Granular Analysis of Social Media Users' Truthfulness Stances Toward Climate Change Factual Claims

Haiqi Zhang, Zhengyuan Zhu, Zeyu Zhang, Jacob Daniel Devasier, Chengkai Li
University of Texas at Arlington
{haiqi.zhang, zhengyuan.zhu, zeyu.zhang, jacob.devasier}@mavs.uta.edu
cli@uta.edu

Abstract

Climate change poses an urgent global problem that requires efficient data analysis mechanisms to provide insights into climate change-related discussions on social media platforms. This paper presents a framework aimed at understanding social media users' perceptions of various climate change topics and uncovering the insights behind these perceptions. Our framework employs a large language model to develop a taxonomy of factual claims related to climate change and build a classification model that detects the truthfulness stance of tweets toward these factual claims. The findings reveal two key conclusions: (1) The public tends to believe the claims are true, regardless of the actual claim veracity; (2) The public shows a lack of discernment between facts and misinformation across different topics, particularly in areas related to politics, economy, and environment. This highlights the need for targeted attention, critical scrutiny, and informed engagement in these discussion areas.

1 Introduction

Climate change is one of the most pressing global challenges of our time, profoundly impacting the environment, economy, and society. Amidst the urgency to address this global crisis, there is a large volume of discourse on climate change across social media platforms, reflecting growing public awareness and engagement. Understanding and analyzing discourse on climate change is crucial for informing public policy, media strategies, and societal awareness. Prior studies have explored various aspects of text analysis on climate change. Coan et al. (2021) constructed a taxonomy of climate contrarian claims to analyze climate change myths and associated factual claims. Topic modeling performed on tweets by Dahal et al. (2019) showed that discussions of climate change span various topics. Stance detection (Aldayel and Magdy, 2019; Upadhyaya et al., 2023b,a) and sentiment analysis (Jost et al., 2019; El Barachi et al., 2021) have also been widely studied to understand people's beliefs and attitudes toward climate change.

In our study, we streamline a framework that involves collecting factual claims, collecting their corresponding social media posts, constructing an automated taxonomy, and detecting truthfulness stances to understand public perceptions of climate change. Specifically, we collect and analyze factual claims related to climate change and employ the Large Language Model (LLM) with human-in-the-loop to automatically construct a taxonomy of important, fact-checked claims. Beyond the taxonomy, we gather discussions related to these factual claims on social media and perform truthfulness stance detection on these social media posts toward their corresponding factual claims in the taxonomy to examine people's judgments on various climate change-related topics.

Our work enhances the understanding of social media users' perceptions of climate change by: 1) providing a framework to understand people's judgments about climate change-related factual claims across different sub-categories of climate change; 2) yielding several significant insights into people's perceptions of climate change, including the observation that the public lacks discernment between facts and misinformation across different topics. Additionally, our findings reveal that the public tends to believe claims are true, regardless of the actual claim veracity, aligning with the findings of previous research by Moravec et al. (2018).

2 Methodology

In the framework, we first collect factual claims from five credible fact-checking websites using the keywords selected from the Environmental Protection Agency (EPA) topics (Section 2.1). Next, we gather corresponding social media posts using key-

Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024), pages 233–240
August 16, 2024 ©2024 Association for Computational Linguistics

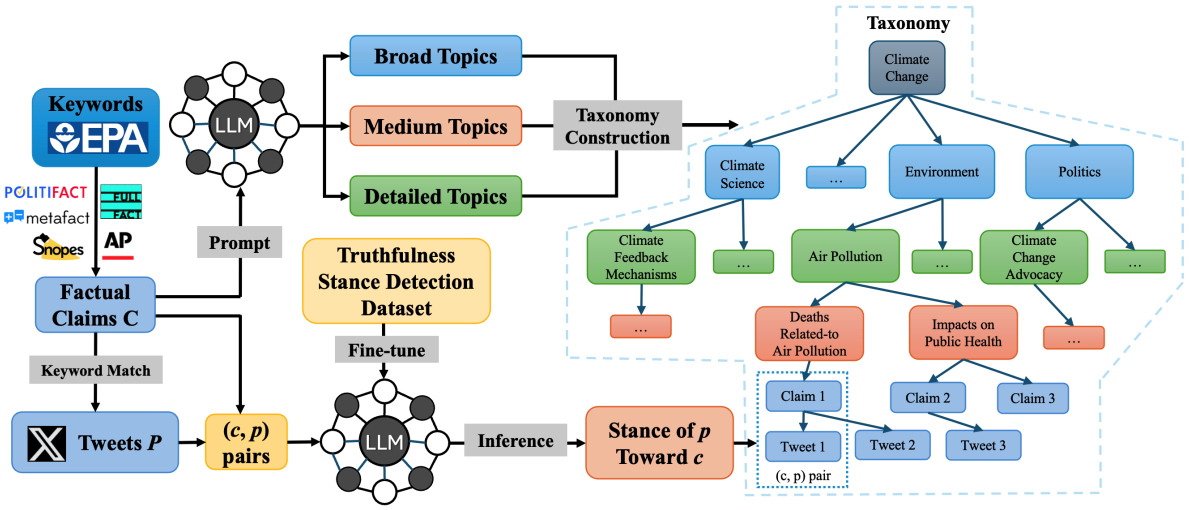

Figure 1: Overview of the framework for analyzing public judgments on climate change-related topics.

words extracted from the collected factual claims (Section 2.2). We then leverage LLM with human-in-the-loop to automatically construct a climate change-related taxonomy (Section 2.3). Finally, we fine-tune a truthfulness stance detection model to assess the truthfulness stances of social media posts toward their corresponding factual claims within the taxonomy (Section 2.4). An overview of the framework is depicted in Figure 1.

2.1 Factual Claim Collection

To identify existing discourse related to climate change, we collect factual claims $\mathcal{C}$ from five fact-checking websites: *PolitiFact*, [1] *Snopes*, [2] *Full Fact*, [3] *Metafact*, [4] and *AP News*. [5] These websites are selected for their popularity and credibility in fact-checking. To collect $\mathcal{C}$, we manually curated a list of climate change-related keywords from the glossary of the Environmental Protection Agency (EPA), such as "global warming" and "greenhouse gas" (full list in Appendix A.1.1). We consider a claim c to be climate change-related if any of the keywords appears in c itself, its fact-checking article's tags (i.e., the topics assigned to the article that categorize its content), or the articles' content. We also collected the verdicts of $\mathcal{C}$ (e.g., "Mostly-true," "False") determined by the fact-checking websites. It is worth noting that the expressions of verdicts vary across different fact-checking websites. Therefore, we categorized them into three unified categories: "Truth," "Uncertain," and "Misinformation" (full verdicts in Appendix A.1.1). After removing duplicates, we obtained $1,409$ unique climate change-related factual claims spanning from November 2007 to May 2024.

[1] politifact.com [2] snopes.com [3] fullfact.org
[4] metafact.io [5] apnews.com

2.2 Tweet Collection

After identifying existing climate change-related factual claims $\mathcal{C}$, we collected corresponding tweets $\mathcal{P}$ discussing those claims to construct (c, p) pairs. This allows us to assess people's judgments of different claims, i.e., whether the tweet c believes the factual claim p is true or false.

To construct (c, p) pairs, we used the tokens extracted from c to collect relevant tweets that discuss each c from X. Specifically, we tokenized and performed part-of-speech tagging for each c using Spacy (ExplosionAI, 2015). We then identified the noun tokens (including proper nouns) in c as token candidates. If fewer than four noun tokens were identified, we added verb tokens to the token candidates. We included adjective tokens if there were still fewer than four token candidates. Claims that resulted in fewer than four tokens after attempting to add verbs and adjectives were disregarded. The final set of tokens formed a search query to collect tweets. In this way, we collected a total of $13,050$ tweets for 729 out of $1,409$ claims. Among these 729 claims, 294 claims had more than 10 tweets.

2.3 Taxonomy Construction

A taxonomy serves as a hierarchical classification structure, organizing topics from broader to more fine-grained levels of granularity. In this framework, we aim to generate a three-level taxonomy from factual claims $\mathcal{C}$ related to climate change. To minimize the manual effort, we prompt LLM, specifically Zephyr (Tunstall et al., 2023), to gen-

erate a set of broad topic, medium topic, and detailed topic, denoted as $\{t^b, t^m, t^d\}$, for each factual claim $c \in \mathcal{C}$. Zephyr is chosen for its competitive performance in language understanding tasks among all 7-billion-parameter LLMs (Chiang et al., 2024). However, the LLM has limitations in consistently producing accurate results based on our initial experiments. For example, the LLM often generates different topics for claims that should be categorized under the same topic. Therefore, we adopt human-in-the-loop to refine the prompt based on the generated topics, enabling multi-round topic generation for optimal results. More specifically, after the LLM generates $\{\hat{t^b}, \hat{t^m}, \hat{t^d}\}$ for all $c \in \mathcal{C}$, humans modify the prompt based on the generated results and then let the LLM generate new topics. This process is repeated until the generated topics are satisfactory.

We start with randomly selecting a subset of claims $\{c_1, c_2, \ldots, c_n\} \subset \mathcal{C}$ ($n = 7$ in our experiments). We manually annotate each c_i with a broad topic t_i^b, a medium topic t_i^m, and a detailed topic t_i^d, as the initial ground truth. These annotated claims and their topics are utilized as learning examples of the prompt for the LLM. Each learning example consists of c_i, all the annotated $\{t^b, t^m, t^d\}$ sets, the question that asks LLM to produce the broad, medium, and detailed topics for c_i, and the answer to the question (i.e., corresponding $\{t_i^b, t_i^m, t_i^d\}$ of c_i). After the LLM learns from the n examples, it is provided with a new claim c_j and asked to generate topics $\{\hat{t_j^b}, \hat{t_j^m}, \hat{t_j^d}\}$ for c_j. Due to the limited context length of LLM, one prompt generates $\{\hat{t_j^b}, \hat{t_j^m}, \hat{t_j^d}\}$ for only one c_j. This generation process is iterated until finishing generating $\{\hat{t_j^b}, \hat{t_j^m}, \hat{t_j^d}\}$ for all $c_j \in \mathcal{C}$. The prompt is detailed in Figure 3 in Appendix B.

After the LLM produces $\{\hat{t_j^b}, \hat{t_j^m}, \hat{t_j^d}\}$ for all $c_j \in \mathcal{C}$, humans scrutinize broad topics that appear frequently (i.e., more than 40 times) and identify the topic sets that contain those frequent broad topics and accurately represent their associated claims. The new topic sets and associated claims are used as new learning examples for the next round of topic generation, continuing until no new frequent broad topics are generated.

2.4 Truthfulness Stance Detection

The task of truthfulness stance detection (Zhu et al., 2022) involves determining the stance of a social media post p toward a factual claim c. The stance can be classified as either believing c is true (*Positive* ($\oplus$)), believing c is false (*Negative* ($\ominus$)), or expressing a neutral stance or no stance toward c (*Neutral/No stance* ($\odot$)). We apply supervised fine-tuning on an LLM to build a classifier, leveraging Zephyr (Tunstall et al., 2023) as the underlying backbone LLM.

An in-house annotated dataset that contains claim-tweet pairs (c, p) and stance labels serves as the ground truth for supervised fine-tuning. The dataset consists of $1,871$ high-quality stance annotations for (c, p) pairs. These pairs were collected using the same method described in Section 2.1 and 2.2, but they are not limited to climate change topics. During the annotation process, annotators provided stance labels for each (c, p) pair. To ensure the dataset's quality, we implemented quality control measures, including screening questions designed to identify low-quality annotators and exclude the annotations from them.

This dataset was chosen because it focuses on p's stance toward c as the target, in contrast to existing datasets where the target is based on topic word (Mohammad et al., 2017, 2016). Additionally, our dataset was annotated with a focus on truthfulness stance toward each factual claim, rather than sentiment stances (Upadhyaya et al., 2023b).

The fine-tuning involves several steps. First, the input (c, p) pair is tokenized using the Byte Pair Encoding (BPE) tokenizer based on SentencePiece (Kudo and Richardson, 2018) and transformed into a dense vector representation. The vector representation is then encoded using the Zephyr encoder and passed through a mean pooler to extract a new vector representation. Finally, the pooled representation is passed through a classification head, consisting of a fully connected layer with a softmax activation layer, to predict the stance. We use cross-entropy as the loss function to update the weight of the classifier. In addition, we apply parameter-efficient fine-tuning using LoRa (Hu et al., 2021), which reduces the number of trainable parameters through low-rank decomposition and speeds up the fine-tuning process.

3 Results

3.1 Results of Climate Change Taxonomy

In our experiments, three rounds of topic generation were conducted. In the first round, 140 broad topics were generated. This was followed by the generation of 111 broad topics in the second round and 98 broad topics in the final round. It is evident

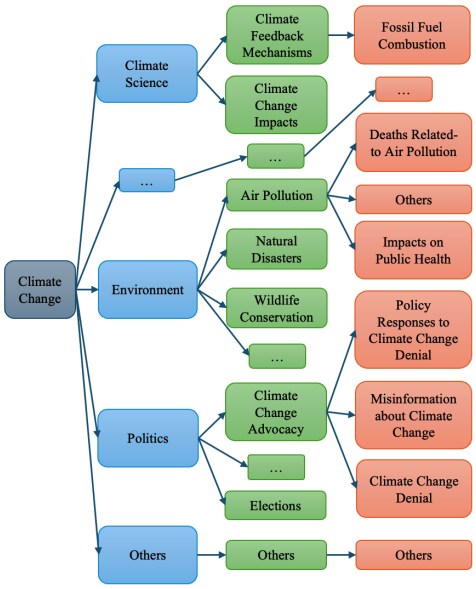

Figure 2: A fragment of climate change taxonomy.

that each successive round produced fewer topics.

In our analysis of the results from the final round, we observed instances where claims that were supposed to belong to the same broad topic were assigned to different topics with subtle differences. For example, some topics had overlapping keywords, e.g., "*Politics, Military*" and "*Politics, Conspiracy Theories*," which could have been merged. These topics usually contained only a couple of claims. To streamline the taxonomy, we merged topics sharing the same initial keyword, as these keywords offered better representation based on our empirical observation, thereby deduplicating the taxonomy. After topic deduplication, certain broad topics were still associated with only a few factual claims. To address this, we grouped such topics into a new broad topic labeled "*Others*." For medium and detailed topics, we retained only those with more than four occurrences, consolidating the rest into the "*Others*" topic within their respective parent topics.

After identifying the topics for each claim, we consolidate the results to construct the taxonomy. Medium topics that share the same broad topic are considered child nodes of that broad topic, and detailed topics are similarly considered child nodes of their respective medium topics. For instance, if one claim has "*Broad topic*: *Environment*; *Medium topic*: *Air Pollution*" and another claim has "*Broad topic*: *Environment*; *Medium topic*: *Natural Disasters*," then "*Air Pllution*" and "*Natural Disasters*" are two child nodes under the broad topic "*Environment*." The final taxonomy comprises 9 broad topics, 33 medium topics, and 13 detailed topics. A subset of the taxonomy is depicted in Figure 2.

To evaluate the produced taxonomy, we randomly selected 100 factual claims from $\mathcal{C}$ and asked two human annotators to categorize them into broad and medium topics based on the taxonomy. Since this is an open-ended problem and a single claim can fit multiple topics, annotators were asked to provide the three most suitable sets of broad and medium topics, including "*Others*." We did not evaluate detailed topics due to the limited number of samples and the specificity, which made them difficult to match accurately. If the generated broad and medium topics appeared in any of the three options provided by the annotators, we considered it correct. The average accuracy of broad topics and medium topics reaches 83% and 62.5%, respectively, indicating the taxonomy is highly effective.

3.2 Results of Truthfulness Stance Detection

	Precision	Recall	Macro F1
$\oplus$	0.863	0.911	0.886
$\odot$	0.783	0.765	0.774
$\ominus$	0.864	0.750	0.803
Avg	0.837	0.808	0.821

Table 1: Performance of truthfulness stance classifier on the annotated dataset. *Positive*, *Neutral/No stance* and *Negative* are denoted as $\oplus$, $\ominus$, $\odot$.

$\oplus$	$\odot$	$\ominus$	**Total**
8,003 (61.33%)	2,668 (20.44%)	2,379 (18.23%)	13,050

Table 2: Truthfulness stance distribution of tweets toward claims.

As shown in Table 1, we assessed the classifier's performance using precision, recall, and macro F1 score on the test set of our truthfulness stance detection dataset, achieving average values of 0.837, 0.808, and 0.821 for precision, recall, and macro F1 score, respectively, indicating robust inference capability. This classifier was applied to collected (c, p) pairs related to climate change. The truthfulness stance distribution of (c, p) pairs in Table 2 reveals that the majority ($8,003$ out of $13,050$ tweets) believe that the claims are true.

In the final results, as indicated in Table 3, each (c, p) pair is associated with a stance, a broad topic,

Claim	Tweet	Stance	Broad Topic	Medium Topic	Detailed Topic
Air pollution linked to greater risk of dementia.	People over 50 in areas with the highest levels of nitrogen oxide in the air showed a 40% greater risk of developing dementia than those with the least NOx #airpollution.	⊕	Health	Air Pollution	Impacts on Brain Health
Sen. Lindsey Graham supports the Green New Deal.	Facebook removed an ad by Adriel Hampton showing Sen. Lindsey Graham backing the Green New Deal.	⊙	Politics	Climate Change Advocacy	Politicians' Stance
The Earth is warming because of the sun's changing distance from the Earth, not because of carbon emissions.	Enough with your pseudo-scientific. Actual science has proven the relationship to human carbon emissions and not cycles of sun /earth distance.	⊖	Climate Science	Climate Feedback Mechanisms	Misconceptions

Table 3: Examples of truthfulness stance detection and their corresponding topics in the taxonomy.

Broad Topic	Truth-⊕	Truth-⊖	Misi-⊕	Misi-⊖	Accuracy	Macro F1
Climate Science	81.7% (524)	18.3% (117)	72.5% (377)	27.5% (143)	0.575	0.524
Economy	70.5% (146)	29.5% (61)	72.5% (351)	27.5% (133)	0.404	0.404
Energy	82.2% (264)	17.8% (57)	74.7% (124)	25.3% (42)	**0.628**	**0.530**
Environment	77.5% (533)	22.5% (155)	74.4% (1040)	25.6% (357)	0.427	0.423
Government Policies	83.2% (183)	16.8% (37)	**69.5%** (205)	**30.5%** (90)	0.530	0.514
Health	**88.7%** (180)	**11.3%** (23)	**77.9%** (169)	**22.1%** (48)	0.543	0.493
Politics	**69%** (363)	**31%** (163)	75.7% (1635)	24.3% (525)	**0.331**	**0.329**
Technology	74.8% (86)	25.2% (29)	69.8% (120)	30.2% (52)	0.481	0.473

Table 4: Stance distribution towards **Truth** and **Misi**information across broad topics. Truth-⊕ and Truth-⊖ denote positive and negative stances towards **Truth**, respectively. Misi-⊕ and Misi-⊖ denote positive and negative stances towards **Misi**information, respectively. Note that the topic "*Others*" is not considered in this analysis.

a medium topic, and a detailed topic. To explore whether social media users can discern true and false claims on various climate change-related topics, we calculated the distribution of positive and negative stances in tweets toward claims with verified verdicts of either true (Truth) or false (Misinformation), as presented in Table 4. We also calculated accuracy to examine how the stances align with the claims' veracity. In addition to accuracy, the macro F1 score was chosen due to the imbalance in the claims' verdicts. We excluded claims from "*Others*" for their small sample size, as well as claims with "Uncertain" verdict and tweets classified as ⊙, as they provide less meaningful insights.

The high percentage of both Truth-⊕ and Misi-⊕ suggests that people tend to believe claims are true regardless of their actual truthfulness. Furthermore, people are more likely to believe claims related to "*Health*," given it has the highest Truth-⊕ (88.7%) and Misi-⊕ (77.9%). The variation in accuracy and macro F1 scores across different topics indicates that people's judgments vary significantly depending on the topics. The low accuracy and macro F1 scores reveal that social media users' judgments of factual claims are not very accurate in the broad topics of "*Politics*" (0.331, 0.329), "*Economy*" (0.404, 0.404), and "*Environment*" (0.427, 0.423) (Table 4), and in the medium topics of "*Elections*" (0.122, 0.117), "*Energy Prices*" (0.221, 0.181), and "*Deforestation*" (0.225, 0.220), as shown in Table 5 in Appendix C. The highest macro F1 score is 0.53 for "*Government Policies*," while most topics' macro F1 score is below 0.5. This suggests that social media users struggle to distinguish between true and false claims. This finding is consistent with the results reported by Moravec et al. (2018) in social science, which suggest that social media users have difficulty detecting fake news and that most users would make more accurate judgments by simply flipping a coin.

4 Conclusion

Our framework provides an effective way to analyze public judgments across multi-level topics related to climate change, aiding in understanding people's perceptions of various climate change topics discussed in online discourse. The results reveal challenges in distinguishing truth from misinformation. More specifically, people tend to accept claims as true, regardless of their accuracy. This issue is particularly evident in discussions on politics, economy, and environment. The findings highlight the need for targeted interventions, such as improved critical thinking education and robust fact-checking, to enhance public discernment and the accuracy of information on social media.

References

Abeer Aldayel and Walid Magdy. 2019. Your stance is exposed! analysing possible factors for stance detection on social media. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *Preprint*, arXiv:2403.04132.

Travis G. Coan, Constantine Boussalis, John Cook, and Mirjam O. Nanko. 2021. Computer-assisted classification of contrarian claims about climate change. *Scientific Reports*, 11(1):22320.

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.

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.

ExplosionAI. 2015. spacy website.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

François Jost, Ann Dale, and Shoshana Schwebel. 2019. How positive is "change" in climate change? a sentiment analysis. *Environmental Science & Policy*, 96:27–36.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 31–41.

Saif M Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)*, 17(3):1–23.

Patricia Moravec, Randall Minas, and Alan R Dennis. 2018. Fake news on social media: People believe what they want to believe when it makes no sense at all. *Kelley School of Business research paper*, (18-87).

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment. *Preprint*, arXiv:2310.16944.

Apoorva Upadhyaya, Marco Fisichella, and Wolfgang Nejdl. 2023a. Intensity-valued emotions help stance detection of climate change twitter data. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 6246–6254. International Joint Conferences on Artificial Intelligence Organization. AI for Good.

Apoorva Upadhyaya, Marco Fisichella, and Wolfgang Nejdl. 2023b. A multi-task model for sentiment aided stance detection of climate change tweets. *Proceedings of the International AAAI Conference on Web and Social Media*, 17(1):854–865.

Zhengyuan Zhu, Zeyu Zhang, Foram Patel, and Chengkai Li. 2022. Detecting stance of tweets toward truthfulness of factual claims. In *Proceedings of the 2022 Computation+Journalism Symposium*.

A Appendix

A.1 Data Collection Details

A.1.1 Key words for collecting factual claims

We curated a list of keywords related to climate change from the glossary of the Environmental Protection Agency (EPA) [6] to collect factual claims from the fact check websites. The keywords include: "*climate change*," "*global warming*," "*greenhouse gas*," "*carbon emission*," "*fossil fuel*," "*ozone*," "*air pollution*," "*carbon dioxide emissions*," "*deforestation*," "*industrial pollution*," "*rising sea levels*," "*extreme weather*," "*melting glaciers*," "*ocean acidification*," "*biodiversity loss*," "*ecosystem disruption*," "*carbon capture*," "*carbon storage*," "*soil carbon*," "*renewable energy*," "*sustainable practices*," "*paris agreement*," "*kyoto protocol*," "*carbon tax*," "*emissions trading schemes*," "*green technology*," "*sustainable technology*," "*environmental change*."

A.1.2 Fact check verdicts and their categories

The verdicts below are categorized into "Truth," "Uncertain," "Misinformation."

- "Truth": True, Correct Attribution, No-Flip, Mostly True, Likely, Near certain.

[6] https://19january2017snapshot.epa.gov/climatechange/glossary-climate-change-terms_.html

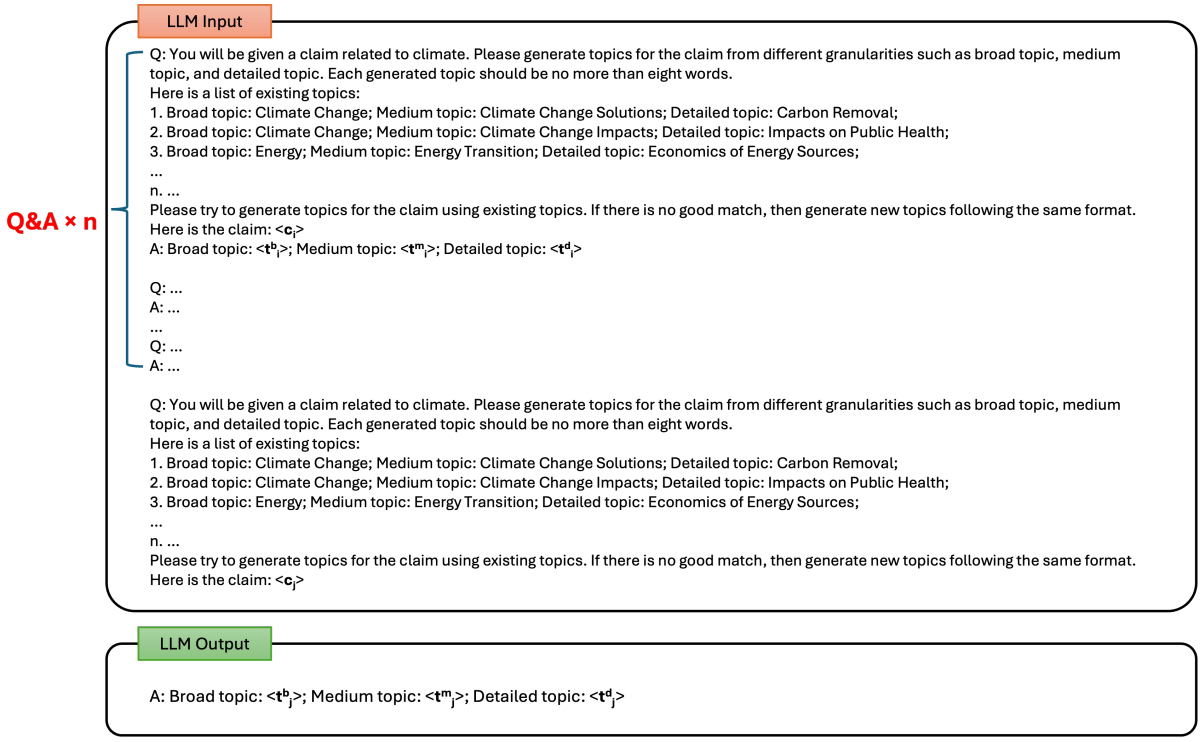

Figure 3: Prompt used to generate topics for each claim

Broad Topic	Medium Topic	Truth-⊕	Truth-⊖	Misi-⊕	Misi-⊖	Accuracy	Macro F1
Politics	Elections	66.7% (2)	33.3% (1)	90.1% (64)	9.9% (7)	0.122	0.117
Environment	Agriculture	50% (2)	50% (2)	85.1% (80)	14.9% (14)	0.163	0.150
Economy	Energy Prices	0 (0)	0 (0)	77.9% (95)	22.1% (27)	0.221	0.181
Environment	Deforestation	60.7% (34)	39.3% (22)	90.1% (154)	9.9% (17)	0.225	0.220
Politics	Others	66.7% (58)	33.3% (29)	75.6% (362)	24.4% (117)	0.309	0.301

Table 5: Examples of relatively inaccurate medium topics in the public's judgments.

Broad Topic	Medium Topic	Truth-⊕	Truth-⊖	Misi-⊕	Misi-⊖	Accuracy	Macro F1
Gov. Policies	Others	94.6% (53)	5.4% (3)	52.9% (9)	47.1% (8)	0.836	0.735
Environment	Energy Policy	100% (3)	0 (0)	23.5% (4)	76.5% (13)	0.800	0.734
Technology	Artificial Intelligence	79.3% (46)	20.7% (12)	46.2% (24)	53.8% (28)	0.673	0.663
Climate Science	Climate Change Impacts	84.5% (49)	15.5% (9)	50% (4)	50% (4)	0.803	0.632
Environment	Climate Change Impacts	92.3% (36)	7.7% (3)	64.8% (35)	35.2% (19)	0.591	0.577

Table 6: Examples of relatively accurate medium topics in the public's judgments.

- "Uncertain": Uncertain, Half True, Research In Progress, Mixture, Unknown, Half-flip, Missing context.
- "Misinformation": False, Pants on Fire, Fake, Full Flop, Labeled Satire, Mostly False, Barely True, False, Unlikely, Extremely Unlikely, Miscaptioned.

B Prompt for Topic Generation

There are n learning examples used to guide the LLM in generating a broad topic, a medium topic, and a detailed topic for each factual claim, as shown in Figure 3. Each prompt example contains a factual claim, a list of topic sets from the n annotated factual claims, considered as "existing topics," a question asking the LLM to generate broad, medium, and detailed topics for the claim, and the answer to the question. In the question, the LLM is instructed to prioritize generating topics from the existing topics. If none of the existing topics align well with the claim, the LLM is then directed to generate new topics. This instruction ensures that the LLM produces a limited number of topics. This prompt is iterated through all the factual claims to generate topics for them.

C Truthfulness Stance Distribution across Medium Topics

Tables 5 and 6 show examples of medium topics where the public's judgments of truth and misinformation are relatively inaccurate and accurate, respectively. In Table 6, medium topics such as "*Others*" under "*Government Policies*," "*Energy Policy*" under "*Environment*," "*Artificial Intelligence*" under "*Technology*," "*Climate Change Impacts*" under "*Climate Science*," and "*Climate Change Impacts*" under "*Environment*" show high accuracy in public judgments with macro F1 scores ranging from 0.577 to 0.735. In contrast, Table 5 presents topics where public judgments are less accurate, indicated by lower Macro F1 scores ranging from 0.117 to 0.301. These topics include "*Elections*" under "*Politics*," "*Agriculture*" under "*Environment*," "*Energy Prices*" under "*Economy*," "*Deforestation*" under "*Environment*," and "*Others*" under "*Politics*."