TSD-CT: A Benchmark Dataset for Truthfulness Stance Detection

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

We present TSD-CT (**T**ruthfulness **S**tance **D**etection–**C**laim and **T**weet), a benchmark dataset designed to advance research in truthfulness stance detection. While prior stance detection datasets focus primarily on political figures, topics, or events, TSD-CT targets truthfulness stance of social media posts toward factual claims. Truthfulness stance reflects whether a post endorses a claim as true, rejects it as false, or expresses no clear position. This focus is particularly valuable for tracking public reactions to misinformation and for enabling applications that analyze belief dynamics in online discourse. TSD-CT comprises 5,331 claim-tweet pairs, each annotated into one of five classes: positive, negative, neutral/no stance, topically different, or problematic. To ensure annotation quality, we introduce a strategy that uses gold-standard labels to compute error scores, evaluate annotator performance, and filter out low-quality contributions. The resulting dataset achieves strong inter-annotator agreement. An error analysis further highlights frequent sources of confusion, particularly between neutral/no stance and other classes. The dataset, along with the annotation interface and codebase, is publicly released to facilitate further research.

CCS Concepts

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

Keywords

Truthfulness Stance, Factual Claim, Social Media, Data Annotation

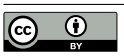
ACM Reference Format:
Zhengyuan Zhu, Haiqi Zhang, Zeyu Zhang, and Chengkai Li. 2025. TSD-CT: A Benchmark Dataset for Truthfulness Stance Detection. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3746252.3761622

1 Introduction

The rise of social media has accelerated the rapid and widespread dissemination of messages and news, influencing public opinion, shaping political discourse, and affecting decision-making. A substantial portion of this content consists of factual claims, which are assertions that can be true or false [11, 18, 19]. Yet, social media users often struggle to distinguish between accurate and inaccurate information, contributing to the spread of misinformation [25]. Understanding how the public engages with and responds to factual claims is therefore critical. Within this context, truthfulness stance detection [29], the task of determining whether a textual utterance affirms, disputes, or remains neutral toward a factual claim, is emerging as a key direction in misinformation analysis research [12, 27, 30].

Despite the growing interest in stance detection, most existing datasets define stance in terms of favorability toward certain entities, such as political figures [15, 17], policies [2], and social issues [3]. In contrast, relatively few datasets focus explicitly on truthfulness stance, in which the factual claims are the targets. Among those that do, some cover a limited number of topics (often a single domain, e.g., COVID-19 misinformation [12]), whereas others focus on news articles rather than social media posts [7, 20]. To address these gaps, this paper introduces the TSD-CT dataset. TSD-CT was constructed using our in-house annotation interface. The dataset consists of 5,331 claim-tweet pairs labeled with five classes: 2,104 *positive* ($\oplus$), 882 *neutral/no stance* ($\odot$), 883 *negative* ($\ominus$), 309 *topically different* (dif), and 1,153 *problematic* (prb). The claims are sourced from PolitiFact (https://www.politifact.com/), and the tweets are retrieved via the Twitter API v2. To ensure label reliability, we implemented a novel quality control strategy that identifies high-quality annotators and tracks their errors using screening pairs with gold-standard labels. The finalized dataset demonstrates strong inter-annotator agreement, with a Gwet's AC2 score of 0.806 and a Krippendorff's alpha of 0.699. We further analyzed the errors made by high-quality annotators on screening pairs and found an overall error rate of 14.59%.

TSD-CT is introduced as a benchmark dataset for training and evaluating machine learning models that detect the truthfulness stance of tweets toward factual claims. With high-quality annotations and broad topical coverage, TSD-CT can be adopted to fine-tune transformer-based models such as BERT [6] and RoBERTa [16], as well as to instruct-tune large language models [1, 22, 23]. Beyond model development, TSD-CT offers a valuable resource for downstream applications such as prioritizing claims for fact-checking, tracking emerging narratives, and analyzing the spread and impact of misinformation on social media [25, 27, 28]. Such applications enable investigations into online polarization and help inform the



CIKM '25, Seoul, Republic of Korea

ACM ISBN 979-8-4007-2040-6/2025/11
https://doi.org/10.1145/3746252.3761622

design of effective strategies to counter misinformation. While this paper focuses on dataset construction and annotation, we note that comprehensive evaluations of a preliminary version of TSD-CT within our recent RATSD framework [30] have already confirmed its effectiveness for both encoder-based and instruction-tuned models.

We also release a web-based annotation interface for truthfulness stance labeling (https://idir.uta.edu/stance_annotation) and its open-source codebase (https://github.com/idirlab/stancedatacollection) to support reproducibility and enable researchers to readily adapt the interface to their own annotation tasks. In line with the FAIR principles [24], the TSD-CT dataset is released under a CC BY 4.0 license on Zenodo (https://doi.org/10.5281/zenodo.15620262).

2 Related Work

Following the conceptual framework of Zhu et al. [30], truthfulness stance is characterized by four components: the *utterance*, the stance *target*, the stance *orientation*, and the stance *type*. While many stance detection studies introduce novel methodologies, we focus on the datasets, comparing differences across these dimensions to contextualize the design and contributions of TSD-CT.

Utterance: from formal news to informal social media. Stance detection tasks differ in the types of utterances they examine. Earlier works such as Emergent [7] and FNC-1 [20], focused on formal news headlines and articles. Other studies, including SemEval-2016 [17], SemEval-2019 [8], P-Stance [15], COVIDLies [12], and this work, target social media content, particularly tweets.

Target: from entities to factual claims. Stance targets span a wide spectrum, from entities and topics (e.g., SemEval-2016, P-Stance) to fact triples (e.g., FactBank [21]), events (e.g., WT-WT [5]), and factual claims (e.g., COVIDLies, Emergent). While many datasets include only a limited set of targets, often fewer than ten, some scale to hundreds or thousands. For instance, COVIDLies comprises 86 claims, Emergent includes 300, and FNC-1 covers 2,542. Our TSD-CT features 2,201 distinct factual claims spanning 845 topics.

Orientation: positive, negative, or neutral/no stance. Stance orientation indicates whether an utterance conveys a positive, negative, or neutral/no stance toward a target. Some datasets, such as SemEval-2019 and [9], distinguish neutral from no stance, but this distinction often suffers from low inter-annotator agreement. Following Zhu et al. [30], TSD-CT merges the two into a single category to reduce ambiguity and improve annotation consistency.

Type: favorability vs. truthfulness. A key distinction among stance detection tasks lies in the stance type. Favorability-based tasks assess support or opposition toward an entity or topic (e.g., SemEval-2016, P-Stance), whereas truthfulness stance detection evaluates whether an utterance affirms, denies, or remains neutral regarding the veracity of a factual claim. Truthfulness stance is both more nuanced and less studied. While COVIDLies also addresses the truthfulness stance of tweets, it is restricted to false claims about the coronavirus. In contrast, TSD-CT encompasses both true and false claims spanning a wide range of topics.

3 Dataset Construction

Claim-tweet pair collection. Factual claims were collected from PolitiFact together with their assigned topics, with each claim potentially associated with one or more topics. For each claim, associated

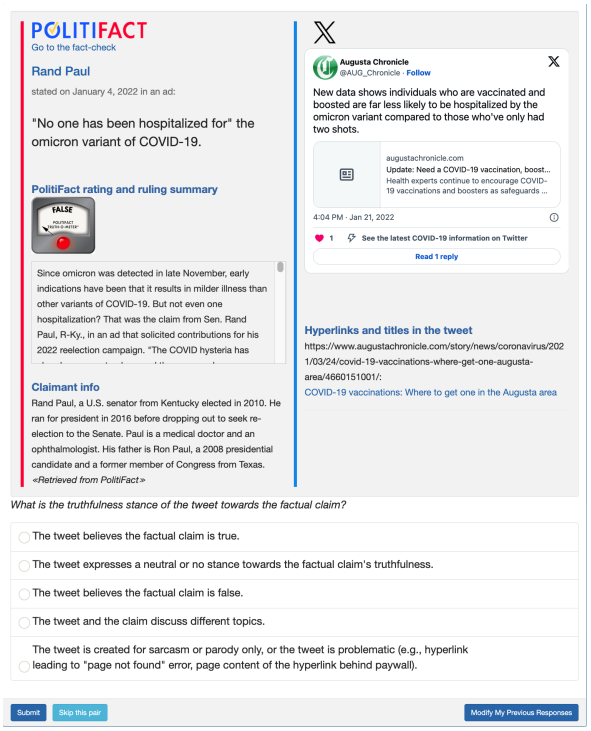


Figure 1: The annotation interface.

tweets were retrieved via the Twitter API v2 using a keyword-based retrieval method. Keywords (nouns, verbs, adjectives, and numerical values) were extracted from the claim to form conjunctive search queries. A temporal window was applied to select tweets posted from one month prior to up to one year after the claim's publication date. To preserve linguistic diversity and reduce redundancy, retweets, replies, and quoted tweets were excluded. Tweets shorter than 30 characters were also removed to eliminate overly brief and ambiguous content.

Annotation interface. We developed a web-based annotation interface (Figure 1) for stance labeling. The interface consists of two primary panels. The left panel presents information about the factual claim from PolitiFact, including the claim text, claimant details, verdict, and a summary of the fact-check review. The right panel displays a corresponding tweet, showing the user who posted it and the timestamp. To assist annotators in verifying external references, the panel also provides clickable page titles that link directly to the sources mentioned in the tweet.

Below the two panels, annotators can choose from five labeling options: *positive* ($\oplus$), *neutral/no stance* ($\odot$), *negative* ($\ominus$), *topically different* (dif), and *problematic* (prb). The $\oplus$ stance applies when a tweet conveys the belief that the claim is true, while $\ominus$ indicates that the tweet asserts the claim is false. The $\odot$ stance is assigned when the tweet either expresses uncertainty about the claim's truthfulness (neutral) or does not explicitly take a position despite discussing the same topic (no stance). The dif class denotes cases where the tweet and claim address different topics. Finally, the prb class applies to tweets that are created solely for sarcasm or parody (as sarcasm and parody detection is outside the scope of this study) or are otherwise problematic (e.g., containing broken hyperlinks or paywalled content). Annotators also have the option

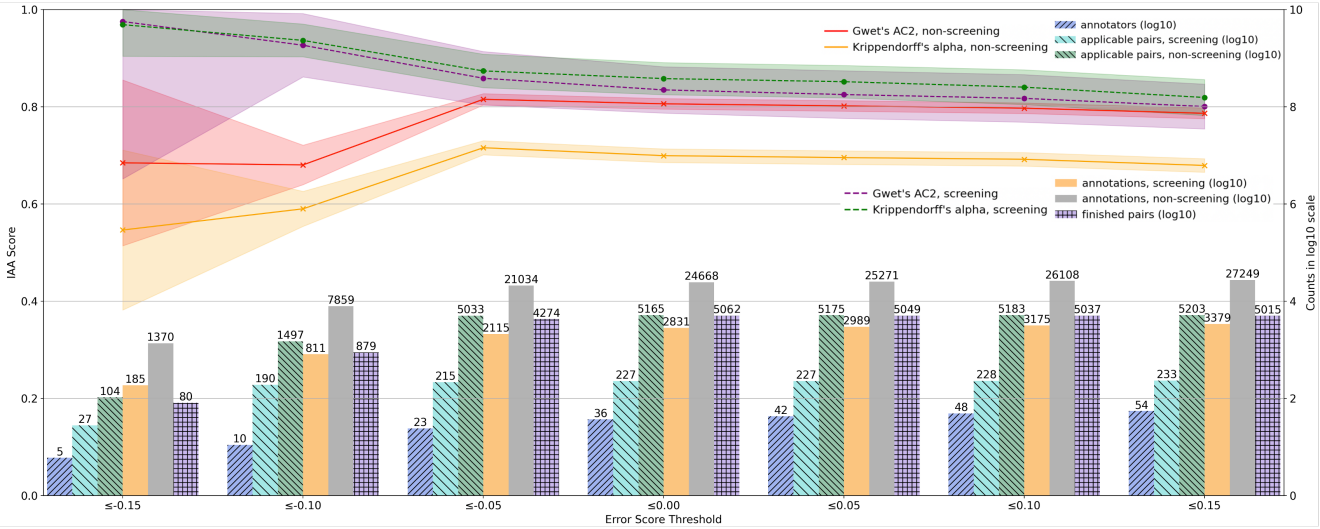


Figure 2: Annotation statistics and IAA scores across error score thresholds.

to skip a pair if they are uncertain about the appropriate class to choose for the pair.

Annotation guideline. Detailed instructions were provided to annotators upon login to the annotation interface. The instructions clearly defined each stance category and illustrated them with examples. The instructions also covered edge cases and special scenarios, such as sarcasm, indirect references, and ambiguous statements. Prior to annotation, annotators completed a training phase using 16 expert-annotated claim-tweet pairs. After labeling each training example, the interface immediately revealed the gold-standard label along with an explanation to help annotators refine their understanding of the annotation guidelines.

Annotator recruitment. We recruited annotators by distributing flyers and sending email announcements across our university. All participants were at least 18 years old and fluent in English. Compensation was provided in the form of gift cards, with earnings determined by both annotation quality and the number of completed annotations. High-performing annotators could earn up to 20 U.S. cents per claim-tweet pair. In total, 214 individuals registered as annotators.

Quality control strategy. To ensure high-quality annotations, we adapted the design of Arslan et al. [4] to implement a screening-based quality control strategy that continuously monitored annotator performance. Four domain experts annotated a set of claim-tweet pairs, among which 253 received unanimous labels. These consensus pairs were used as screening items and randomly inserted into the annotation stream. Initially, each annotator had a 30% chance of receiving a screening pair, which gradually decreased to 10% as they completed more tasks.

We assessed annotation quality using a weighted error score (E_w), computed from an annotator's screening pair responses: $E_w = \frac{-0.2 \cdot N_{\text{correct}} + 0.5 \cdot N_{\text{mild}} + 1.0 \cdot N_{\text{moderate}} + 2.0 \cdot N_{\text{severe}}}{N_{screen}}$ where N_{screen} is the number of completed screening pairs by the annotator. $N_{correct}$, N_{mild}, $N_{moderate}$, and N_{severe} denote the counts of exact matches and misclassifications of increasing severity. For example, labeling $\oplus$ as $\ominus$ contributes to the severe error count (N_{severe}), labeling $\odot$ as $\oplus$ contributes to the moderate error count ($N_{moderate}$), and labeling prb as $\odot$ contributes to the mild error count (N_{mild}). (Refer to [26] for the full specification of the scenarios corresponding to each severity level.) We tied annotation quality to monetary compensation through a reward function: $R = 2.0 \cdot Q(E_w) \cdot \left(\frac{A_{\text{ann}}}{A_{\text{all}}}\right)^2 \cdot 0.6^{\frac{N_{\text{skip}}}{N_{\text{ann}}}} \cdot \frac{N_{\text{ann}}}{100}$ where N_{ann} is the number of claim-tweet pairs annotated by the annotator, A_{ann} is the average token length of those pairs, and A_{all} is the average token length per pair across the whole dataset, and N_{skip} is the number of pairs the annotator chose to skip. The quality multiplier $Q(E_w)$ rewards low-error annotators and penalizes high-error ones, defined as:

$$Q(E_w) = \begin{cases} 3 - \frac{7 \cdot E_w}{0.2}, & \text{if } E_w \le 0 \\ \left(\frac{0.3 - E_w}{0.3}\right)^{2.5} \cdot 3, & \text{if } 0 < E_w \le 0.3 \\ 0, & \text{otherwise} \end{cases}$$

This design ensured that annotators who consistently maintained low error rates and engaged with longer, more substantive content were fairly and proportionally rewarded.

Stopping condition. A claim-tweet pair was considered finished once sufficient agreement was reached among high-quality annotators (those with $E_w < 0$ in our setting). Only annotations from this group were used when determining completion. Specifically, the stopping condition was satisfied if any one of the four labels—$\oplus$, $\odot$, $\ominus$, or dif—met all of the following criteria based solely on high-quality annotations: (1) It was selected by at least three annotators; (2) It exceeded each of the other three labels by at least two votes; (3) Its vote count was greater than or equal to half the combined total of the other three labels. As an exception, the annotation process for a pair stopped immediately if any high-quality annotator selects the prb label.

4 Dataset Analysis

Figure 2 illustrates how key statistics of TSD-CT vary across error score thresholds (E_w). These statistics include the number of finished pairs and the number of annotators classified as high-quality under the threshold. All values reported in the figure are derived exclusively from annotations provided by high-quality annotators.

The top portion of Figure 2 reports inter-annotator agreement (IAA) scores, including Krippendorff's alpha [14] and Gwet's AC2 [10], both of which are suitable for incomplete data—in this case, the fact that not every annotator labeled every pair. IAA scores were calculated for the ⊕, ⊙, and ⊖ classes, which are treated as ordinal variables. The *dif* and *prb* classes were excluded since they lie outside the stance detection task scope. Furthermore, *prb* can be assigned based on the judgment of a single high-quality annotator (as discussed in Section 3), making agreement computation inapplicable. The IAA scores were computed separately for screening pairs and for non-screening pairs (i.e., real task annotations). The shaded areas in the figure represent 95% confidence intervals.

The bottom portion of Figure 2 presents counts on a logarithmic scale, including annotators, finished pairs (i.e., pairs meeting the stopping condition), applicable screening and non-screening pairs (i.e., those with at least two annotations), and annotations for screening and non-screening pairs.

Threshold-based quality filtering and dataset growth. As the threshold E_w is relaxed (from ≤−0.15 to ≤0.15), more annotators qualify as high-quality, resulting in sharp increases in both annotations and finished pairs. At the strictest setting (≤−0.15), only five annotators are retained, yielding 1,370 annotations across just 80 finished pairs. At a more moderate threshold (≤0.00), 36 annotators contribute 24,668 annotations and 5,062 finished pairs. Beyond this point, however, the number of finished pairs declines, suggesting that the inclusion of lower-quality annotators reduces the likelihood of reaching agreement.

Inter-annotator agreement. At strict thresholds (≤−0.15 or ≤−0.10), the IAA scores for non-screening pairs are low due to the small number of applicable pairs—many claim–tweet pairs have too few annotations to yield reliable agreement, leading to wide confidence intervals. For example, at ≤−0.15, both Krippendorff's alpha and Gwet's AC2 for screening pairs reach near-perfect values because of the small sample size and high consistency, yet their confidence intervals are large and less reliable. As the threshold increases to ≤−0.05, agreement peaks, with Gwet's AC2 at 0.815 and Krippendorff's alpha at 0.716. Beyond this point, agreement scores decline as lower-quality annotators are included, though they remain strong and stable. The number of finished pairs reaches its maximum at the threshold of 0.00, where Gwet's AC2 is 0.803 and Krippendorff's alpha is 0.699. To balance annotation quality with dataset size, we adopted 0.00 as the threshold for final dataset.

Error analysis. To better understand annotator errors, we analyzed the 2,831 annotations on screening pairs contributed by high-quality annotators (those with $E_w \leq 0$). Figure 3 presents the confusion matrix between gold-standard labels (columns) and annotators' responses (rows). Off-diagonal cells report the count and column-wise percentage of misclassifications, while diagonal cells show the number of correct annotations together with the recall for the class represented by the corresponding column.

Overall, we observed 413 misclassifications, yielding an error rate of 14.59%. The *prb* class had the lowest recall at 48.6%, largely because annotators often skipped URL verification and found it difficult to distinguish sarcasm or parody. By contrast, ⊕, ⊙, and ⊖ achieved high recall at 92.5%, 81.4%, and 91.4%, respectively.

The ⊙ class was the most frequent misclassification across all other classes, both in absolute count and percentage. For instance,

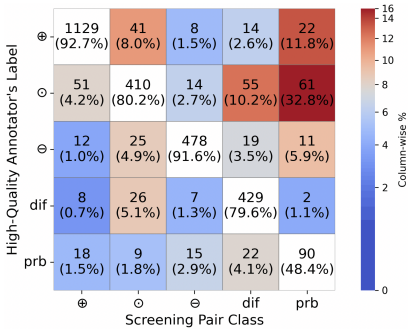


Figure 3: Confusion matrix for screening pair annotations by high-quality annotators.

dif was mislabeled as ⊙ 52 times, and *prb* as ⊙ 54 times. This pattern suggests that annotators often default to ⊙—the middle-ground stance—as a "safe" choice when uncertain, consistent with prior findings in stance detection [13]. This tendency underscores the inherent ambiguity of the neutral/no-stance class. While it poses challenges for modeling, it also mirrors the uncertainty present in real discourse, making it a valuable testbed for robustness.

Final dataset composition. Under the chosen threshold of $E_w \leq 0$, TSD-CT comprises 5,331 finalized claim-tweet pairs (including 5,062 non-screening pairs, 253 screening pairs, and 16 training pairs), covering 2,201 distinct factual claims across 845 topics as assigned by PolitiFact. The class distribution is: 2,104 (39.47%) ⊕, 882 (16.54%) ⊙, 883 (16.56%) ⊖, 309 (5.80%) *dif*, and 1,153 (21.63%) *prb*. The dataset also reflects diverse claim veracity according to PolitiFact rulings: 722 (32.80%) false claims, 353 (16.04%) "pants on fire," 340 (15.45%) barely true, 292 (13.27%) half true, 287 (13.04%) mostly true, and 207 (9.40%) true. On average, each claim–tweet pair contains 34.61 tokens and 261.76 characters. Moreover, 2,169 pairs (40.71%) include at least one hyperlink, indicating external reference or contextual support. Among the 845 claim topics, the dataset is dominated by discussions of coronavirus (1,573 pairs; 29.51%) and public health (870; 16.32%), followed by Donald Trump (583; 10.94%), elections (438; 8.22%), economy (390; 7.32%), health care (329; 6.17%), crime (295; 5.53%), government regulation (255; 4.78%), drugs (240; 4.50%), science (240; 4.50%), among others.

Ethical consideration. The data annotation for TSD-CT was approved by the Institutional Review Board at the University of Texas at Arlington (Protocol No. 2023-0093.2). In compliance with X's content-sharing policy, the released dataset provides only tweet IDs and associated metadata, rather than raw tweet text.

5 Conclusion

This paper introduces the TSD-CT dataset, constructed through systematic claim collection, tweet retrieval, annotation, and quality control. With broad topical coverage and high-quality annotations, TSD-CT supports both model development and analyses of online discourse. Future extensions include multilingual and multimodal versions to broaden its applicability for studying truthfulness stance in diverse contexts.

Acknowledgments

This work is partially supported by the U.S. National Science Foundation award #2346261.

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] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Ana Aleksandric, Henry Isaac Anderson, Anisha Dangal, Gabriela Mustata Wilson, and Shirin Nilizadeh. 2024. Analyzing the Stance of Facebook Posts on Abortion Considering State-Level Health and Social Compositions. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 18. 15–28.

[3] Emily Allaway and Kathleen R. McKeown. 2020. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. 8913–8931.

[4] Fatma Arslan, Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2020. A Benchmark Dataset of Check-Worthy Factual Claims. In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media (ICWSM)*. 821–829.

[5] 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*. 1715–1724.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*. 4171–4186.

[7] William Ferreira and Andreas Vlachos. 2016. Emergent: A Novel Dataset for Stance Classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*. 1163–1168.

[8] 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*. 845–854.

[9] Lara Grimminger and Roman Klinger. 2021. Hate Towards the Political Opponent: A Twitter Corpus Study of the 2020 US Elections on the Basis of Offensive Speech and Stance Detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. 171–180.

[10] Kilem Li Gwet. 2008. Computing Inter-Rater Reliability and its Variance in the Presence of High Agreement. *Brit. J. Math. Statist. Psych.* 61, 1 (2008), 29–48.

[11] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward Automated Fact-Checking: Detecting Check-worthy Factual Claims by ClaimBuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. 1803–1812.

[12] 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*.

[13] Kenneth Joseph, Lisa Friedland, William Hobbs, Oren Tsur, and David Lazer. 2017. ConStance: Modeling Annotation Contexts to Improve Stance Classification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 1115–1124.

[14] Klaus Krippendorff. 1970. Estimating the Reliability, Systematic Error and Random Error of Interval Data. *Educational and Psychological Measurement* 30, 1 (1970), 61–70.

[15] Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 2021. P-Stance: A Large Dataset for Stance Detection in the Political Domain. In *Findings of the Association for Computational Linguistics*. 2355–2365.

[16] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692* (2019).

[17] 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*. 31–41.

[18] Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated Fact-Checking for Assisting Human Fact-Checkers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. 4551–4558. Survey Track.

[19] Jingwei Ni, Minjing Shi, Dominik Stammbach, Mrinmaya Sachan, Elliott Ash, and Markus Leippold. 2024. AFaCTA: Assisting the Annotation of Factual Claim Detection with Reliable LLM Annotators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Bangkok, Thailand, 1890–1912.

[20] Dean Pomerleau and Delip Rao. 2017. Fake News Challenge Stage 1 (FNC-1): Stance Detection. http://www.fakenewschallenge.org.

[21] Roser Saurí and James Pustejovsky. 2009. FactBank: A Corpus Annotated With Event Factuality. *Language Resources and Evaluation* 43 (2009), 227–268.

[22] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint arXiv:2312.11805* (2023).

[23] Hugo Touvron, Louis Martin, Kevin Lu, Olatunji Ruwase, Shruti Bhosale, Jinjing Jiang, Armand Joulin, Myle Ott, and Yann LeCun. 2024. LLaMA 3: Open Foundation and Instruction Models. https://ai.meta.com/blog/meta-llama-3/. Meta AI Blog.

[24] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data* 3, 1 (2016), 1–9.

[25] 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.

[26] Zhengyuan Zhu. 2025. *Understanding Misinformation on Social Media through Truthfulness Stance*. Ph. D. Dissertation. University of Texas at Arlington.

[27] 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*. Association for Computational Linguistics, Online, 99–105.

[28] 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)*. ACM, Seoul, Republic of Korea. https://doi.org/10.1145/3746252.3761490

[29] 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*.

[30] 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.