TrustMap: Mapping Truthfulness Stance of Social Media Posts on Factual Claims for Geographical Analysis

Zhengyuan Zhu
zhengyuan.zhu@mavs.uta.edu
The University of Texas at Arlington
Arlington, TX, USA

Haiqi Zhang
haiqi.zhang@mavs.uta.edu
The University of Texas at Arlington
Arlington, TX, USA

Zeyu Zhang
zeyu.zhang@mavs.uta.edu
The University of Texas at Arlington
Arlington, TX, USA

Chengkai Li
cli@uta.edu
The University of Texas at Arlington
Arlington, TX, USA

Abstract

Factual claims and misinformation circulate widely on social media, shaping public opinion and decision-making. The concept of *truthfulness stance* refers to whether a text affirms a claim as true, rejects it as false, or takes no clear position. Capturing such stances is essential for understanding how the public engages with and propagates misinformation. We present the **tru**thfulness **st**ance **map** (TrustMap), an application that identifies and visualizes stances of tweets toward factual claims. Users may input factual claims or select claims from a curated set. For each claim, TrustMap retrieves relevant social media posts and applies a retrieval-augmented approach with fine-tuned language models to classify stance. Posts are classified as positive, negative, or neutral/no stance. These classifications are then aggregated by location to reveal regional variations in public opinion. To enhance interpretability, TrustMap uses large language models to generate stance explanations for individual posts and to produce regional stance summaries. By integrating retrieval-augmented truthfulness stance detection with geographical visualization, TrustMap provides the first tool of its kind for exploring how belief in factual claims varies across regions.

CCS Concepts

• **Computing methodologies → Natural language processing**; • **Information systems → Geographic information systems**; **Social networks**; • **Applied computing → Sociology**.

Keywords

Truthfulness Stance, Factual Claim, Social Media, Large Language Model

ACM Reference Format:
Zhengyuan Zhu, Haiqi Zhang, Zeyu Zhang, and Chengkai Li. 2025. TrustMap: Mapping Truthfulness Stance of Social Media Posts on Factual Claims for Geographical Analysis. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3746252.3761490

CIKM '25, Seoul, Republic of Korea

ACM ISBN 979-8-4007-2040-6/2025/11
https://doi.org/10.1145/3746252.3761490

1 Introduction

In recent years, the dissemination of factual claims, public narratives, and misinformation has intensified across domains such as health [21], environment [22], and politics [23]. Social media platforms such as X (formerly Twitter) play a central role in this ecosystem, not only enabling rapid circulation of content [5, 19, 30] but also accelerating the spread of misleading claims and conspiracy theories, which can shape public opinion and decision-making [2, 4, 26]. Social media users frequently react to factual claims by endorsing their accuracy, disputing their validity, or expressing uncertainty. Understanding these responses, referred to as *truthfulness stances*, is essential for analyzing how misinformation and public messaging influence discourse across topics. While prior research in stance detection has examined specific domains such as health misinformation [10], news claims [18], and other controversial issues [28], relatively little work has focused on truthfulness stance toward general factual claims spanning diverse topics [33, 34]. Even fewer studies have incorporated geolocation-based analyses, which are crucial for understanding how perceptions of claim veracity vary across communities and how narratives resonate at regional levels.

This paper introduces the **tru**thfulness **st**ance **map** (TrustMap), accessible at https://idir.uta.edu/trustmap. TrustMap provides a visual representation of stance distributions toward factual claims across geographical regions in the United States. It aggregates social media users' perceptions of claim veracity into three categories—*positive*, *negative*, and *neutral/no stance*. A positive stance indicates support for a claim's truthfulness, a negative stance denotes that the claim is disputed as false, and a neutral/no stance reflects either uncertainty or the absence of an explicit endorsement or refutation.

We collected 136,040 tweets related to 2,216 distinct claims for static exploration, of which 24,262 tweets contain geolocation information. Each claim-tweet pair was classified using RATSD (Retrieval Augmented Truthfulness Stance Detection) [34], a model that leverages fine-tuned large language models (LLMs) with retrieval-augmented generation (RAG) [13]. In TrustMap, users can either explore predefined topics and their associated claims or enter a custom claim for real-time analysis. For real-time queries, the system retrieves relevant tweets via X's API and applies the RATSD model to classify their stances. The results can then be explored on an interactive U.S. map, complemented by stance distribution visualizations, LLM-generated explanations for individual tweet classifications, and regional summary reports. Analysis of the collected claim-tweet pairs reveals that social media users frequently endorse claims as true, irrespective of their actual veracity. At the geographic level, users in certain states show greater difficulty in distinguishing true claims from false ones.

While our prior work [28] analyzed truthfulness stances in the climate change domain, it did not provide a user interface. TrustMap builds on this foundation by extending coverage to multiple domains and enabling real-time claim exploration. To the best of our knowledge, it is the first application that integrates truthfulness stance detection with interactive geographical analytics. Beyond its technical contributions, TrustMap provides practical value: it helps fact-checkers, journalists, and policymakers identify regions most vulnerable to misinformation, supports monitoring of emerging claims, and offers the research community a reproducible tool for studying geographical variation in truthfulness stance. The codebase and data are available at https://github.com/idirlab/trustmap, with a video demonstration at https://vimeo.com/1094263767.

2 Related Work

Although numerous online narrative monitoring and analytics applications exist, most emphasize broad data collection and general analytics [6, 31] or concentrate on popular tasks such as sentiment analysis [1, 14]. In contrast, the incorporation of truthfulness stance into interactive maps or dashboards remains relatively unexplored. Prior systems typically address stance types different from those examined in our work. For example, StanceVis Prime [12] supports the analysis and visualization of sentiment and stance in temporal social media data. It processes documents from multiple text streams and applies sentiment and stance classification, generating data series linked to the source texts. However, its stance detection relies on identifying seven linguistic modifiers [20], which differs fundamentally from our definition of truthfulness stance. Similarly, Liew et al. [15] developed a dashboard for monitoring and analyzing online narratives, but it focuses on *sentiment*-oriented stance toward *general* topics (e.g., vaccine side effects), in contrast to our emphasis on *truthfulness*-oriented stance toward *specific* claims.

Existing applications of truthfulness stance detection primarily stem from our previous work. We first developed a dashboard for the COVID-19 misinfodemic [32], which identified stances toward COVID-19-related claims. However, stance detection was not the system's primary focus, and its underlying method was less advanced than RATSD. In contrast, TrustMap employs the RATSD model and extends coverage to a broader set of factual claims, including both verified facts and misinformation across multiple topics beyond COVID-19. We also created a framework to analyze social media users' truthfulness stances toward climate change-related claims [28], and more recently developed the novel truthfulness stance detection methodology RATSD [34]. Together, these earlier efforts form the foundation on which TrustMap is built.

3 Design and Implementation of TrustMap

TrustMap is built on a three-stage data pipeline: (1) data ingestion, (2) truthfulness stance classification, and (3) data exploration. In the data ingestion stage, factual claims were collected from PolitiFact (https://www.politifact.com/), and claim-tweet pairs were generated by querying X's APIs to retrieve tweets related to each claim. In the classification stage, the RATSD model assigned each claim-tweet pair one of three stance classes—positive, negative, or neutral/no stance—to capture the stance of the tweet regarding the corresponding claim's veracity. For the exploration stage, the labeled pairs were aggregated by tweet geolocation, enabling users to interactively examine stance distributions and patterns through the interface.

3.1 Data Ingestion

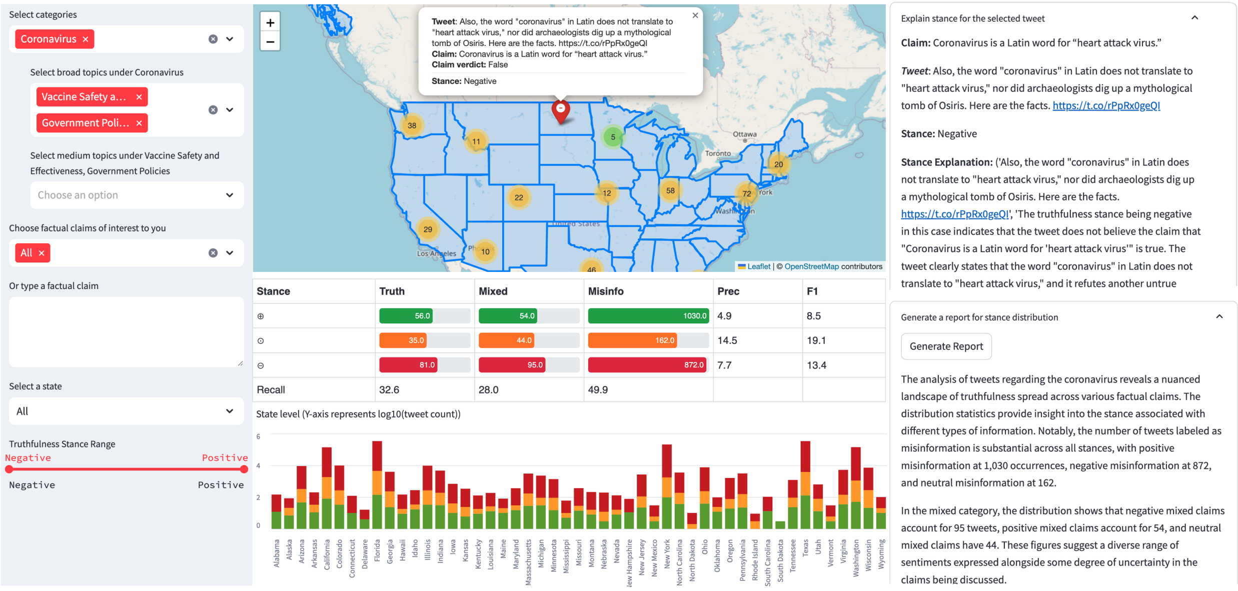
We collected factual claims from PolitiFact using an in-house fact-check collection tool. PolitiFact was chosen because it offers the most comprehensive coverage, with 25,694 claims spanning a wide range of topics. For each claim, we retrieved tweets by constructing queries to X's API (https://docs.x.com/x-api/posts/search/integrate/build-a-query) using a keyword-based strategy that extracted query terms—typically nouns, verbs, adjectives, and numbers—from the claim text. Tweets were collected within a defined time window extending from one month before to one year after the publication date of the fact-check corresponding to each claim. Tweets shorter than 30 characters were excluded to reduce noise.

We used two strategies to collect geolocation information from tweets, which is key to rendering claim-tweet pairs on TrustMap. The first strategy was to include location operators (e.g., latitude, longitude, and radius) in the query (e.g., "[keywords] geocode:[latitude], [longitude], [radius]"). For this, we selected 78 representative cities across the United States and used their latitude and longitude in the queries. However, this approach yielded limited data, as X users rarely disclose their location in tweets. The second strategy was to retrieve tweets and their metadata, then extract geolocation information from X users' profile descriptions. All retrieved geolocation strings were normalized using Geopy [9], which converts unstructured text into structured fields such as city, county, state, and country. Among the 136,040 claim-related tweets collected, 24,262 contained U.S. geolocation information obtained using the two strategies described above. All tweets, regardless of geolocation, were included in stance classification and aggregate analyses (Table 2). However, only tweets with U.S. geolocation were used in the TrustMap interactive map and in regional comparisons (Table 3). It is important to note that geolocated tweets represent only a subset of all posts and may introduce demographic bias, since users who disclose or enable location information can differ systematically from the broader population.

3.2 Truthfulness Stance Detection Model

We applied RATSD [34] to classify the stance of each tweet toward its associated factual claim. RATSD incorporates contextual knowledge to improve classification accuracy. This enhancement is particularly important because 1) both claims and tweets are often standalone sentences that lack sufficient context for informed classification, and 2) tweets frequently contain acronyms, hashtags, and slang that complicate interpretation.

RATSD enhances stance detection through two components: *context retrieval* and *stance analysis generation.* For each claim-tweet pair, the model first retrieves relevant contextual information using RAG. Contextual documents are constructed separately: fact-check articles are used for claims, while X's metadata (e.g., user profiles) is used for tweets. During training and inference, the model selects



Select categories: Coronavirus ×

Select broad topics under Coronavirus: Vaccine Safety a... × Government Poli... ×

Select medium topics under Vaccine Safety and Effectiveness, Government Policies: Choose an option

Choose factual claims of interest to you: All ×

Or type a factual claim

Select a state: All

Truthfulness Stance Range: Negative – Positive

Tweet: Also, the word "coronavirus" in Latin does not translate to "heart attack virus," nor did archaeologists dig up a mythological tomb of Osiris. Here are the facts. https://t.co/rPpRx0geQI
Claim: Coronavirus is a Latin word for “heart attack virus.”
Claim verdict: False

Stance: Negative

Stance	Truth	Mixed	Misinfo	Prec	F1
⊕	56.0	54.0	1030.0	4.9	8.5
⊙	35.0	44.0	162.0	14.5	19.1
⊖	81.0	95.0	872.0	7.7	13.4
Recall	32.6	28.0	49.9		

State level (Y-axis represents log10(tweet count))

Explain stance for the selected tweet

Claim: Coronavirus is a Latin word for “heart attack virus.”

Tweet: Also, the word "coronavirus" in Latin does not translate to "heart attack virus," nor did archaeologists dig up a mythological tomb of Osiris. Here are the facts. https://t.co/rPpRx0geQI

Stance: Negative

Stance Explanation: ('Also, the word "coronavirus" in Latin does not translate to "heart attack virus," nor did archaeologists dig up a mythological tomb of Osiris. Here are the facts. https://t.co/rPpRx0geQI', 'The truthfulness stance being negative in this case indicates that the tweet does not believe the claim that "Coronavirus is a Latin word for 'heart attack virus'" is true. The tweet clearly states that the word "coronavirus" in Latin does not translate to "heart attack virus," and it refutes another untrue

Generate a report for stance distribution

Generate Report

The analysis of tweets regarding the coronavirus reveals a nuanced landscape of truthfulness spread across various factual claims. The distribution statistics provide insight into the stance associated with different types of information. Notably, the number of tweets labeled as misinformation is substantial across all stances, with positive misinformation at 1,030 occurrences, negative misinformation at 872, and neutral misinformation at 162.

In the mixed category, the distribution shows that negative mixed claims account for 95 tweets, positive mixed claims account for 54, and neutral mixed claims have 44. These figures suggest a diverse range of sentiments expressed alongside some degree of uncertainty in the claims being discussed.

Figure 1: The user interface of TrustMap.

documents relevant to the input claim and tweet, retrieves semantically aligned chunks using BGE embeddings [25], and prompts an LLM to summarize them into coherent context.

Next, RATSD prompts an LLM to generate a stance analysis, which is a natural language explanation of the tweet’s stance with respect to the claim. This generated analysis replaces the original tweet in the classifier’s input, enabling the model to learn from structured, context-rich representations rather than noisy, raw social media text.

Finally, RATSD employs a fine-tuned LLM that encodes an input sequence consisting of the claim, tweet, stance analysis, and contextual knowledge. The model produces a stance prediction by computing a probability distribution over the stance classes through a softmax layer. Training was performed using cross-entropy loss with L2 regularization, and parameters were optimized with the Adam optimizer [11].

3.3 Data Exploration in TrustMap

The truthfulness stance results are presented through an interactive user interface (Figure 1) that enables filtering and exploration across regions, topics, and claims. Built with Streamlit (https://streamlit.io/), the interface consists of a control panel, an interactive map, statistical charts, and panels with LLM-generated explanations.

The control panel on the left allows users to select claims within one or more broad topics. For topics with many claims, users can refine their selection by choosing medium- or detailed-level subtopics. The multi-level claim topic taxonomy is generated using the LLM-Taxo method [29]. Once topics are selected, a multi-select dropdown is populated with the corresponding factual claims, enabling users to choose specific claims for exploration. In addition to static claims exploration, TrustMap supports real-time analysis: users can manually enter a factual claim in a text input field. TrustMap then automatically formulates a query to X’s API, retrieves relevant tweets, and applies RATSD to assess each tweet’s stance toward the entered claim. This feature enables users to explore new or emerging claims. Users can further narrow the exploration results by focusing on specific regions, either by selecting a U.S. state from a dropdown menu or by clicking directly on the map. Upon selection, the map will zoom to the chosen state. A slider at the bottom provides additional filtering, allowing users to view tweets by stance class.

On the map, claim-tweet pairs are clustered by geographic location to reduce visual clutter. As users zoom in, these clusters break apart into individual markers. Clicking a marker opens a pop-up that displays the full text of both the claim and the associated tweet, along with metadata such as the tweet’s location, the claim’s veracity rating from PolitiFact (e.g., “True,” “False,” “Pants on Fire”), and the predicted truthfulness stance of the tweet regarding the claim.

Beneath the map, two interactive charts provide aggregated insights. The first chart reports the number of claim-tweet pairs by stance class, along with performance metrics (precision, recall, and F1 score) comparing stance predictions with PolitiFact’s verdicts. The second chart presents stance distributions across U.S. states, helping users identify geographical patterns and regional polarization in public opinion. To support interpretation, TrustMap offers two LLM-based explanation features powered by GPT-4o-mini [17]. At the individual level, users can click a button to generate a rationale explaining why a stance was assigned to a specific claim-tweet pair. At the aggregate level, users can request a summary report highlighting overall stance distribution, dominant narratives, and notable regional deviations. Together, these features help contextualize stance detection results and enhance the interpretation of public discourse.

4 Results and Analysis

4.1 Performance Evaluation and Results

The RATSD model was trained on the dataset from Zhu et al. [34], which contains 2,220 labeled claim-tweet pairs: 1,262 positive stance examples (stance class denoted as ⊕), 451 neutral/no stance examples (⊙), and 507 negative stance examples (⊖). To evaluate

Model	$F_{\oplus}$	$F_{\odot}$	$F_{\ominus}$	F_M
BUT-FIT	83.38	72.00	65.11	80.11
BLCU_NLP	85.37	71.43	63.29	73.36
BERTSCORE+NLI	88.68	72.53	81.04	80.75
BART+NLI	88.00	73.42	74.25	78.56
TESTED	84.09	72.37	67.90	74.75
RATSD$_{\text{Zephyr}}$	88.67	77.38	80.28	82.10
RATSD$_{\text{GPT-3.5}}$	**93.27**	**80.24**	**87.90**	**87.13**

Table 1: Truthfulness stance detection model performance.

performance, RATSD was benchmarked against several state-of-the-art stance detection models, including fine-tuned language models such as pre-trained models (e.g., BUT-FIT [8]), generative pre-trained models (e.g., BLCU_NLP [27]), and domain-adaptive pre-trained models (e.g., BERTSCORE+NLI [10], BART+NLI [18], and TESTED [3]). RATSD utilizes two alternative fine-tuned LLMs: the open-source Zephyr [24] and the proprietary GPT-3.5 [7]. Evaluation metrics include F1 scores for each stance class (denoted as $F_{\oplus}$, $F_{\odot}$ and $F_{\ominus}$) and the macro F1 score (F_M). The evaluation results in Table 1, reproduced from Zhu et al. [34], show that RATSD achieves consistently strong performance across all stance classes.

Given RATSD's strong performance, we rely on its predicted stance labels to evaluate the accuracy of X users' judgments about claims. Specifically, for each claim-tweet pair, we compare PolitiFact's veracity verdict for the claim with the stance expressed in the tweet. A user's stance is considered accurate if it aligns with PolitiFact's assessment of the claim's truthfulness. PolitiFact assigns one of six verdicts to each claim: "True," "Mostly True," "Half True," "Mostly False," "False," and "Pants on Fire." To ensure clarity and avoid overly sparse categories, we consolidated these into three broader groups: "True" and "Mostly True" were mapped to "Truth", "Half True" was renamed "Mixed", and "Mostly False," "False," and "Pants on Fire" were mapped to "Misinformation." This mapping enables us to assess how well social media users collectively discern the truthfulness of claims. Importantly, this evaluation measures users' judgments, not model accuracy, which is separately reported in Table 1. Following this setup, we computed precision and F1 scores of users' judgments for each stance class, as well as recall for each claim verdict category (Table 2).

Stance	Truth	Mixed	Misinfo	Precision	F1
⊕	6,754	5,094	64,643	9.0	15.6
⊙	1,398	1,350	9,677	10.9	11.7
⊖	3,494	4,453	39,177	83.1	48.8
Recall	58.0	12.4	34.5	-	-

Table 2: Distribution of X users' truthfulness stances toward Truth, Mixed, and Misinformation, with precision, recall, and F1 score reported for each stance class.

The results reveal a strong tendency for X users to believe claims are true, regardless of their actual veracity. This is consistent with recent findings [16, 28]. Specifically, nearly 57.0% of misinformation (64,643 out of 64,643+9,677+39,177) is judged as true by users, yielding a recall of only 34.5% for "Misinformation." At the same time, users exhibit considerable skepticism toward true claims: only 58.0% of tweets about true claims convey a positive stance, while more than one-third dispute their validity. Mixed-veracity claims ("Half True") further illustrate the challenge, as users rarely adopt neutral or uncertain positions. Instead, their responses are strongly polarized, with most tweets either endorsing these claims as true or rejecting them as false.

State	Truth-⊕	Truth-⊖	Misinfo-⊕	Misinfo-⊖	Acc	F1
Washington	74.1% (177)	25.9% (62)	**42.7% (719)**	**57.3% (963)**	**65.7**	**51.2**
Florida	69.4% (120)	30.6% (53)	**66.6% (1,147)**	**33.4% (576)**	**51.4**	**32.8**
Texas	70.3% (204)	29.7% (86)	62.7% (1,006)	37.3% (599)	53.8	39.8
New York	75.4% (138)	24.6% (45)	54.9% (850)	45.1% (698)	60.3	42.3
California	67.1% (94)	32.9% (46)	62.7% (827)	37.3% (491)	52.2	35.3
Arizona	**78.0% (46)**	**22.0% (13)**	63.3% (321)	36.7% (186)	57.3	37.1
Colorado	**66.7% (24)**	**33.3% (12)**	58.1% (299)	41.9% (216)	54.3	35.8
Virginia	73.2% (60)	26.8% (22)	62.3% (288)	37.7% (173)	55.3	40.3
U.S.	70.0% (1,444)	30.0% (619)	59.9% (10,299)	40.1% (6,886)	55.0	38.3
All	65.9% (6,754)	34.1% (3,494)	62.3% (64,643)	37.7% (39,177)	51.8	35.0

Table 3: Truthfulness stance distribution toward Truth and Misinformation across U.S. states. Truth-⊕ and Truth-⊖ denote positive and negative stances toward Truth, respectively, while Misinfo-⊕ and Misinfo-⊖ denote positive and negative stances toward Misinformation.

4.2 Geographical Analysis

We analyzed X users' stance by geographical location. As shown in Table 3, Florida records both the highest count and percentage of Misinfo-⊕ (66.6%), indicating that misinformation is widely endorsed in this region. Florida X users also have the lowest accuracy (51.4) and macro F1 score (32.8). In contrast, Washington stands out with the highest accuracy (65.7) and macro F1 score (51.2), as well as the lowest Misinfo-⊕ rate (42.7%), suggesting that relatively more X users in this state push back against misinformation. Yet even in stronger-performing states such as Washington, F1 scores remain modest. This observation underscores that, while some regional variation exists, users overall struggle to discern the truthfulness of claims. This challenge is reflected in the national-level ("United States") results, where accuracy is 55.0 and macro F1 is just 38.3. A comparison between "United States" and "All" (which includes tweets from inside and outside the U.S., as well as those without geolocation) shows that U.S.-based users perform slightly better at distinguishing true from false claims, as evidenced by higher accuracy and macro F1 scores.

5 Conclusion

TrustMap is an interactive application for analyzing how social media users across regions respond to factual claims. By integrating truthfulness stance detection with geospatial analysis, it provides a nuanced view of public reactions to both accurate information and misinformation. The results highlight notable regional variations in how claims are perceived. Overall, this work illustrates the value of combining truthfulness stance classification with visual analytics to better understand public engagement with factual claims.

Acknowledgments

This work is partially supported by the U.S. National Science Foundation award #2346261. The authors acknowledge the Texas Advanced Computing Center (TACC) for providing HPC resources that contributed to the research results reported within this paper.

GenAI USAGE DISCLOSURE

We fully disclose the use of generative AI tools in accordance with CIKM 2025 and ACM guidelines. OpenAI's GPT-4o (via ChatGPT) was employed to enhance the clarity, grammar, and flow of the manuscript. GitHub Copilot assisted with writing, refactoring, and debugging code for dataset processing, annotation tools, and visualizations. All AI-generated content was carefully reviewed, verified, and revised by the authors. Importantly, no core research ideas, experimental design, results, or analysis were produced by AI. All intellectual contributions and final decisions remain solely the responsibility of the authors.

References

[1] Amit Agarwal, Ritu Singh, and Durga Toshniwal. 2018. Geospatial Sentiment Analysis Using Twitter Data for UK-EU Referendum. *Journal of Information and Optimization Sciences* 39, 1 (2018), 303–317.

[2] Hunt Allcott, Matthew Gentzkow, and Chuan Yu. 2019. Trends in the Diffusion of Misinformation on Social Media. *Research & Politics* 6, 2 (2019).

[3] Erik Arakelyan, Arnav Arora, and Isabelle Augenstein. 2023. Topic-Guided Sampling For Data-Efficient Multi-Domain Stance Detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. 13448–13464.

[4] Abu Muna Almaududi Ausat. 2023. The Role of Social Media in Shaping Public Opinion and Its Influence on Economic Decisions. *Technology and Society Perspectives (TACIT)* 1, 1 (2023), 35–44.

[5] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. 2012. The Role of Social Networks in Information Diffusion. In *Proceedings of the 21st International Conference on World Wide Web*. 519–528.

[6] Erik Borra and Bernhard Rieder. 2014. Programmed Method: Developing a Toolset for Capturing and Analyzing Tweets. *Aslib Journal of Information Management* 66, 3 (2014), 262–278.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models Are Few-Shot Learners. *Advances in Neural Information Processing Systems* 33 (2020), 1877–1901.

[8] Martin Fajcik, Pavel Smrz, and Lukás Burget. 2019. BUT-FIT at SemEval-2019 Task 7: Determining the Rumour Stance with Pre-Trained Deep Bidirectional Transformers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. 1097–1104.

[9] GeoPy. 2023. *Geopy: Geocoding library for Python*.

[10] 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*.

[11] Diederik P Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980* (2014).

[12] Kostiantyn Kucher, Rafael M Martins, Carita Paradis, and Andreas Kerren. 2020. StanceVis Prime: Visual Analysis of Sentiment and Stance in Social Media Texts. *Journal of Visualization* 23, 6 (2020), 1015–1034.

[13] 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 (2020), 9459–9474.

[14] Andreas Kilde Lien, Lars Martin Randem, Hans Petter Fauchald Taralrud, and Maryam Edalati. 2022. OSN Dashboard Tool For Sentiment Analysis. *arXiv preprint arXiv:2206.06935* (2022).

[15] Xin Yu Liew, Nazia Hameed, Jeremie Clos, and Joel E Fischer. 2024. Designing and Evaluating a Discourse Analysis Dashboard. In *Proceedings of the Second International Symposium on Trustworthy Autonomous Systems*. 1–5.

[16] Patricia L. Moravec, R. K. Minas, and Alan R. Dennis. 2019. Fake News on Social Media: People Believe What They Want to Believe When It Makes No Sense at All. *MIS Quarterly* 43, 4 (2019), 1343–1360.

[17] OpenAI. 2024. GPT-4o-mini: Optimized Mini Version of GPT-4. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/ Accessed: 2025-02-10.

[18] Revanth Gangi Reddy, Sai Chetan, Zhenhailong Wang, Yi R Fung, Kathryn Conger, Ahmed Elsayed, Martha Palmer, Preslav Nakov, Eduard Hovy, Kevin Small, et al. 2022. NewsClaims: A New Benchmark for Claim Detection From News With Attribute Knowledge. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 6002–6018.

[19] Zhan Shi, Huaxia Rui, and Andrew B Whinston. 2014. Content Sharing in a Social Broadcasting Environment: Evidence From Twitter. *MIS quarterly* 38, 1 (2014), 123–142.

[20] Maria Skeppstedt, Vasiliki Simaki, Carita Paradis, and Andreas Kerren. 2017. Detection of Stance and Sentiment Modifiers in Political Blogs. In *International Conference on Speech and Computer*. 302–311.

[21] Victor Suarez-Lledo and Javier Alvarez-Galvez. 2021. Prevalence of Health Misinformation on Social Media: Systematic Review. *Journal of Medical Internet Research* 23, 1 (2021), e17187.

[22] Kathie M d'I Treen, Hywel TP Williams, and Saffron J O'Neill. 2020. Online Misinformation About Climate Change. *Wiley Interdisciplinary Reviews: Climate Change* 11, 5 (2020), e665.

[23] Joshua A. Tucker, Andrew M. Guess, Pablo Barberá, Cristian Vaccari, Alexandra A. Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. 2018. Social Media, Political Polarization, and Political Disinformation: a Review of the Scientific Literature. *Social Science Research Network (SSRN)* (2018). https://ssrn.com/abstract=3144139

[24] 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* (2023).

[25] Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-Pack: Packed Resources for General Chinese Embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 641–649.

[26] Harry Yaojun Yan, Garrett Morrow, Kai-Cheng Yang, and John Wihbey. 2025. The Origin of Public Concerns Over AI Supercharging Misinformation in the 2024 US Presidential Election. *Harvard Kennedy School Misinformation Review* (2025).

[27] Ruoyao Yang, Wanying Xie, Chunhua Liu, and Dong Yu. 2019. BLCU_NLP at SemEval-2019 Task 7: An Inference Chain-Based GPT Model for Rumour Evaluation. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. 1090–1096.

[28] Haiqi Zhang, Zhengyuan Zhu, Zeyu Zhang, Jacob Devasier, and Chengkai Li. 2024. 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)*. 233–240.

[29] Haiqi Zhang, Zhengyuan Zhu, Zeyu Zhang, and Chengkai Li. 2025. LLMTaxo: Leveraging Large Language Models for Constructing Taxonomy of Factual Claims from Social Media. In *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics, Vienna, Austria, 19627–19641. https://doi.org/10.18653/v1/2025.findings-acl.1007

[30] Zeyu Zhang, Zhengyuan Zhu, Haiqi Zhang, and Chengkai Li. 2024. Exploring Behavioral Tendencies on Social Media: A Perspective Through Claim Check-Worthiness. In *International Conference on Advances in Social Networks Analysis and Mining*. 373–390.

[31] Zeyu Zhang, Zhengyuan Zhu, Haiqi Zhang, Foram Patel, Josue Caraballo, Patrick Hennecke, and Chengkai Li. 2024. Wildfire: A Twitter Social Sensing Platform for Layperson. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*. 1106–1109.

[32] Zhengyuan Zhu, Kevin Meng, Josue Caraballo, Israa Jaradat, Xiao Shi, Zeyu Zhang, Farahnaz Akrami, Haojin Liao, Fatma Arslan, Damian Jimenez, et al. 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*. 99–105.

[33] Zhengyuan Zhu, Haiqi Zhang, Zeyu Zhang, and Chengkai Li. 2025. TSD-CT: A Benchmark Dataset for Truthfulness Stance Detection. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25)*. ACM, Seoul, Republic of Korea. https://doi.org/10.1145/3746252.3761622

[34] Zhengyuan Zhu, Zeyu Zhang, Haiqi Zhang, and Chengkai Li. 2025. RATSD: Retrieval Augmented Truthfulness Stance Detection from Social Media Posts Toward Factual Claims. In *Findings of the Association for Computational Linguistics: NAACL 2025*. 3366–3381.