.

Charles Spearman. 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

Manfred Stede and Ronny Patz. 2021. The climate change debate and natural language processing. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 8–18.

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Mathieu Valette. 2024. What does perspectivism mean? An ethical and methodological countercriticism. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 111–115, Torino, Italia. ELRA and ICCL.

Anne M Van Valkengoed and Linda Steg. 2019. Meta-analyses of factors motivating climate change adaptation behaviour. *Nature Climate Change*, 9(2):158–163.

Ivana Vrselja, Mario Pandžić, Martina Lotar Rihtarić, and Maria Ojala. 2024. Media exposure to climate change information and pro-environmental behavior: The role of climate change risk judgment. *BMC Psychology*, 12(1):262.

Leon Weber-Genzel, Siyao Peng, Marie-Catherine De Marneffe, and Barbara Plank. 2024. VariErr NLI: Separating annotation error from human label variation. *arXiv preprint arXiv:2403.01931*.

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*.

Jin Xu, Mariët Theune, and Daniel Braun. 2024. Leveraging annotator disagreement for text classification. *arXiv preprint arXiv:2409.17577*.

Musarat Yasmin et al. 2024. Framing vulnerability: An ecolinguistic analysis of gender and climate change discourse. *Current Research in Environmental Sustainability*, 7:100258.

Dominika Zaremba, Jarosław M Michałowski, Christian A Klöckner, Artur Marchewka, and Małgorzata Wierzba. 2024. Correction: Development and validation of the emotional climate change stories (eccs) stimuli set. *Behavior Research Methods*, 56(7):8158.

A Temporal Analysis

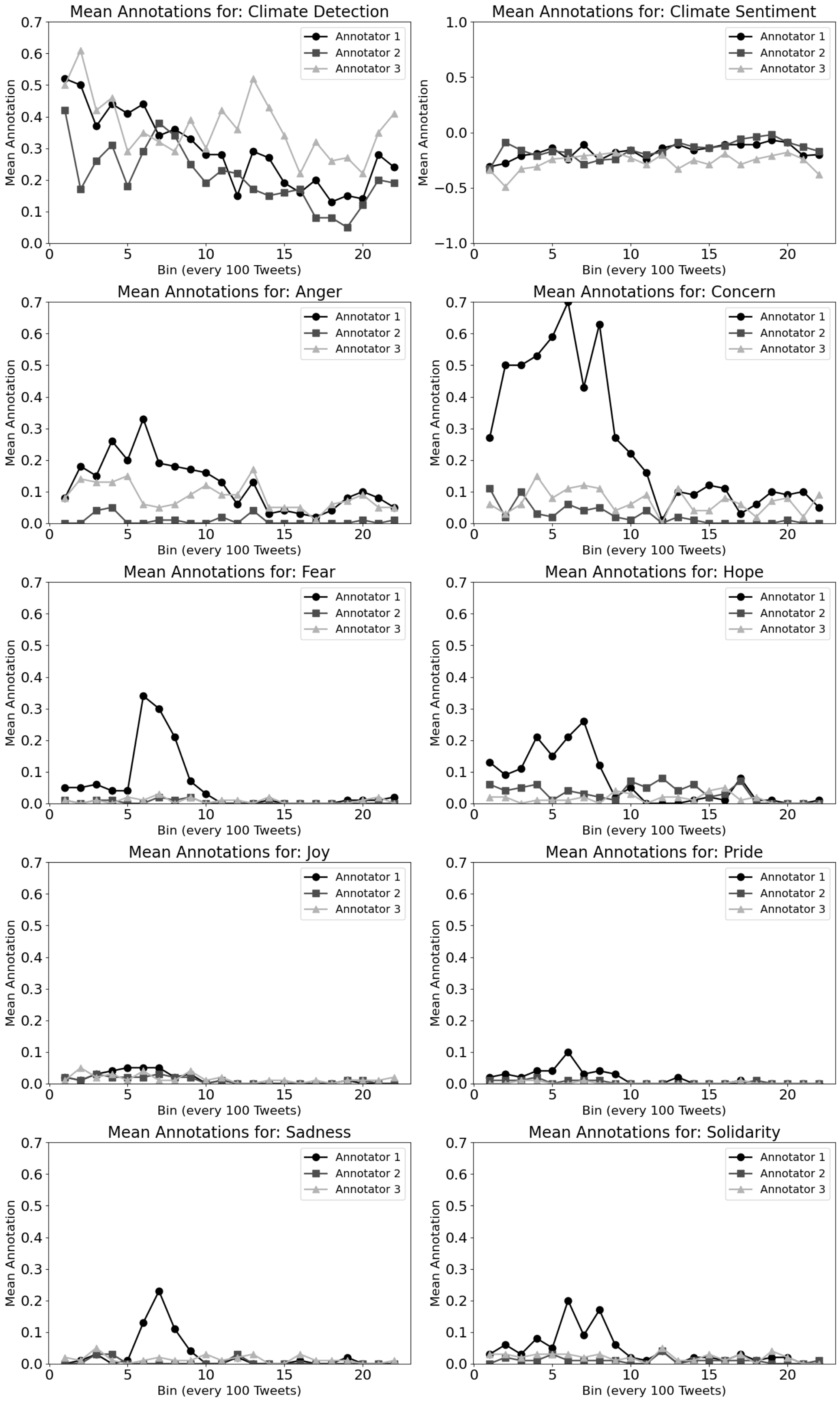

Figure 4: Set of plots showing the **distribution of true labels** assigned by each annotator across specific categories, illustrating the amount of labels given per category **over time**.

B Clustering-Based Topic Modeling

Component	Setting
Embedding Model	SentenceTransformer("all-MiniLM-L6-v2")
UMAP Configuration	random_state=777, n_neighbors=29

HDBSCAN Configuration	metric='euclidean', min_cluster_size=31,
	cluster_selection_method='eom',
	prediction_data=True, min_samples=5

Table 5: Parameter settings used for BERTopic modeling (Grootendorst, 2022).

B.1 Topics for Anger and Concern

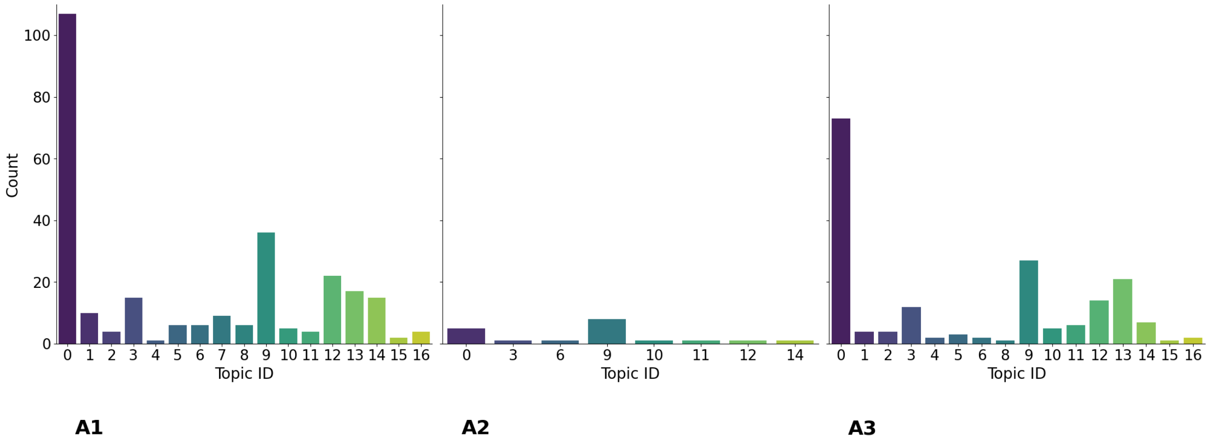

Figure 5: Plots showing the **count of tweets by topic** labeled with the emotion ANGER per annotator (A1, A2, A3).

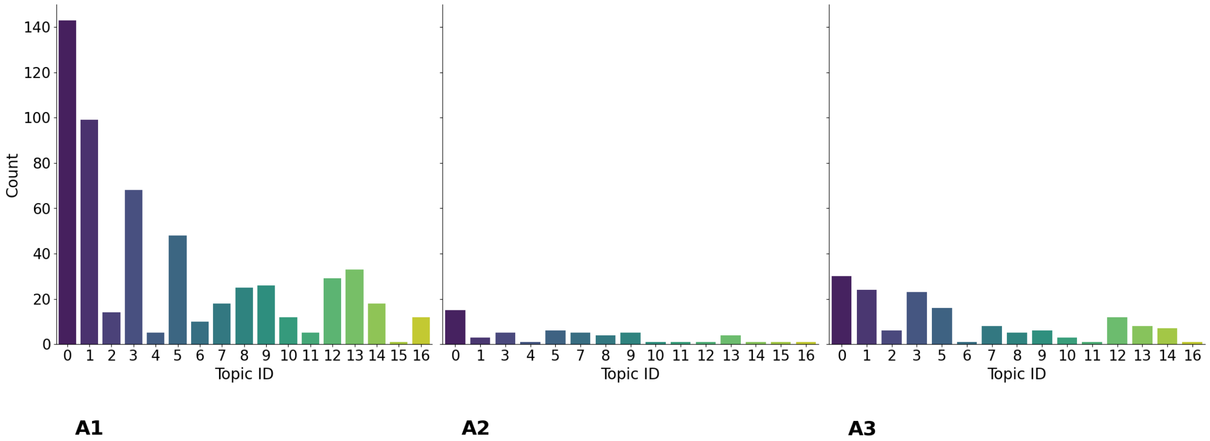

Figure 6: Plots showing the **count of tweets by topic** labeled with the emotion CONCERN per annotator (A1, A2, A3).

B.2 Inter-Annotator Agreement for A1 and A2 by Topics

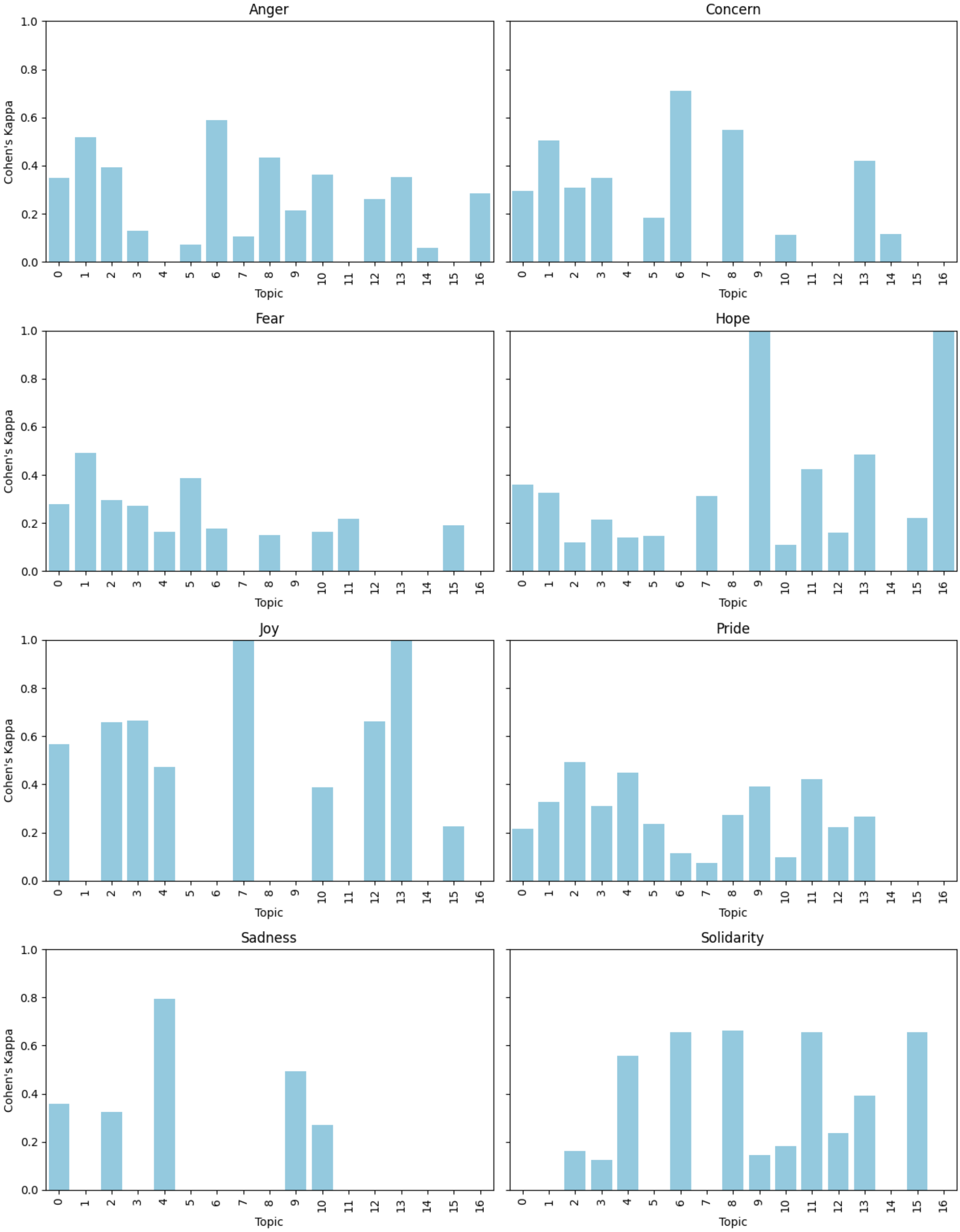

Figure 7: Set of plots showing the calculated **Cohen's Kappa** (Cohen, 1960) **values per topic** for annotator pair A1 and A2.

B.3 Inter-Annotator Agreement for A1 and A3 by Topics

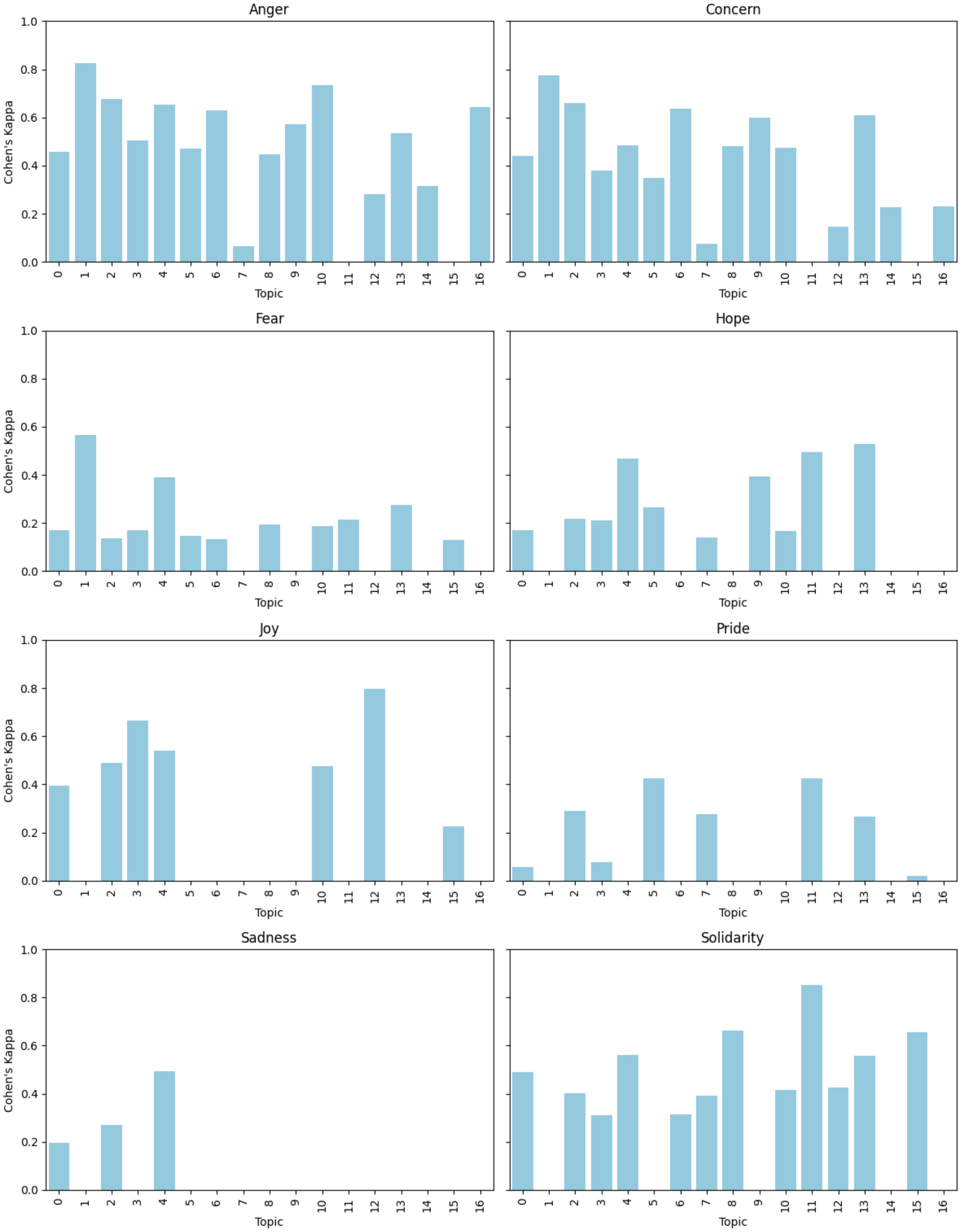

Figure 8: Set of plots showing the calculated **Cohen's Kappa** (Cohen, 1960) **values per topic** for annotator pair A1 and A3.

B.4 Inter-Annotator Agreement for A2 and A3 by Topics

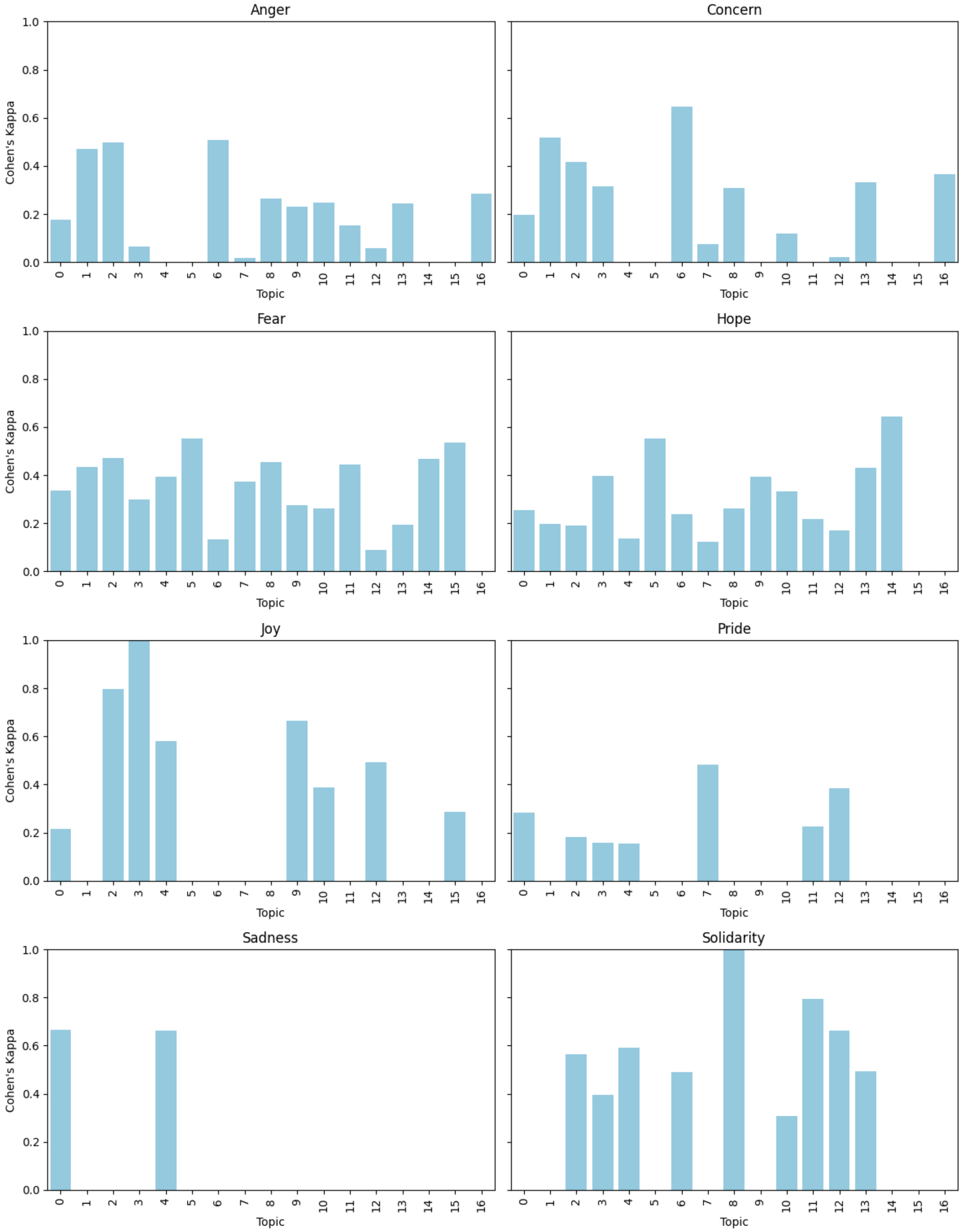

Figure 9: Set of plots showing the calculated **Cohen's Kappa** (Cohen, 1960) **values per topic** for annotator pair A2 and A3.

C Label Spearman Correlation Matrices

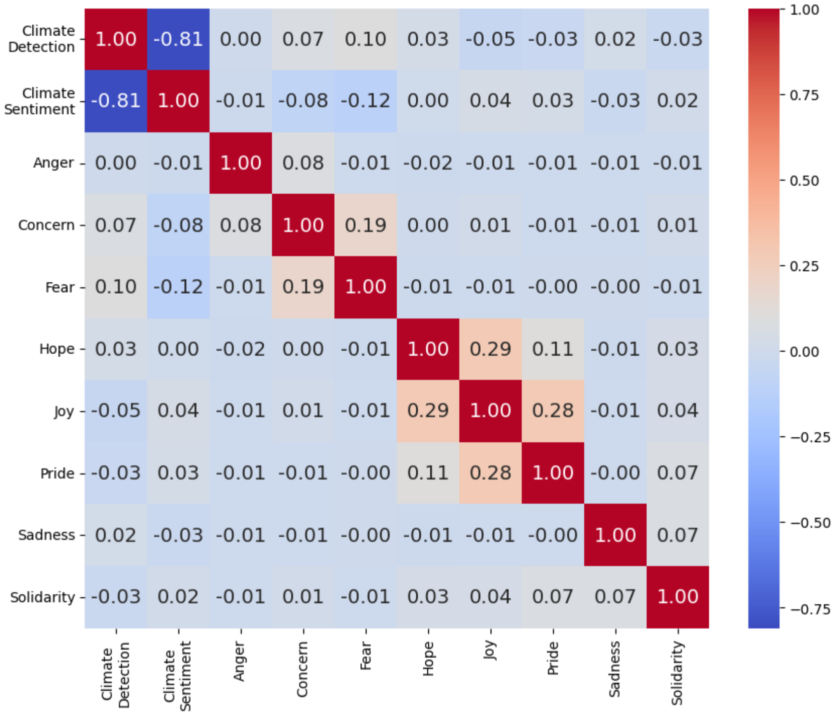

Figure 10: Spearman correlation (Spearman, 1904) matrix of the categories labeled by A2, showing the common occurrences of the labels.

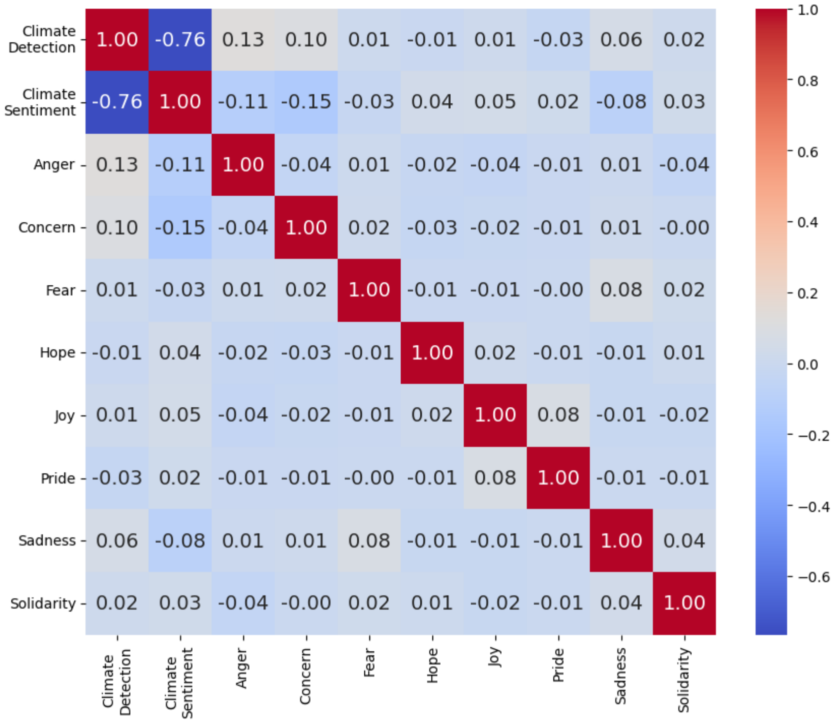

Figure 11: Spearman correlation (Spearman, 1904) matrix of the categories labeled by A3, showing the common occurrences of the labels.