Thesis Proposal: Understanding Misinformation on Social Media Through Truthfulness Stance

Zhengyuan Zhu
University of Texas at Arlington
zhengyuan.zhu@mavs.uta.edu

Abstract

The rapid spread of misinformation on social media underscores the need for effective tools to analyze public discourse. In this thesis proposal, we focus on the concept of truthfulness stance and propose a truthfulness stance detection framework that leverages large language models (LLMs) with retrieval-augmented generation (RAG) to enhance the contextual understanding of tweets in relation to claims. The framework is evaluated on a newly developed dataset and established benchmark datasets. Experimental results demonstrate that our approach outperforms state-of-the-art methods, significantly improving the Macro-F1 score. To highlight the practical impact of truthfulness stance detection in mitigating misinformation, we showcase several real-world applications and potential future directions.

1 Introduction

Social media platforms serve as vital spaces for discussions on topics such as politics, health, and societal issues; however, they also facilitate the rapid spread of misinformation. Posts on these platforms provide valuable insights into public perceptions and opinions, offering a lens through which societal trends, beliefs, and behaviors can be analyzed (Sobkowicz et al., 2012; Zhang et al., 2018; Willaert et al., 2020). To capture public perceptions and opinions regarding factual claims, this thesis proposal introduces the concept of truthfulness stance(Zhu et al., 2022; Zhang et al., 2024a). Figure1 illustrates examples of tweets expressing positive, neutral, negative, and no stance toward different factual claims. The proposed research on truthfulness stance detection has significant applications across multiple domains. It serves as a crucial tool for analyzing how misinformation spreads (Ecker et al., 2022) and influences decision-making in politics (Ognyanova et al., 2020), public health (Suarez-Lledo and Alvarez-Galvez, 2021) and environmental concern (Zhang et al., 2024a).

Following the concept of truthfulness stance, we introduce a new benchmark dataset specifically designed for training and evaluating truthfulness stance detection models. To address the challenges inherent in this task, we propose a large language model (LLM)-based framework, detailed in Section 4, leveraging retrieval-augmented generation (RAG) to enhance contextual understanding.

Our experimental evaluation conducted on our annotated dataset alongside three established stance detection datasets—SemEval-2019 (Gorrell et al., 2019), WT-WT (Conforti et al., 2020), and COVIDLies (Hossain et al., 2020)—shows that our framework utilizing GPT-3.5 outperforms state-of-the-art models.

We also explore potential applications of truthfulness stance detection, including an ongoing study on a truthfulness stance map for election-related factual claims. Finally, we outline future directions that can further benefit misinformation research and public discourse analysis.

The overarching goal of this thesis proposal is to automate misinformation analysis to assess public perceptions of social media. The specific objectives include:

- Constructing and annotating a large-scale dataset for truthfulness stance detection.
- Developing novel computational methods for stance classification using LLMs.
- Designing and implementing real-world applications that leverage truthfulness stance detection to counter misinformation.

2 Task Definition

Stance, in the field of sociolinguistics, is defined as the speakers taking up positions concerning the expressive, referential, interactional, and social implications of their speech (Jaffe, 2009). In the context of our work, given a factual claim c and a tweet t,

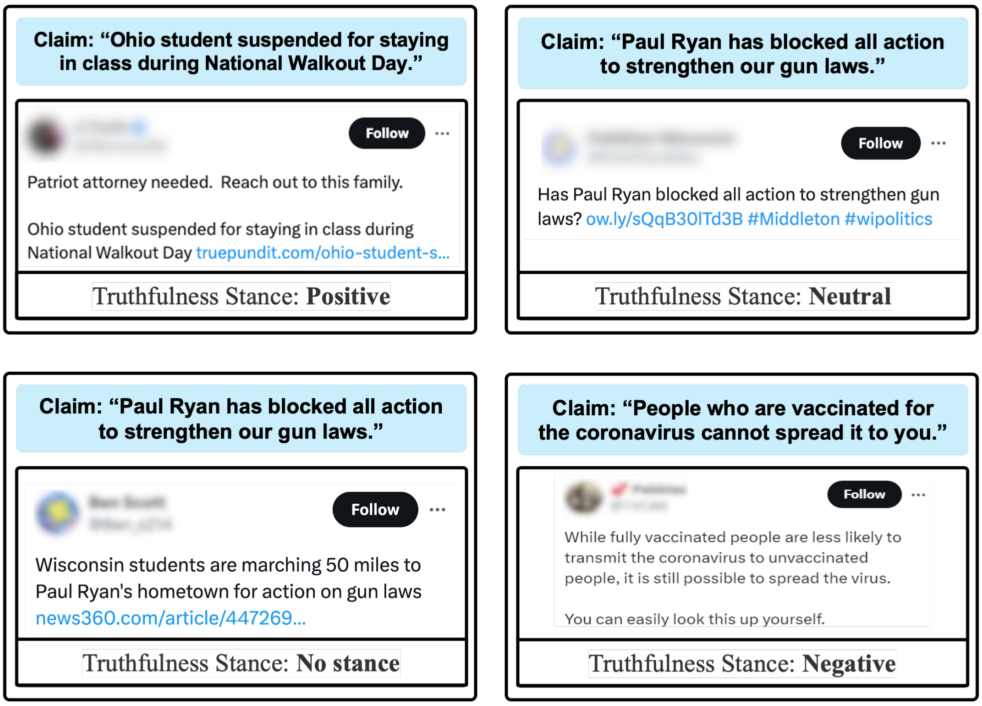

Figure 1: Four tweets expressing different truthfulness stances toward different factual claims.

the task of truthfulness stance detection is to return a classification label $s(c, t)$ from one of three distinct classes — *positive* ($\oplus$), *negative* ($\ominus$), and *neutral/no stance* ($\odot$). The *positive* stance ($\oplus$) is for when t conveys the belief that c is true. Conversely, the *negative* stance ($\ominus$) indicates that t believes c is false. The *neutral/no stance* ($\odot$) signifies that t expresses uncertainty about the truthfulness of c (*neutral*), or t does not express any opinion about c's truthfulness despite both t and c discussing the same topic (*no stance*).

3 Data Collection

3.1 Fact-check Collection

We developed a tool to collect the fact-checks from seven well-known fact-checking websites, including AFP Fact Check, AP Fact Check, FactCheck.org, FullFact, Metafact, PolitiFact, and Snopes. Specifically, we employ *XPath* to navigate through the source HTML code and extract relevant text elements. Specifically, *XPath* expressions allow us to pinpoint specific nodes in the HTML document structure, such as article headlines, claim summaries, and verdicts. To ensure our dataset remains up-to-date with the latest fact-checks, we use *Crontab* to schedule a daily task that fetches the data. The collected data is coded using Claim-Review's data schema, a widely adopted standard for structuring fact-checks.

3.2 Dataset Collection and Annotation

The creation of our truthfulness stance detection dataset involved collecting factual claims and corresponding tweets, followed by annotating claim-tweet pairs. We first collected factual claims from Politifact and then retrieved relevant tweets discussing these claims. We extracted keywords from the claims and used them to retrieve tweets through the Twitter API v2, resulting in 36,154 claim-tweet pairs. To reduce redundancy, we sanitized the dataset by removing similar or duplicate tweets, leaving 2,283 unique claims paired with 5,793 tweets for annotation. Data annotation was performed using an in-house annotation website equipped with annotation quality control measures that can filter out annotations from low-quality annotators to ensure data quality. Claim-tweet pairs were annotated with five stance labels: positive, neutral/no stance, negative, different topics, and problematic.

To identify high-quality annotators, we used 287 carefully selected screening pairs. Five researchers consistently labeled each pair. These pairs were mixed with the pairs that needed real annotation. They were randomly chosen and presented to an annotator without the annotator's knowledge at an average frequency of one in every ten pairs. Annotators were scored based on how well their labels match the experts' labels on the screening pairs. Annotations from low-quality annotators were excluded from the dataset.

A total of 18,584 annotations were collected, 13,594 of which came from high-quality annotators. This resulted in 3,105 completed pairs from high-quality annotators, containing 1,520 unique claims. Of all annotators, 30 out of 206 were classified as high-quality. Notably, of the completed pairs, 216 were labeled as "different topic" and 669 as "problematic." While both categories are included in the released dataset, they are excluded from model training and evaluation.

4 Proposed Methodology

Recent advancements have demonstrated the effectiveness of retrieval-augmented generation (RAG) in knowledge retrieval (Lewis et al., 2020; Wang et al., 2023) and the success of large language models (LLMs) in text analysis (Tang et al., 2023). Building on these advancements, we propose a data augmentation strategy that leverages LLMs for two key purposes: (1) utilizing RAG to retrieve relevant contextual information from external knowledge corpora, thereby mitigating the inherent lack of context in standalone tweets and claims, and (2) synthesizing an analysis of a given tweet t in relation to its associated factual claim c.

4.1 Knowledge Corpora Construction

Two knowledge corpora were constructed for supplying contextual knowledge to other components in our frameowrk, for claims and tweets, respectively. The first knowledge corpus, denoted $\mathcal{D}_C$, encompasses $52,596$ synthesized documents for factual claims. Given a claim c, the corresponding synthesized document d_c was constructed by concatenating excerpts from fact-checks on the claim. The second knowledge corpus, $\mathcal{D}_T$, consists of $8,236$ synthesized documents for tweets posted from 2010 to 2023.

4.2 Contextual Knowledge Generation

The framework generates contextual knowledge in the form of two documents, e_c and e_t, corresponding to a claim c and a tweet t, respectively, within a given claim-tweet pair. These contextual documents play a crucial role in ensuring accurate truthfulness stance detection. The generation process consists of three interrelated steps: relevant document selection, relevant chunk retrieval, and prompting the LLM.

For selecting relevant documents, a keyword-based approach was used to retrieve pertinent texts from the claim knowledge corpus $\mathcal{D}_C$. Nouns, verbs, and adjectives were extracted from each claim, and Jaccard similarity was computed between the extracted terms and the documents in the corpus. Based on similarity scores, the ten most relevant documents were selected for each claim. A similar procedure was followed for tweets, where the ten most similar documents were identified from the tweet knowledge corpus $\mathcal{D}_T$.

Once the relevant documents were identified, they were segmented into smaller textual chunks, each consisting of 512 tokens, to facilitate efficient retrieval of highly relevant information. The segmentation was followed by an embedding-based retrieval process using the BAAI General Embedding (BGE) model (Xiao et al., 2023). Cosine similarity was applied to measure the alignment between each chunk and the query, allowing the system to retrieve the top ten most relevant chunks. The prompt used for retrieval was designed to be consistent with the prompt instruction later used in the LLM generation step. To generate the contextual knowledge documents e_c and e_t, the LLM was prompted with structured input that included both the claim and the tweet, along with the retrieved relevant chunks. The prompt provided explicit instructions to the model to synthesize contextual information that enhances the factual grounding of the claim-tweet relationship.

Building on the contextual knowledge obtained in the previous steps, the framework generates a stance analysis that captures the truthfulness of a tweet's stance toward a claim. An LLM is prompted with the claim, the tweet, and their corresponding contextual knowledge documents, e_c and e_t, to produce a detailed narrative assessing the tweet's stance. The output of this process is denoted as a.

4.3 Classification Model

Our framework produces the final stance label by using a fine-tuned LLM as a classifier. Given a claim-tweet pair (c, t) as well as the corresponding a, e_c and e_t generated by other components described earlier, the LLM converts the i-th input into a vector representation. The vector is fed into a single fully connected layer and a softmax layer to produce the probability distribution of stance orientation labels. The model parameters are fine-tuned during training and optimized using the Adam optimizer (Kingma and Ba, 2015).

5 Evaluation

We applied our framework to three widely used benchmark datasets—SemEval-2019 (Gorrell et al., 2019), WT-WT (Conforti et al., 2020), and COVIDLies (Hossain et al., 2020)—for performance comparison, along with our own dataset. However, since the stance and class categories vary in definition and naming across these datasets, we standardized the labels by merging and renaming them to ensure a fair comparison of model performance. We evaluated the performance of two types of stance detection models: LM-based and LLM-based. Consistent with previous studies, we used F1 scores for each class—denoted as $F_{\oplus}$, $F_{\odot}$, and $F_{\ominus}$—and the Macro F1 score (F_M) as our evaluation metrics. We evaluated the performance of our framework by comparing it to several state-of-the-art stance detection models. In our framework, we utilize two fine-tuned LLMs: the open-source model Zephyr (Tunstall et al., 2023) and the proprietary model GPT-3.5. Our framework demonstrates strong performance across all datasets compared to other stance detection models. Our framework based on GPT-3.5 achieves the highest scores across all metrics on our dataset.

6 Real-World Applications

Truthfulness stance detection has the potential to inspire a new line of research. Its key applications fall into two primary categories: dashboard systems and social analytics.

Dashboards for Misinformation Monitoring. Dashboards provide an effective means of visualizing stance trends and identifying misinformation hotspots. For example, Zhu et al. (2021) developed a geolocation-based dashboard during the COVID-19 pandemic to track public stance toward health-related claims sourced from trusted organizations such as the WHO and CDC. Such a system helps health organizations identify communities vulnerable to misinformation, enabling targeted interventions.

A similar approach can be applied to election-related misinformation using our proposed framework, which analyzes factual claims and stance patterns in election-related tweets. Figure 2 illustrates a truthfulness stance map, an interactive interface that visually represents the distribution of truthfulness stances toward election-related factual claims across the United States. Developed using Streamlit (Streamlit, 2019), the map allows users to select claims from a sidebar and explore public stance distributions.

This tool benefits multiple stakeholders. Political strategists can use it to assess public reactions to their candidates' statements and refine campaign strategies (Dwivedi et al., 2021) and track public sentiment toward political figures over time (Dimitrova and Matthes, 2018). Social scientists can analyze the spread and impact of conspiracy theories, testing hypotheses on election-related misinformation. By integrating granular geolocation data with an intuitive, user-friendly interface, this work offers a valuable resource for understanding and addressing misinformation in the context of the 2024 election.

Social Analytics. The truthfulness stance can help social scientists analyze societal beliefs. For example, Zhang et al. (2024a) proposed a framework to address the urgent global challenge of understanding public perceptions of climate change topics on social media. This framework utilizes an LLM to construct a taxonomy of factual claims related to climate change and develop a truthfulness stance detection model to classify the stance of tweets toward these claims. Findings reveal that the public generally believes claims to be true, regardless of their actual veracity, and that there is a notable lack of discernment between facts and misinformation, particularly in discussions related to politics, economics, and the environment. These insights emphasize the need for greater critical scrutiny and targeted attention in climate change discourse.

Truthfulness stance detection can also be integrated into social analytics platforms. The proposed model can be deployed as an application programming interface (API), enabling incorporation into platforms such as open-source social sensing systems (Zhang et al., 2024b). This integration allows researchers, policymakers, and analysts to monitor stance patterns in real-time, enhancing their ability to detect and respond to misinformation. Embedding truthfulness stance detection into these systems contributes to misinformation tracking, public discourse analysis, and data-driven decision-making across various fields.

7 Future Directions

While our proposed framework demonstrates strong performance in truthfulness stance detec-

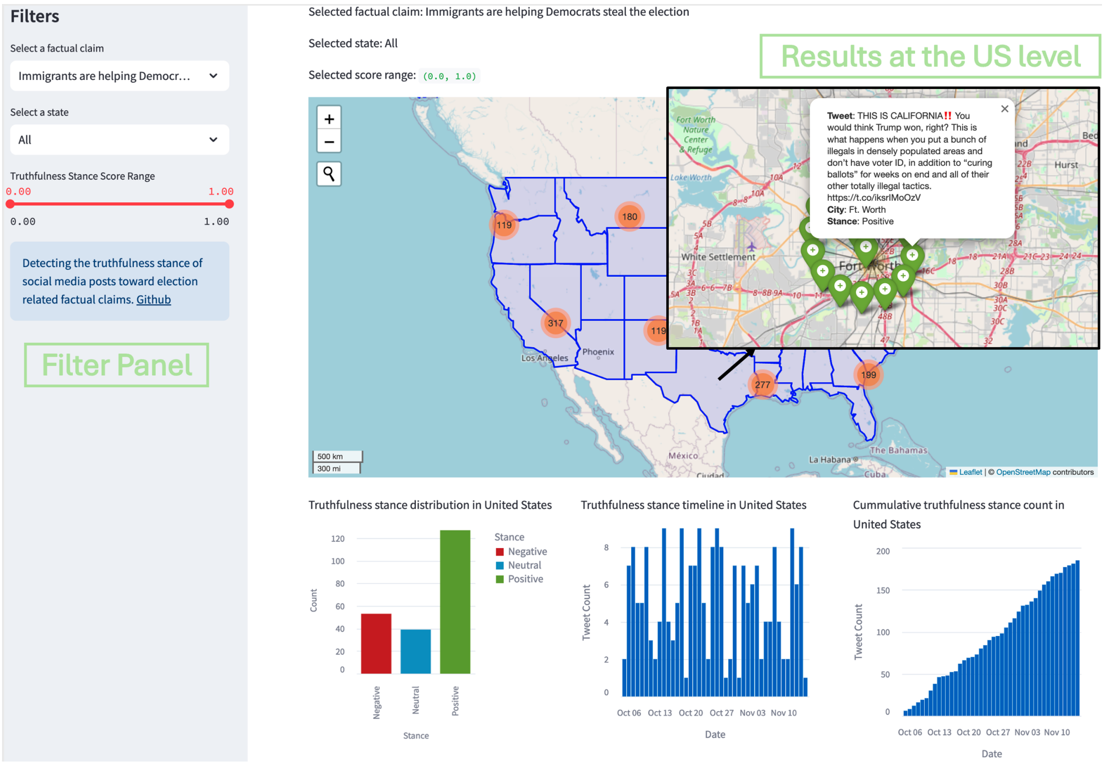

Figure 2: A truthfulness stance map for election-related misinformation.

tion, several promising directions remain for future research to enhance its capabilities further.

Multimodal Truthfulness Stance Detection. Currently, our framework processes textual claims and tweets in isolation. However, social media posts frequently contain multimodal content, including images, videos, and hyperlinks, which can provide critical context for understanding stance. Future research could explore the integration of multimodal learning techniques, leveraging vision-language models such as CLIP (Radford et al., 2021) and BLIP (Li et al., 2022) to capture both visual and textual cues. Incorporating multimodal features may help disambiguate subtle stance expressions, particularly in cases involving sarcasm, memes, or manipulated media.

Handling Sarcasm and Irrelevant Claim-Tweet Pairs. The current framework primarily focuses on explicit stance expressions. However, sarcasm and indirect language can obscure the intended meaning of tweets, making stance detection more challenging. Additionally, our annotation process excluded claim-tweet pairs deemed irrelevant. Future research should investigate methods to detect and address sarcasm, irony, and other forms of implicit stance expression. Techniques such as contrastive learning and contextual embeddings from conversation history may enhance robustness against these linguistic challenges.

Incorporating Conversational Context. Truthfulness stance detection often requires considering broader conversational context, as a single tweet may be part of an ongoing discussion thread. Future research could explore incorporating conversational history to capture evolving stance shifts. Understanding how stance evolves across multiple turns in a conversation can provide deeper insights into misinformation propagation and belief reinforcement.

Addressing Multiple Claims Within a Tweet. Some tweets contain multiple factual claims, each of which may require separate stance evaluations. The current framework assumes a one-to-one correspondence between tweets and claims, which limits its applicability in cases where users express opinions on multiple claims simultaneously. Future studies could explore methods for claim segmentation and multi-label stance classification, enabling a more granular analysis of complex tweets.

References

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020. Will-they-won't-they: A very large dataset for stance detection on twitter. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1724.

Daniela V Dimitrova and Jörg Matthes. 2018. Social media in political campaigning around the world: Theoretical and methodological challenges.

Yogesh K Dwivedi, Elvira Ismagilova, D Laurie Hughes, Jamie Carlson, Raffaele Filieri, Jenna Jacobson, Varsha Jain, Heikki Karjaluoto, Hajer Kefi, Anjala S Krishen, et al. 2021. Setting the future of digital and social media marketing research: Perspectives and research propositions. *International journal of information management*, 59:102168.

Ullrich KH Ecker, Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K Fazio, Nadia Brashier, Panayiota Kendeou, Emily K Vraga, and Michelle A Amazeen. 2022. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1):13–29.

Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854.

Tamanna Hossain, Robert L. Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. 2020. COVIDLies: Detecting COVID-19 misinformation on social media. In *Proceedings of the 1st Workshop on NLP for COVID-19*.

Alexandra Jaffe. 2009. *Stance: sociolinguistic perspectives*. Oup Usa.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Katherine Ognyanova, David Lazer, Ronald E Robertson, and Christo Wilson. 2020. Misinformation in action: Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power. *Harvard Kennedy School Misinformation Review*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Pawel Sobkowicz, Michael Kaschesky, and Guillaume Bouchard. 2012. Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web. *Government Information Quarterly*, 29(4):470–479.

Inc. Streamlit. 2019. Streamlit: A faster way to build and share data apps. https://streamlit.io/. Accessed: 2024-11-20.

Victor Suarez-Lledo and Javier Alvarez-Galvez. 2021. Prevalence of health misinformation on social media: systematic review. *Journal of Medical Internet Research*, 23(1):e17187.

Tianyi Tang, Hongyuan Lu, Yuchen Eleanor Jiang, Haoyang Huang, Dongdong Zhang, Wayne Xin Zhao, and Furu Wei. 2023. Not all metrics are guilty: Improving nlg evaluation with LLM paraphrasing. *arXiv preprint arXiv:2305.15067*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, et al. 2023. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv preprint arXiv:2310.07521*.

Tom Willaert, Paul Van Eecke, Katrien Beuls, and Luc Steels. 2020. Building social media observatories for monitoring online opinion dynamics. *Social Media+ Society*, 6(2):2056305119898778.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighof. 2023. C-pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597*.

Daniel Yue Zhang, Jose Badilla, Yang Zhang, and Dong Wang. 2018. Towards reliable missing truth discovery in online social media sensing applications. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 143–150.

Haiqi Zhang, Zhengyuan Zhu, Zeyu Zhang, Jacob Devasier, and Chengkai Li. 2024a. Granular analysis of social media users' truthfulness stances toward climate change factual claims. In *Proceedings of the*

1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024), pages 233–240, Bangkok, Thailand. Association for Computational Linguistics.

Zeyu Zhang, Zhengyuan Zhu, Haiqi Zhang, Foram Patel, Josue Caraballo, Patrick Hennecke, and Chengkai Li. 2024b. Wildfire: A twitter social sensing platform for layperson. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 1106–1109.

Zhengyuan Zhu, Kevin Meng, Josue Caraballo, Israa Jaradat, Xiao Shi, Zeyu Zhang, Farahnaz Akrami, Haojin Liao, Fatma Arslan, Damian Jimenez, Mohammed Samiul Saeef, Paras Pathak, and Chengkai Li. 2021. A dashboard for mitigating the COVID-19 misinfodemic. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 99–105, Online. Association for Computational Linguistics.

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*.