

UNDERSTANDING MISINFORMATION ON SOCIAL MEDIA THROUGH
TRUTHFULNESS STANCE

by

ZHENGYUAN ZHU

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington in Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON
August 2025

Copyright © by Zhengyuan Zhu 2025

All Rights Reserved

To all the human beings on the resonant social media and within the silent universe.

“Each of us carries a unique belief system—that is the beauty of humanity. No one can claim with certainty to fully know another or even themselves. Yet, in seeking to understand ourselves as whole and complete beings, we edge closer to true knowing.”

— Zhengyuan Zhu

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my Ph.D. advisor, Dr. Chengkai Li, for his unwavering support, insightful guidance, and dedication throughout my doctoral journey. His commitment to research excellence, innovative thinking, and meticulous attention to detail has profoundly shaped my academic and professional growth. Under his mentorship, I have developed strong analytical and problem-solving skills, along with a deep appreciation for scholarly rigor and integrity.

I am also sincerely thankful to Dr. Vassilis Athitsos, Dr. Shirin Nilizadeh, and Dr. Gautam Das for serving on my dissertation committee. Their valuable feedback, encouragement, and thoughtful suggestions were instrumental in shaping the direction and quality of this work.

I wish to extend my heartfelt thanks to the faculty and staff of the Computer Science and Engineering Department at the University of Texas at Arlington for providing an enriching research environment and the necessary infrastructure to support my studies. In particular, I am grateful to Mrs. Ginger Dickens, Mrs. Sherri Gotcher, Mrs. Jade Myers, and Mrs. Samantha Oliva for their constant assistance, patience, and readiness to help whenever I needed guidance.

My sincere appreciation also goes to my friends and labmates at the IDIR Lab, including Penny, Allen, Xiao, Nasim, Jacob, Qing, Fatma, Josue, Sami, Damian, and all the dedicated master's and bachelor's students. I am grateful for their stimulating conversations, collaborative spirit, and unwavering support throughout this journey.

July 25th, 2025

ABSTRACT

UNDERSTANDING MISINFORMATION ON SOCIAL MEDIA THROUGH TRUTHFULNESS STANCE

Zhengyuan Zhu, Ph.D.

The University of Texas at Arlington, 2025

Supervising Professor: Chengkai Li

Misinformation on social media has become a pervasive issue that profoundly influences public opinion and decision-making. As false or misleading claims circulate widely online, there is a critical need for analytical tools to understand how people react to such claims. This dissertation introduces the concept of truthfulness stance as a key lens for social sensing. In essence, truthfulness stance assesses whether a textual utterance believes a factual claim to be true, false, or expresses a neutral stance or no stance toward the claim. Leveraging stance in this manner fills an important gap in misinformation research: it enables us to gauge the public's collective judgment on what is true or false, thereby offering a proxy measure of misinformation acceptance or rejection at scale. Understanding these stances is vital, as they can reveal how misinformation spreads and influences society, informing strategies in political decision-making and public health interventions.

Despite a rich body of work on stance detection, prior studies have often used varying definitions for “stance,” leading to a fragmented understanding of the concept. This dissertation provides the first in-depth conceptual framework unifying these

definitions. In doing so, it delineates how truthfulness stance relates to or differs from other stance types. To place truthfulness stance in context, we propose a conceptual framework that generalizes stance definitions across different studies. Typically, any stance instance involves four components: (1) an utterance (e.g., a social media post or a news article in which the stance is expressed), (2) a target (the entity, topic, or factual claim that the stance is about), (3) an orientation of the stance (e.g., positive, neutral/no stance, negative), and (4) the type of stance being expressed (e.g., favorability, likelihood, or truthfulness).

Since truthfulness stance differs from conventional definitions of stance, there is a pressing need for a dedicated dataset that captures this distinction. To address this gap, this dissertation introduces a novel labeled dataset, TSD-CT, which stands for **T**ruthfulness **S**tance **D**etection for **C**laim–**T**weet pairs. TSD-CT contains 5,331 claim–tweet pairs, each consisting of a factual claim and a corresponding social media post (specifically, a tweet) expressing a truthfulness stance toward that claim. The claims were sourced from PolitiFact, a popular fact-checking website. For each claim, we collected tweets that mention or discuss the claim. We then collected human annotations that label the stance of each tweet as “positive,” “negative,” “neutral/no stance,” or “different topic.” These labels correspond to whether the tweet believes the claim is true, believes it is false, is unsure or neutral, does not take a stance, or if the tweet is unrelated to the claim. We included a fifth category, “problematic,” during annotation to flag tweets that were nonsensical (e.g., pure sarcasm or invalid content). The annotations were performed using an in-house web platform with rigorous quality control. Annotators, who were recruited and trained from the university community, were provided with detailed guidelines and examples for each stance category. Multiple mechanisms were in place to monitor agreement and consistency, including an administrative dashboard to track annotator performance in

real time. The resulting dataset is publicly available on Zenodo, providing a resource for advancing research in computational social science and social media analysis.

At the core of our approach to identifying the truthfulness stance of social media posts is **RATSD**, a novel framework for automated stance classification with respect to factual claims. **RATSD** stands for **R**etrieval-**A**ugmented **T**ruthfulness **S**tance **D**etection, reflecting its hybrid design that combines large language models (LLMs) with information retrieval techniques. The motivation behind **RATSD** is to overcome the key challenge in stance detection: contextual understanding. Social media posts such as Tweets are often short, filled with colloquialisms or sarcasm, and may lack context, which makes it hard for a model to determine a tweet’s stance towards a claim in isolation. Our idea is to empower the model with additional knowledge and a reasoning process before it makes a stance judgment. Notably, to the best of our knowledge, this work is the first to apply RAG techniques to stance detection. Our experiment results demonstrate that injecting contextual knowledge in this way substantially improves a model’s performance on the task.

We rigorously evaluated **RATSD** on multiple datasets to validate its effectiveness. First, we tested on our **TSD-CT** dataset, using it as a benchmark for truthfulness stance detection. Additionally, we included three existing stance datasets in the evaluation: SemEval-2019 (a collection of tweets from a shared task on stance toward the veracity of rumors), the **WT-WT** (Will-They-Won’t-They) dataset (a large stance dataset focused on predicting outcomes in the financial domain), and **COVIDLies** (tweets annotated for stance regarding COVID-19 related misinformation). We compared **RATSD** with several state-of-the-art baseline models from recent literature. The results show a clear advantage for **RATSD**. Our framework outperforms state-of-the-art methods on all datasets, with particularly strong gains on the **TSD-CT** dataset.

An ablation study confirmed that the components of the framework each contribute to these gains: removing any component led to notable drops in performance.

A key aspect of this research is demonstrating how truthfulness stance detection can power practical tools for misinformation monitoring and analysis. We developed and deployed several proof-of-concept applications that utilize our stance detection approach to provide actionable insights. We created **TrustMap**, a web-based interactive tool that visualizes the aggregate truthfulness stances of social media posts across different geographic regions. **TrustMap** ingests streams of tweets about various factual claims and applies our RATSD framework to label each tweet as positive, neutral/no stance, or negative. It clusters and displays these tweets on a map of the United States, allowing users to explore how different regions respond to specific claims under certain topics. The geographical patterns unveiled by **TrustMap** help researchers and policymakers identify regional variations in misinformation belief. By connecting stance detection with geospatial analysis, **TrustMap** offers a novel perspective on public engagement with factual claims. We also applied our stance detection framework to the topic of climate change. We collected climate-related factual claims and tweets related to them. Using a variant of RATSD framework, we analyzed public perceptions on climate-related issues. The findings were telling: the public tends to believe most climate-related claims are true, regardless of the claims' actual veracity, indicating a concerning bias toward accepting information at face value. By identifying where people are overly credulous or confused about facts, stakeholders (such as science communicators or environmental agencies) can better target their educational efforts. Another application is a COVID-19 misinformation dashboard. This dashboard was developed during the COVID-19 pandemic to help track and mitigate the so-called “COVID-19 misinfodemic.” We curated a catalog of known COVID-19 facts (e.g., “Vaccines reduce transmission”) and debunked myths (e.g., “5G spreads the virus”),

and then monitored Twitter to observe how frequently these pieces of information were appearing and whether users were endorsing or rejecting them. Using a BERT-based stance detection model (an early version of our system), the dashboard could match tweets to the closest factual claim or myth and determine the tweet’s stance toward it. This enabled the platform to display, for any U.S. region, which COVID-related rumors were most prevalent and whether the local Twitter discourse was pushing back against those rumors or amplifying them. Such a tool provided health officials and the public with situational awareness of the misinformation landscape during the pandemic. Collectively, these applications show that the methods developed in this dissertation are not merely theoretical. They can be operationalized to tackle real-world challenges.

In conclusion, this dissertation contributes a novel perspective and a toolkit for understanding misinformation through the lens of truthfulness stance. We demonstrated a full pipeline from conceptual foundations and data resources to algorithms and deployment. The approach has proven effective in controlled experiments and useful in practical scenarios. Going forward, we envision truthfulness stance analysis becoming an integral part of misinformation research, helping scholars and practitioners chart the landscape of truth and lies in social media discourse. Ultimately, by shedding light on collective truth perceptions, our work aids in fostering a more informed and resilient society in the digital age.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	iv
ABSTRACT	v
LIST OF ILLUSTRATIONS	xiii
LIST OF TABLES	xvi
Chapter	Page
1. INTRODUCTION	1
1.1 Background	1
1.1.1 The Challenge of Misinformation in Online Discourse	1
1.1.2 On Combating Misinformation	2
1.2 Truthfulness Stance: Gauging Belief in Claims	5
1.3 The TSD-CT Dataset: A Benchmark for Truthfulness Stance Detection	8
1.4 The RATSD Framework: Retrieval-Augmented Stance Detection . . .	10
1.5 Real World Applications	11
1.6 Significance and Outlook	13
2. CONCEPTUAL FRAMEWORK OF TRUTHFULNESS STANCE DETECTION	15
2.1 Orientation of Stance	16
2.2 Utterance of Stance	19
2.3 Target of Stance	20
2.4 Type of Stance	21
3. THE TRUTHFULNESS STANCE DETECTION DATASET	23
3.1 Fact-Check Collection	24

3.2	Claim-Tweet Pair Collection	26
3.3	Claim-Tweet Pair Sanitization	27
4.	THE TRUTHFULNESS STANCE DATA ANNOTATION	30
4.1	Annotator Recruitment	30
4.2	Annotation Interface	31
4.3	Quality Control Strategy	39
4.4	Stopping Condition	42
4.5	Dataset Statistics	43
5.	EXPLORING TRUTHFULNESS STANCE DETECTION APPROACHES	49
5.1	Early Approaches	49
5.2	Neutral Network Models	50
5.3	Pre-Trained Language Models	51
5.4	The RATSD Framework	52
5.4.1	Knowledge Corpora Construction	54
5.4.2	Contextual Knowledge Generation	55
5.4.3	Classification Model	58
6.	EVALUATION	59
6.1	Experiment Datasets	59
6.2	Implementation Details	59
6.3	Experiment Results	60
6.4	Ablation Study	64
7.	APPLICATIONS: MAPPING TRUTHFULNESS STANCE OF SOCIAL MEDIA POSTS ON FACTUAL CLAIMS	65
7.1	Overview of the Applications	65
7.2	Related System	68
7.3	Overview of the Framework	69

7.4	Framework Design	70
7.4.1	Data Ingestion	70
7.4.2	Taxonomy Construction	73
7.4.3	Truthfulness Stance Detection Model	74
7.4.4	Data Exploration	74
7.5	Results and Analysis	81
7.5.1	Results of Truthfulness Stance Detection in TrustMap	81
7.5.2	Results of Truthfulness Stance Detection in ClimateStance	84
7.5.3	Result of Taxonomy Generation in ClimateStance	84
7.5.4	Detailed Analysis for TrustMap	87
7.5.5	Detailed Analysis for Climate Stance	89
7.6	Limitations and Ethics Statement of TrustMap	91
7.7	A Dashboard for Mitigating the COVID-19 Misinfodemic	92
7.7.1	Background	93
7.7.2	The Dashboard	96
7.7.3	The Datasets	97
7.7.4	Matching Tweets with Facts and Stance Detection	100
7.7.5	Evaluation and Results	103
8.	CONCLUSIONS AND FUTURE WORK	108
8.1	Societal Impact	109
8.2	Future Work	110
APPENDIX		
A.	IMPLEMENTATION DETAILS	114
	Bibliography	128
	BIOGRAPHICAL STATEMENT	149

LIST OF ILLUSTRATIONS

Figure	Page
1.1 Four tweets expressing different truthfulness stances toward the same factual claim.	6
2.1 A claim c and a tweet t may be related or unrelated. Related claim-tweet pairs are partitioned into four cases of stance. The cases within the dashed, purple frame (positive, neutral, negative) are applicable when t expresses a stance regarding c 's truthfulness. Though conceptually different, for practical purposes neutral and no stance are combined into one class \odot in our consideration.	17
2.2 A few samples of challenging cases in truthfulness stance annotation. .	18
3.1 An example of Politifact fact-check.	25
4.1 User login page of the annotation interface.	31
4.2 The registration form after clicking the “Sign up” button.	32
4.3 The annotation instruction page.	33
4.4 The main annotation page.	34
4.5 The page for modifying previous annotations.	35
4.6 The leaderboard pop-up in the annotation interface.	36
4.7 A table showing annotation progress by subset.	37
4.8 A table showing label distribution of each annotator.	38
4.9 A table showing annotation quality of each annotator.	39
4.10 A table showing annotation details for claim–tweet pairs.	40

4.11 Annotation statistics and IAA scores by varying error score thresholds. Line plots show Krippendorff’s alpha and Gwet’s AC2 for screening and non-screening pairs with confidence intervals. Bar plots indicate counts on a logarithmic scale for annotations, screening annotations, finished pairs, applicable pairs, and annotators.	44
4.12 Confusion matrix for annotations on screening pairs by high-quality an- notators. Diagonal values indicate correct answers; off-diagonal values indicate errors with both raw count and recall (column-wise percentage).	47
5.1 The RATSD framework.	54
5.2 An example of stance analysis.	57
7.1 Geographic distribution of collected tweets across U.S. states.	72
7.2 The overall user interface of TrustMap.	75
7.3 The control panel in TrustMap.	76
7.4 The map view in TrustMap.	78
7.5 An example of a claim-tweet pair and its truthfulness stance displayed on TrustMap.	79
7.6 The statistical chart in TrustMap.	80
7.7 The stance explanation for an individual tweet in TrustMap.	81
7.8 The stance distribution explanation in TrustMap.	81
7.9 A fragment of climate change taxonomy.	85
7.10 The user interface of the dashboard for mitigating the COVID-19 mis- infodemic.	95
7.11 Matching tweets with facts and stance detection.	99
7.12 An entry of the catalog stored in RDF.	101
7.13 Six countries with the most misinformation tweets.	105

A.1	The prompt used to generate contextual knowledge for each claim–tweet pair.	123
A.2	The contextual knowledge about the claim.	124
A.3	The contextual knowledge about the tweet.	125
A.4	Prompt used to generate topics for each claim.	126

LIST OF TABLES

Table	Page
2.1 Various definitions of stance differ in the type, utterance, and target of stance.	19
3.1 Field statistics from the fact-check collection. * Not all websites follow a consistent structure in their fact-checks. For instance, AFP Fact Check and FactCheck.org do not separately list the claims they fact-check.	26
4.1 Distribution of stance labels in the finalized TSD-CT dataset.	48
6.1 Label distribution of SemEval-2019, WT-WT, COVIDLies and TSD-CT datasets.	60
6.2 Performance comparison on the TSD-CT dataset.	62
6.3 Performance comparison on the SemEval-2019 dataset.	62
6.4 Performance comparison on the WT-WT dataset.	63
6.5 Performance comparison on the COVIDLies dataset.	63
6.6 The zero-shot model performance comparison and ablation study on the TSD-CT dataset.	64
7.1 Distribution of X users' truthfulness stances toward true, mixed, and false claims, along with the precision, recall, and F1-score for each stance category in TrustMap.	82

7.2	Truthfulness stance distribution towards Truth and Misinformation across PolitiFact’s topics. Truth-⊕ and Truth-⊖ denote positive and negative stances towards Truth , respectively. Misinfo-⊕ and Misinfo-⊖ denote positive and negative stances towards Misinformation , respectively. Note that “All” includes every topic, not just the selected eight, so the total count for the eight topics does not equal the count for “All.”	82
7.3	Performance of truthfulness stance classifier on the annotated dataset.	84
7.4	Distribution of tweets by truthfulness stance toward claims.	84
7.5	Truthfulness stance distribution toward Truth and Misinformation across U.S. states.	87
7.6	Truthfulness stance distribution by political leaning.	88
7.7	Examples of truthfulness stance detection and their corresponding topics in the taxonomy.	90
7.8	Stance distribution towards Truth and Misinformation across broad topics. Truth-⊕ and Truth-⊖ denote positive and negative stances towards Truth , respectively. Misi-⊕ and Misi-⊖ denote positive and negative stances towards Misinformation , respectively. Note that the topic “ <i>Others</i> ” is not considered in this analysis.	91
7.9	Example results of matching tweets with facts and stance detection. .	100
7.10	Performance of claim matching on the 800 tweet-fact pairs.	103
7.11	Performance of Stance-BERT on the FNC-1 test dataset and 200 matched tweet-fact pairs.	104
7.12	Correlation between the percentage of confirmed/deceased/recovered cases and the percentage of misinformation tweets. The number of recovered cases in U.K. after April 13th is missing from the data source.	106
7.13	Most frequent categories of misinformation tweets.	107

- A.1 Examples of relatively inaccurate medium topics in the public's judgments. 125
- A.2 Examples of relatively accurate medium topics in the public's judgments. 125

CHAPTER 1

INTRODUCTION

1.1 Background

1.1.1 The Challenge of Misinformation in Online Discourse

In today's digital age, social media and other online platforms serve as primary channels for news and information dissemination (Bakshy et al., 2012; Shi et al., 2014). While these platforms enable instant sharing and broad distribution, they also contribute to the rapid spread of misinformation—false or misleading content that can misinform the public (Lazer et al., 2018). Studies (Vosoughi et al., 2018; Dizikes, 2018) have shown that misinformation spreads farther and faster than fact on social media. In particular, as reported in (Dizikes, 2018), an analysis of millions of tweets revealed that false stories were about 70% more likely to be retweeted than true ones, and reached large audiences roughly six times faster.

The virality of misinformation is driven not solely by automated bots, but largely by human users' tendency to share novel or sensational claims. The spreading of misinformation poses serious threats: it can sway political opinions (Sobkowicz et al., 2012; Tucker et al., 2018), undermine public health measures (Do Nascimento et al., 2022; Suarez-Lledo and Alvarez-Galvez, 2021), and erode trust in institutions (Ognyanova et al., 2020).

Public trust in online information has been shaken. Surveys indicate high levels of skepticism toward news encountered on social media. A 2025 study (Talker Research, 2025) of 2,000 U.S. adults revealed that only 41% of online content was

viewed as completely accurate, 23% as completely false and purposely inaccurate or misleading, and 36% as somewhere in between. Additionally, 75% of the participants reported trusting the Internet less than ever before, and 78% agreed that the Internet has “never been worse” when it comes to distinguishing between what is real and what is artificial. Globally, trust in media varies significantly, ranging from approximately 19% to 69% across countries, with a median around 38.5%, according to the Reuters Institute’s 2023 Digital News Report (Reuters, 2024).

One reason misinformation is pervasive is the sheer volume of “factual” content being shared. Online discussions are replete with **factual claims**, which we define as assertions that could be either true or false. These range from simple statements of fact to complex narratives. Many web users struggle to distinguish truth from falsehood, inadvertently amplifying unverified or false factual claims. In this context, it is critical to develop tools for gauging the veracity of widely circulated claims and for understanding how people react to them. Before introducing the concept of truthfulness stance, which we propose as a means of understanding public reactions to factual claims, we first examine AI-driven approaches to combating misinformation through automated fact-checking.

1.1.2 On Combating Misinformation

Professional fact-checking has emerged as a crucial response to the challenge of misinformation. This process involves systematically scrutinizing and verifying the veracity of factual claims using authoritative evidence (Graves, 2017). However, the manual nature of traditional fact-checking means it cannot scale to address the sheer volume and velocity of misinformation circulating on social media platforms (Graves, 2018). A single viral rumor can be seen by millions within hours, far outpacing what human fact-checkers can verify in that time. This limitation has spurred growing

interest in automated fact-checking—developing computational methods to identify, verify, and contextualize dubious claims at scale (Vosoughi et al., 2018). Such approaches are not intended to replace human fact-checkers, but to complement them, and have become a core focus of efforts to combat online misinformation. An ideal automated fact-checking pipeline is commonly conceptualized (Guo et al., 2022) as comprising several key stages, as follows.

Claim Check-Worthiness Detection. This stage identifies which claims are worth fact-checking among the vast amount of online content. It relies on the notion of “check-worthiness,” i.e., the likelihood that verifying a claim would be of interest to the public or important for accountability (Hassan et al., 2015). In practice, systems either make a binary decision or rank claims by priority (Hassan et al., 2017; Nakov et al., 2018; Mansour et al., 2023). Prior work has typically focused on factual statements that can be verified, avoiding content that is subjective or based solely on personal experience (Konstantinovskiy et al., 2021; Ni et al., 2024). By filtering out non-check-worthy content, this stage helps ensure that limited fact-checking resources focus on the most consequential claims (Zeng et al., 2021; Guo et al., 2022; Das et al., 2023).

Evidence Retrieval. Once a claim is selected for checking, the next stage is to gather relevant evidence from trusted sources that can support or refute it. Early automated systems sometimes skipped this stage, attempting to predict veracity from the claim text alone (Wang, 2017), but relying solely on a claim’s wording is often insufficient to identify cleverly crafted misinformation. Modern systems therefore rely on retrieved documents, databases, or other information sources (e.g., news articles, Wikipedia entries) that provide factual context for a claim. Such evidence is essential

because it grounds verification in real-world information and enables the system to back up its decision with facts (Thorne et al., 2018; Zellers et al., 2019).

Verdict Prediction. Given a claim and the collected evidence, the system predicts a verdict assessing the claim’s truthfulness. This is typically formulated as a classification task that assigns the claim a label such as “True” or “False” (or “Supported”/“Refuted” when emphasizing the evidence basis) (Thorne et al., 2018). Some approaches use binary or three-class verdicts (e.g., “True,” “False,” or “Not Enough Info” if evidence is insufficient), while others adopt fine-grained multi-class scales (akin to journalistic ratings such as “mostly true” or “pants on fire”) (Wang, 2017; Shahi and Nandini, 2020).

Justification Production. The final stage generates a justification explaining how the system reached its verdict. In professional fact-checking, providing a rationale is vital: simply declaring a claim true or false without explanation often fails to convince the audience and can even backfire by entrenching false beliefs (Lewandowsky et al., 2012). Similarly, an automated system should explain its reasoning both to persuade users and to maintain transparency, especially since machine learning models can be “black boxes” (Lipton, 2018). Approaches to justification production include highlighting the key evidence that led to the verdict (e.g., using attention weights to show which facts were considered most relevant) (Popat et al., 2018), using logical or rule-based inference methods that produce human-readable reasoning steps (Gad-Elrab et al., 2019), or generating a natural-language explanation that summarizes how the evidence supports the verdict (Atanasova, 2024).

Among these stages, stance detection is often treated as part of the evidence retrieval stage in automated fact-checking pipelines (Ferreira and Vlachos, 2016). The rationale is that analyzing how people respond to a claim, or what stance news articles

adopt, can serve as a proxy for truth assessment (Kou et al., 2021; Saeed et al., 2022). In other words, the collective judgment of the crowd or various information sources may indicate a claim’s veracity (Xu et al., 2021).

However, most existing stance detection research, especially before 2020, did not explicitly detect the truthfulness stance of social media posts toward general factual claims. Instead, they typically examined the stance of news articles toward rumors for fact-checking purposes. The distinction is twofold: (1) **Source type**: social media posts are authored by a diverse set of individuals with varying credibility, whereas news articles are produced by professional outlets or journalists, often following editorial norms; (2) **Target type**: general factual claims can be any verifiable statement on a wide range of topics, while rumors in this context usually refer to unverified or disputed stories circulating at a specific time. This gap in the literature motivated our work. We view stance detection as a social sensing approach to map public perception of factual claims, encompassing both misinformation and verified facts. In this dissertation, online users are treated as “human sensors” whose collective reactions provide signals about the influence and dissemination of factual claims. Understanding how such claims spread and how the public receives them is essential. If a false claim goes viral and most users respond without skepticism, that indicates failures in information discernment; conversely, if a verified fact is met with widespread disbelief, that signals an underlying issue of trust. The following section presents a detailed discussion of the truthfulness stance.

1.2 Truthfulness Stance: Gauging Belief in Claims

As noted in Section 1.1, online information provides a valuable lens for gauging people’s perceptions and opinions, offering insights into societal trends, beliefs, and behaviors that shape human society (Sobkowicz et al., 2012; Zhang et al., 2018;

Claim: “California introduces new bill that would allow mothers to kill their babies up to 7 days after birth.”

 Lee @HistoryDayNow Follow ... <p>This is beyond sick, it's called murder.</p> <p>默默地哭</p> <p>California introduces new bill that would allow mothers to kill their babies up to 7 days after birth - Miami Standard</p>	 Yvette @Yvettewrites Follow ... <p>Is this really true ? I know that California is insane in many ways , but this ?</p> <p>California introduces new bill that would allow mothers to kill their babies up to 7 days after birth -- Society's Child --</p>
<p>Truthfulness Stance: Positive</p>	<p>Truthfulness Stance: Neutral</p>
<p>Truthfulness Stance: Negative</p>	<p>Truthfulness Stance: No Stance</p>

Figure 1.1: Four tweets expressing different truthfulness stances toward the same factual claim.

Willaert et al., 2020). This dissertation focuses on the concept of *truthfulness stance* which, given a factual claim, assesses whether a textual utterance affirms its truth, disputes it as false, or expresses a neutral or indeterminate position. Specifically, this dissertation examines social media posts, focusing on tweets from Twitter (now rebranded as X) as the primary form of textual utterance.

Figure 1.1 presents examples of tweets that express positive, neutral, negative, or no stance regarding the truthfulness of the same factual claim. For the claim “California introduces new bill that would allow mothers to kill their babies up to 7 days after birth,” these social media posts respond differently, as follows.

- The first tweet expresses outrage, stating “This is beyond sick, it’s called murder.” This strong emotional reaction suggests that the user believes the claim is true—that California did introduce such a bill.
- The second tweet questions the claim by asking, “Is this really true?” This expression of uncertainty indicates that the user is unsure whether California actually introduced such a bill.
- The third tweet explicitly refutes the claim, stating that “Assembly Bill 223 does not legalize infanticide.” This act of correction implies that the user believes the claim is false.
- The last tweet states, “they want to pass a bill,” which clearly relates to the claim’s topic, but the post does not explicitly express a belief about whether the bill has been introduced or whether the claim is true or false. Therefore, no clear stance is conveyed.

Truthfulness stance can be a useful tool in discerning how misinformation spreads (Ecker et al., 2022; Allcott et al., 2019) and shapes decision-making in political discourse (Ognyanova et al., 2020; Yan et al., 2025), economic decisions (Ausat, 2023), and health-related contexts (Suarez-Lledo and Alvarez-Galvez, 2021). Such insights can help social scientists assess the impact of misinformation and develop effective interventions (Watts et al., 2021). Additionally, health organizations can utilize this information to gauge public opinion and identify communities in specific geographic regions that may be more susceptible to health-related misinformation (Loomba et al., 2021; Zhu et al., 2021). Truthfulness stance can also aid marketers and media strategists in evaluating the effectiveness of their campaigns (Dwivedi et al., 2021) and tracking shifts in public perception regarding a product or political figure (Dimitrova and Matthes, 2018).

As noted in Section 1.1.2, definitions of stance vary across studies. Despite these differences, they share a common conceptual framework, wherein a declared stance comprises four components: a textual **utterance** expressing the stance (e.g., a news article or a social media post), a **target** that receives the stance (e.g., an entity, a topic, an event, or a factual claim), the **orientation** of the stance (e.g., positive, neutral, or negative), and the **type of stance**, which specifies what the stance is about (e.g., favorability toward the target entity, the likelihood of an event, or the target claim’s truthfulness). Chapter 2 presents this conceptual framework of stance definitions in greater detail. While prior studies (Küçük and Can, 2020; Hardalov et al., 2022a; Alturayeif et al., 2023) have addressed various aspects of stance, our systematic articulation of the conceptual framework represents a significant contribution to the field. This level of nuance and fine-grained analysis has been largely absent from the literature. Regarding the specific definition of stance examined in this dissertation, ours is the first to focus on the stance of social media posts regarding the truthfulness of *general* claims.

1.3 The TSD-CT Dataset: A Benchmark for Truthfulness Stance Detection

Most existing datasets define stance in terms of favorability toward certain entities, such as political figures (Mohammad et al., 2016; Li et al., 2021), policies (Aleksandric et al., 2024), and social issues (Allaway and McKeown, 2020). 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 (Hossain et al., 2020)), whereas others focus on news articles rather than social media posts (Pomerleau and Rao, 2017; Ferreira and Vlachos, 2016). To address these gaps, Chapter 3 introduces TSD-CT (Truthfulness Stance Detection–Claim and Tweet), a new dataset constructed

using our in-house annotation interface. The dataset consists of 5,331 claim-tweet pairs labeled with five classes: 2,104 *positive stance* (\oplus), 882 *neutral/no stance* (\odot), 883 *negative stance* (\ominus), 309 *different topics* (*dif*), and 1,153 *problematic pairs* (*prb*). The claims are sourced from PolitiFact (<https://www.politifact.com/>), and the tweets were retrieved via the Twitter API v2. To ensure label reliability, we implemented a 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 0.803 and a Krippendorff’s alpha of 0.699, indicating both robustness of the annotation process and the reliability of labels. An error analysis of high-quality annotators on screening pairs revealed an overall error rate of 14.59%. The most frequent mistakes occurred when annotators selected the \odot class, particularly in cases where the correct label was one of \oplus , *dif*, or *prb*.

TSD-CT serves as a benchmark dataset for training and evaluating machine learning models that detect the truthfulness stance of tweets toward factual claims. With its high-quality annotations and broad topical coverage, TSD-CT can be used to fine-tune transformer-based models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), or to instruct-tune large language models (LLMs) (Achiam et al., 2023; Touvron et al., 2024; Team et al., 2023). Beyond model training, TSD-CT is a valuable resource for applications that study the spread and impact of misinformation on social media (Zhu et al., 2021, 2025a). These applications support investigations into online polarization and help inform the design of effective strategies to mitigate misinformation. We also provide the web-based annotation interface for truthfulness stance labeling, publicly available at https://idir.uta.edu/stance_annotation, along with an open-source codebase (<https://github.com/idirlab/stancedatacollection>) to support reproducibility

and enable other researchers to easily customize their annotation tasks. In line with the FAIR principles (Wilkinson et al., 2016), the TSD-CT dataset is published under a CC BY 4.0 license and is available in standardized formats with detailed documentation at Zenodo (<https://doi.org/10.5281/zenodo.15620262>).

1.4 The RATSD Framework: Retrieval-Augmented Stance Detection

Chapter 5 provides an overview of baseline approaches and state-of-the-art methods for stance detection, and introduces a novel framework for truthfulness stance detection. This framework is designed to identify the orientation of the stance expressed in a given tweet regarding the truthfulness of a specific claim. Particularly, we have developed an LLM-empowered framework **RATSD** (**r**etrieval **a**ugmented **t**ruthfulness **s**tance **d**etection). The framework employs LLMs, including open-source models such as Zephyr (Tunstall et al., 2023) and proprietary models such as GPT-3.5 (Achiam et al., 2023), for three purposes. *First*, RATSD generates contextual knowledge related to factual claims and tweets using the approach of retrieval augmented generation (RAG) (Lewis et al., 2020). Incorporating contextual knowledge enables the framework to access relevant, up-to-date information, thereby enhancing the accuracy and contextual awareness of stance detection models. *Second*, RATSD produces stance analyses—narratives of tweets’ stance regarding claims—by prompting LLMs with the generated contextual knowledge. This directly leverages LLMs’ reasoning capability in analyzing tweets’ stance, incorporating contextual information into the learning process. *Third*, RATSD includes a classifier based on a fine-tuned language model which, taking a claim, the tweet analysis and the contextual knowledge as input, returns as the classification label the orientation of the tweet’s truthfulness stance toward the claim. To our knowledge, this work pioneers the application of RAG to stance detection, demonstrating the utility of contextual knowledge for the

task. To evaluate the effectiveness of RATSD, Chapter 6 discusses our experiments with RATSD on TSD-CT and three existing benchmark datasets, SemEval-2019 (Gorrell et al., 2019), WT-WT (Conforti et al., 2020), and COVIDLies (Hossain et al., 2020), in varying experimental settings and choices of LLMs. The results show that RATSD utilizing GPT-3.5 outperformed the state-of-the-art models (Reddy et al., 2022; Arakelyan et al., 2023), achieving a 6.38 point increase in Macro-F1 score on TSD-CT. Our ablation study revealed that contextual knowledge and stance analyses play pivotal roles in the models’ performance.

1.5 Real World Applications

To demonstrate the practical impact of this research, we developed three key real-world applications as detailed in Chapter 7: (1) **TrustMap**, a comprehensive truthfulness stance mapping tool; (2) **ClimateStance**, a climate change stance analysis framework; and (3) a COVID-19 misinformation dashboard. These applications showcase how the methods and insights from the dissertation can be applied across different domains and use cases, from general social sensing applications to domain-specific belief tracking, thereby bridging the gap between research and practice.

TrustMap. TrustMap is our most comprehensive application. It visualizes the aggregate truthfulness stance of social media posts toward factual claims across U.S. regions. Building on RATSD, TrustMap integrates stance detection with geospatial analysis, enabling users to explore stance patterns by claim and location and to gain insights into how narratives and misinformation spread across communities. To make the results more interpretable for non-technical audiences, TrustMap includes: (1) Stance explanations: Explanations for each claim–tweet pair, which are short, LLM-generated rationales describing why the post was assigned its stance label; (2) Stance

distribution summaries: LLM-generated overviews that describe the overall stance breakdown for user-selected claims.

ClimateStance. ClimateStance is a stance detection framework for climate change-related discourse. It collects fact-checked climate change-related claims and relevant social media discussions, then employs an LLM to organize these claims into a topical taxonomy of climate-related issues and classify the tweets’ stances regarding their corresponding claims. By analyzing truthfulness stance, the framework provides insights into public perceptions of climate-related issues. Our empirical analysis revealed that social media users often accept claims as true regardless of the claims’ actual veracity, and that the public struggles to distinguish facts from misinformation in climate-related topics such as politics, economy, and environment. These findings underscore the broader challenges of misinformation in the climate domain and demonstrate the value of stance detection in identifying gaps in public discernment.

COVID-19 Misinformation Dashboard. The COVID-19 misinformation dashboard focuses on a critical health domain. This public dashboard combines stance detection with semantic claim matching to monitor the “infodemic” during the COVID-19 pandemic. It uses a curated catalog of verified facts and debunked misinformation about COVID-19, matches relevant tweets to these known claims using BERT-based semantic similarity, and then detects each tweet’s stance toward the claim. The dashboard features an interactive map and an analytics panel that highlight prevalent misinformation or facts in selected U.S. regions. By integrating automated claim–tweet matching with stance classification, it enables public health officials and researchers to understand the spread of false COVID-19 narratives over time and geography.

1.6 Significance and Outlook

The development of the TSD-CT dataset, the RATSD framework, and the real-world applications presented in this dissertation form a toolkit for understanding and combating online misinformation. By focusing on how people react to claims, we introduce a new dimension to both social sensing and fact-checking research. This work is significant in several ways:

Advancing Theoretical Understanding. We offer a refined definition of stance in the context of truthfulness, situating it within a general stance taxonomy. This theoretical contribution clarifies how truthfulness stance detection relates to, yet differs from, traditional stance detection or sentiment analysis tasks. It lays the groundwork for future research to more precisely define the stance type under study.

Data Resource for the Community. The TSD-CT dataset is a significant resource contribution. It addresses the scarcity of well-annotated data for truthfulness stance on social media. We anticipate that TSD-CT will become a benchmark dataset for this niche, enabling head-to-head comparisons of models and stimulating research in both computational journalism and computational social science on misinformation-related topics. Researchers may also extend TSD-CT or create analogous datasets for other languages or platforms by following the data collection and annotation methodology described in this work.

Innovative Methodology. The RATSD approach demonstrates a novel way to combine retrieval, generation, and classification for stance detection. Beyond our specific task, it suggests that combining external knowledge with LLM reasoning may substantially improve performance in other social media content analysis problems (e.g., detecting hate speech context, identifying sarcasm). Demonstrating RATSD’s

effectiveness highlights a path for integrating LLMs into misinformation mitigation in a controlled, evidence-based manner.

Practical Tools and Applications. Ultimately, our research has significant practical value. The ability to automatically gauge truthfulness stance enables a range of applications, as demonstrated in this dissertation through several real-world case studies. These examples illustrate the versatility of truthfulness stance detection, from enhancing situational awareness for journalists, to aiding sociologists in measuring polarization, to empowering platforms and policymakers to identify communities at risk of influence by false narratives.

In closing this introduction, we emphasize that tackling misinformation is a multifaceted challenge; no single technique—whether human or automated—can suffice on its own. Our work on truthfulness stance adds a vital piece to this puzzle by illuminating the reception dimension of misinformation. When integrated with claim verification and other components, it brings us closer to building comprehensive automated fact-checking systems that can operate at scale and speed while still engaging meaningfully with human judgments, such as those expressed in stances and fact-checks. Through this work, we aim to demonstrate that mitigating misinformation requires understanding not only what is true or false, but also what people believe to be true or false. By developing datasets and models for detecting truthfulness stance, we take a substantial step toward that goal, contributing both knowledge and tools to support researchers and practitioners striving to uphold truth in online discourse.

CHAPTER 2

CONCEPTUAL FRAMEWORK OF TRUTHFULNESS STANCE DETECTION

The task of stance detection has garnered significant interest in natural language processing, but its definition and scope vary widely across studies. Prior surveys of stance detection (Alturayeif et al., 2023; Hardalov et al., 2022b; Küçük and Can, 2020) have examined various aspects of the concept of stance detection, yet none have provided a fine-grained conceptual framework to unify the different types of stance considered in the literature. This lack of a unified framework represents a theoretical gap: researchers often use the term “stance” to mean different things, ranging from favoring or opposing a target to believing or disbelieving a claim’s veracity. In particular, the concept of truthfulness stance—whether an utterance conveys that a claim is true, false, or neither—has not been explicitly distinguished from other stance types in most prior work.

There is thus a clear theoretical and practical motivation to delineate truthfulness stance as a distinct category of stance. First, from a theoretical standpoint, failing to distinguish between stance types—favorability, likelihood, and truthfulness—can lead to confusion. A person’s stance can concern different attributes of the target. For example, one might oppose a policy (negative favorability stance) while still believing it to be true (positive truthfulness stance), as illustrated in the first example in Figure 1.1. The literature so far has not systematically categorized these differences. This dissertation addresses this gap by providing the first in-depth exploration of truthfulness stance. The next section lays out the conceptual foundations of truthfulness stance detection, clarifying how we define stance in this work and how

it compares to previous formulations in the literature. Let t denote a social media tweet (e.g., a tweet) and c denote a factual claim. Together, t and c constitute a claim-tweet pair. A **positive** stance applies when t conveys the belief that c is true. A **negative** stance indicates that t believes c is false. A **neutral/no stance** signifies that t either expresses uncertainty about the truthfulness of c (**neutral**) or does not explicitly take a position on c 's truthfulness, even though both t and c discuss the same topic (**no stance**).

In our framework, a stance is characterized by four components: (1) the utterance that expresses the stance (e.g., a tweet or a sentence in a news article), (2) the target of the stance, which is the entity, topic, event, or claim that the stance is about, (3) the orientation of the stance (the evaluative label such as positive, negative, or neutral/no stance toward the target), and (4) the type of stance, which specifies in what sense the stance is being expressed (for instance, favorability, likelihood, or truthfulness). This conceptual breakdown builds upon prior work (Alturayeif et al., 2023; Hardalov et al., 2022b; Küçük and Can, 2020; Zhu et al., 2022) but extends it by explicitly distinguishing the different stance types and their implications. Below, we discuss each component in detail and illustrate how truthfulness stance fits into and differs from existing stance definitions.

2.1 Orientation of Stance

Figure 2.1 illustrates the relationship among all stance *orientation* labels. Note that our truthfulness stance detection model does not consider unrelated pairs, because detecting the relevance between c and t pertains to research areas such as textual semantic similarity (Wang and Dong, 2020; Gomaa et al., 2013), which is beyond the scope of this dissertation. We did annotate unrelated pairs while creating

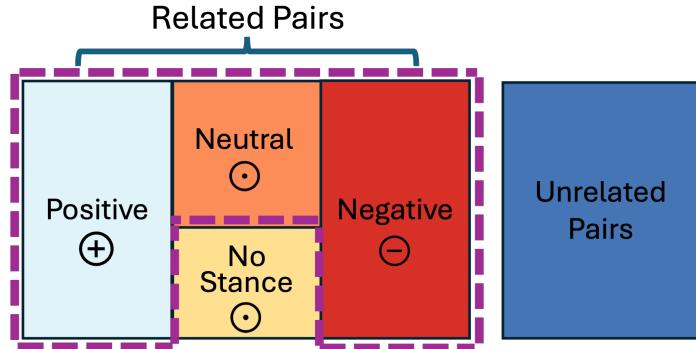


Figure 2.1: A claim c and a tweet t may be related or unrelated. Related claim-tweet pairs are partitioned into four cases of stance. The cases within the dashed, purple frame (positive, neutral, negative) are applicable when t expresses a stance regarding c 's truthfulness. Though conceptually different, for practical purposes neutral and no stance are combined into one class \odot in our consideration.

TSD-CT, though, in order to exclude such pairs in training and evaluating detection models.

Conceptually, we recognize the difference between a neutral stance and no stance. A tweet holds a neutral stance if it expresses a mixed verdict or uncertainty about a claim's truthfulness. On the other hand, a tweet has no stance if, while being related to the claim in terms of topic, it does not express an intentional stance reflecting beliefs (Dretske, 1988) regarding the claim's truthfulness. This distinction is similarly recognized in some existing studies, such as SemEval-19 and (Grimminger and Klinger, 2021), though they use different terminology for stance labels.

In practice, though, discerning no stance is highly challenging. Example (3) in Figure 2.2 demonstrates one such challenging case. Further explanations of the remaining challenging examples are provided in Appendix A.2. Although the tweet is highly pertinent to the claim, as it mentions Paul Ryan, gun laws, and “action,” it does not indicate whether Paul Ryan has blocked such actions or not. Its stance is not neutral; rather, it does not express any stance on the claim’s truthfulness. Neutral

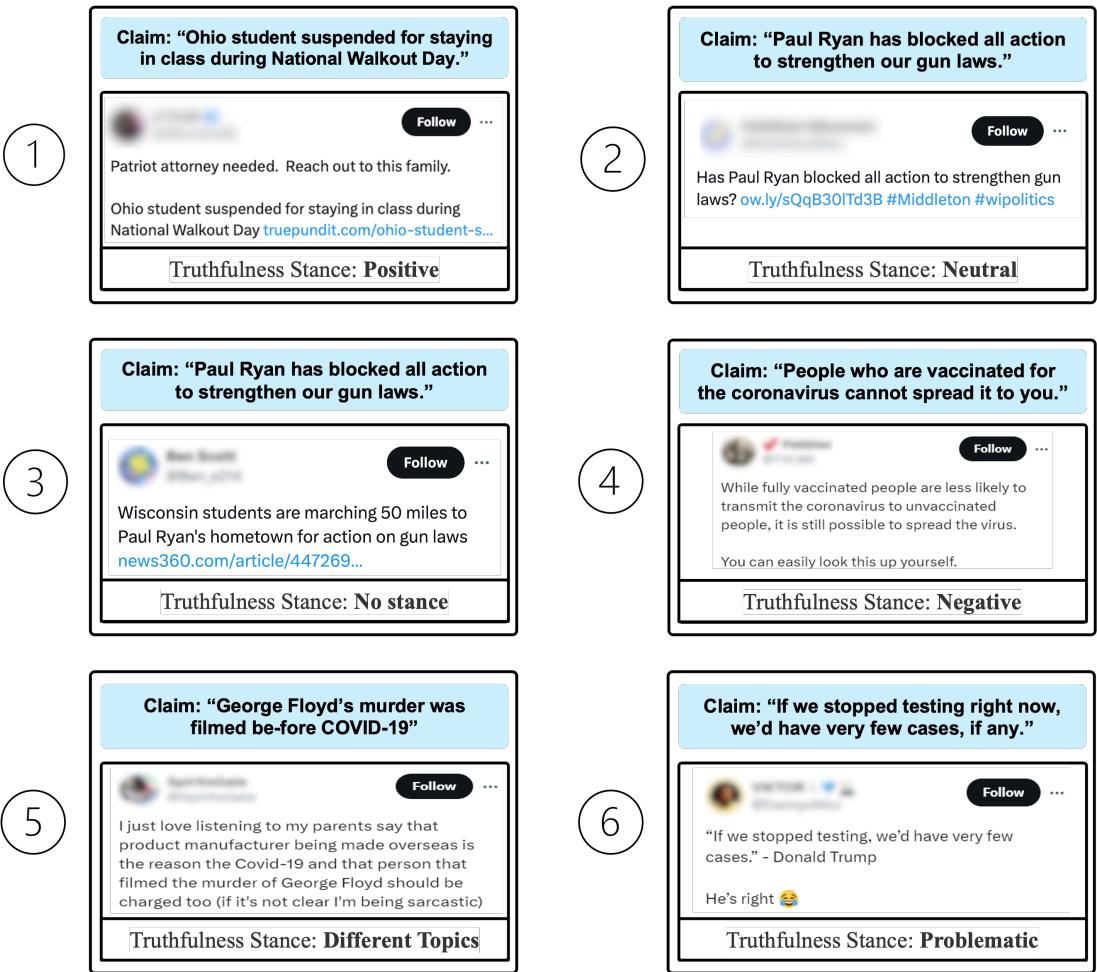


Figure 2.2: A few samples of challenging cases in truthfulness stance annotation.

stance and no stance often exhibit strong similarities. This difficulty is evident in our preparation of the TSD-CT dataset where, among all pairs of stance labels, the (neutral stance, no stance) pair received the lowest inter-annotator agreement among expert annotators.

Given this intrinsic challenge, we chose to merge neutral stance and no stance into a single class \odot for both dataset creation and detection model development. A similar approach was used in sentiment analysis, where Koppel and Schler (2006) categorized documents' neutrality sentiment into two types. The first type of neu-

trality sentiment (analogous to no stance in our framework) applies to documents that present objective information without expressing a clear sentiment. The second type (akin to neutral stance) applies to documents that convey a mix of positive and negative sentiment.

Type of stance	Target of stance			
	Entities or Topics	Events or Rumors	Fact Triples	Factual Claims
Favorability	SemEval-2016 (Mohammad et al., 2016); VAST (Allaway and McKeown, 2020); P-Stance (Li et al., 2021); (Grimminger and Klinger, 2021); (Aleksandric et al., 2024)	MGTAB (Shi et al., 2023)		
Likelihood		WT-WT (Conforti et al., 2020)		
Truthfulness	PHEME (Zubiaga et al., 2016); SemEval-2017 (Derczynski et al., 2017); SemEval-2019 (Gorrell et al., 2019)	NewsClaims (Reddy et al., 2022); Fact-Bank (Saurí and Pustejovsky, 2009); Diab et al., 2009	Emergent (Ferreira and Vlachos, 2016); FNC-1 (Pomerleau and Rao, 2017); COVIDLies (Hossain et al., 2020); This work (TSD-CT)	

Table 2.1: Various definitions of stance differ in the type, utterance, and target of stance.

2.2 Utterance of Stance

Table 2.1 compares the definitions of stance across existing datasets, listing dataset names (if available) and their corresponding references. Our ensuing discussion refers to these names whenever applicable. Researchers have developed various methods and models for these datasets and their respective stance detection tasks. Such models are referenced throughout this paper but not necessarily in Table 2.1.

To distinguish between *utterances* in existing definitions and our own, Table 2.1 uses two colors—brown for news articles and blue for social media posts (primarily tweets, though SemEval-2019 includes Reddit posts and VAST considers comments on news websites). In stance detection, the prevalence of informal language traits, such as slang, abbreviations, and misspellings, poses greater challenges (Al Qundus et al., 2020; Smirnov, 2017) compared to news articles, which predominantly adhere to formal language conventions.

2.3 Target of Stance

Table 2.1 identifies four primary types of stance *targets* in prior studies: 1) entities (e.g., Hillary Clinton) and topics (e.g., “legalization of abortion”) in SemEval-2016, VAST, P-Stance, (Grimminger and Klinger, 2021), and (Aleksandric et al., 2024); 2) events (e.g., mergers and acquisitions of companies in WT-WT and Japan’s nuclear wastewater release in MGTAB) and rumors—true or false eventually—in PHEME, SemEval-2017 and SemEval-2019 (e.g., the rumor about a second shooter in the 2014 Parliament Hill shootings in Ottawa); 3) factual claims (e.g., news claims in Emergent, news headlines in FNC-1, and COVID-19 related misconceptions in COVIDLies); and 4) fact triples (i.e., subject-predicate-object triples) extracted from the utterance itself. For example, in NewsClaims and FactBank (Saurí and Pustejovsky, 2009), the stance is about whether an utterance affirms or refutes a particular fact triple, e.g., (Vitamin C, cure, COVID-19 virus). Similarly, Diab et al. (2009) explored *committed belief* by evaluating whether a writer conveys belief in the truth of a fact triple, such as (GM, layoff, workers), within their utterance.

Note that the datasets also vary in the number of targets, ranging from fewer than ten to several thousand. Some datasets have a small number of targets, such as MGTAB (1 target), P-Stance and (Grimminger and Klinger, 2021) (3 targets), WT-

WT (5 targets), and SemEval-2016 (6 targets). Others contain dozens to hundreds of targets, including COVIDLies (86 misconceptions), Emergent (300 news claims), NewsClaims (889 fact triples), and PHEME, SemEval-2017 and SemEval-201 (several hundred latent rumors as their source tweets, each mentioning a rumor, are on such a scale). Finally, some datasets feature thousands of targets, such as our TSD-CT (2,201 factual claims), FNC-1 (2,542 news headlines), FactBank (4,801 fact triples), and VAST (5,634 topics).

2.4 Type of Stance

The *type* of stance in various existing definitions falls into three main categories: 1) likelihood of target events occurring (e.g., WT-WT); 2) favorability—determining whether the stance expressed in an utterance is in favor of or against a given target (e.g., SemEval-2016, VAST, P-Stance, (Grimminger and Klinger, 2021), (Aleksandric et al., 2024), and MGTAB); 3) the truthfulness of a rumor (PHEME, SemEval-2017 and SemEval-2019), a news headline (FNC-1), a fact triple (NewsClaims, FactBank and (Diab et al., 2009)), or a claim (Emergent and COVIDLies). These stance types are not equivalent and therefore require distinct detection models. This is clearly illustrated by the upper-left example in Figure 1.1—the tweet conveys a negative favorability stance but a positive truthfulness stance toward the claim.

Among the aforementioned datasets and definitions of stance, TSD-CT most closely resembles COVIDLies, as both focus on tweets’ stance toward the truthfulness of factual claims. One key distinction is that COVIDLies exclusively focuses on COVID-19-related misconception claims, which were manually examined and rephrased and tend to be simple and short. On the contrary, the claims in TSD-CT, sourced from Politifact, are more complex both syntactically and semantically, covering a broader range of topics relevant to fact-checkers. It is also worth noting

that COVIDLies only includes false claims as targets, whereas TSD-CT contains a mix of true and false claims.

In summary, this chapter established the motivation for investigating truthfulness stance and introduced a conceptual framework delineating its key components. We highlighted how truthfulness stance is defined in relation to other stance types and why this distinction matters theoretically and for misinformation research. In the following chapters, we build on this foundation: Chapter 3 will describe the creation of a dataset for truthfulness stance detection, and Chapter 4 will delve into the analysis of stance data, while Chapter 5 will present models for automatically detecting truthfulness stance.

CHAPTER 3

THE TRUTHFULNESS STANCE DETECTION DATASET

As discussed in Chapter 2, the stance definition central to this dissertation differs from previous definitions in existing literature. Consequently, it became essential to construct a new dataset tailored to the definition of *truthfulness stance* presented in this research.

Given that the **target** of our truthfulness stance detection framework is factual claims, it was crucial to select high-quality, widely discussed claims prevalent on social media platforms. Fact-checks provide a reliable and rich source of such claims, and among various fact-checking websites, PolitiFact (<https://www.politifact.com/>) is recognized as one of the most comprehensive and trusted platforms available (see the statistics in Table 3.1). Considering the constraints of limited human annotation resources, we strategically chose PolitiFact as the primary source for factual claims utilized in our annotation process. It is important to highlight that our data collection effort extended beyond merely collecting factual claims from PolitiFact. To facilitate advanced model development discussed in Chapter 5, we also systematically gathered comprehensive contextual knowledge, including fact-check articles, summaries, claimants, and other metadata from seven prominent fact-checking websites.

Regarding the **utterances**, we specifically selected tweets from X (formerly Twitter), due to its status as the most widely used and data-rich social media platform. Notably, X provides extensive user-generated content and, crucially, detailed geolocation information, which is essential for applications such as TrustMap, as described in Chapter 7. Such geolocation data is typically unavailable or significantly

limited on other social media platforms such as Facebook (<http://facebook.com/>) and Bluesky (<https://bsky.app/>).

The resulting dataset, named TSD-CT (Truthfulness Stance Detection–Claim and Tweet), consists of curated pairs of factual claims and corresponding tweets. The following sections detail the processes involved in the creation and structuring of the TSD-CT dataset.

3.1 Fact-Check Collection

We developed a web-scraping 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. The collected data is coded using ClaimReview’s data schema, a widely adopted standard for structuring fact-checks. It enables interoperability between different platforms, making it easier to share, search, and analyze fact-checking information across the web. The data schema includes fields such as “Publisher,” “Claim Reviewed,” “Review Summary,” “Review,” “Verdict,” “Author,” “Claim Published Date,” “Fact-check Published Date,” “Thumbnail URL,” “URL,” and “Tags.” The details of the meaning of each field can be found in Appendix A.3. An example of a PolitiFact fact-check is illustrated in Figure 3.1.

Table 3.1 summarizes the field-level statistics of our collected fact-check dataset across seven fact-checking websites. The structure and completeness of data fields vary notably across sources. PolitiFact maintains a highly structured format with the most comprehensive coverage across all fields, including 25,660 claims, 25,663 reviews, and 25,694 URLs, and assigns verdict labels to nearly all entries. Snopes provides a large number of factual claims (17,925) and reviews (19,128), along with verdicts such as “True,” “Mostly True,” “Mixture,” “Unproven,” “False,” “Scam,” and

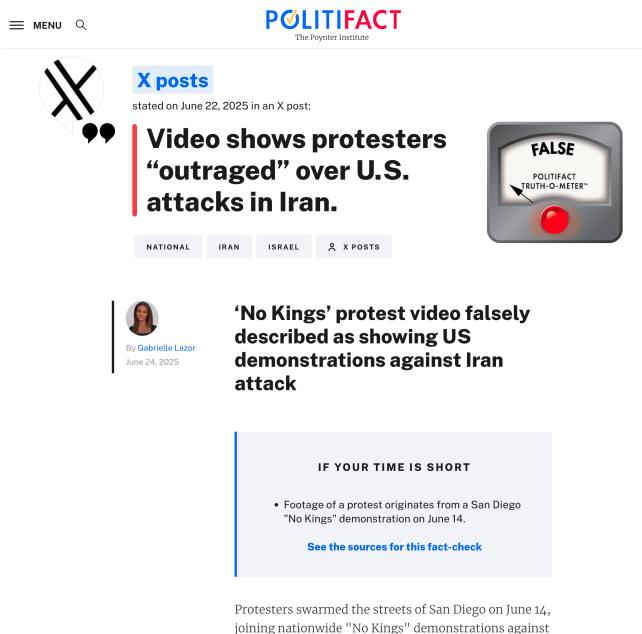


Figure 3.1: An example of Politifact fact-check.

others. However, it includes relatively few verdict labels (45) and summaries (2,638). Metafact differs from traditional outlets by allowing users to submit questions, which are converted into affirmative claims and verified by domain experts. Although it includes 3,404 full reviews, it includes five types of verdicts in total and does not provide summaries. FullFact, managed by independent fact-checkers, includes substantial review content (3,995) and summaries (3,924), but lacks formal one-word verdict labels, instead favoring narrative assessments. FactCheck.org also lacks discrete claim and verdict fields, offering instead comprehensive reviews (3,696) and tags (2,881) to aid interpretation. AFP Fact Check and FactCheck.org lack explicit claim fields altogether and do not assign verdicts. AFP includes extensive reviews (8,160) and images (8,183), yet provides no discrete verdict labels. AP Fact Check, which began using structured formats around mid-2021, includes both claims (2,669) and verdicts (26), but many entries still lack consistent tagging or summary fields.

Across all sources, we collected a total of 63,281 fact-check articles with accessible URLs. PolitiFact contributes the most with 25,694 entries, followed by Snopes (18,996), AFP Fact Check (8,176), and FullFact (3,995). The remaining sources (FactCheck.org, AP Fact Check, and Metafact) provide 3,696, 2,825, and 889 articles, respectively.

Field	AFP Fact Check	AP Fact Check	FactCheck.org	FullFact	Metafact	PolitiFact	Snopes
Claim	0*	2,669	0*	3,989	889	25,660	17,925
Review Summary	8,052	2,665	3,624	3,924	0*	25,578	2,638
Review	8,160	2,824	3,696	3,995	3,404	25,663	19,128
Verdict	0*	26	0*	0*	5	9	45
Author	0*	130	197	0*	2,139	4,788	19
Claim Published Date	0*	0*	0*	0*	310	5,654	0*
Factcheck Published Date	1,241	2,825	2,481	1,623	536	4,935	5,132
Image URL	8,183	0*	2,642	3,661	0*	8,425	13,398
Factcheck URL	8,176	2,825	3,696	3,995	889	25,694	18,996
Tags	65	87	2,881	96	240	4,987	81

Table 3.1: Field statistics from the fact-check collection. * Not all websites follow a consistent structure in their fact-checks. For instance, AFP Fact Check and FactCheck.org do not separately list the claims they fact-check.

3.2 Claim-Tweet Pair Collection

To construct the claim-tweet pair collection, we selected factual claims exclusively from PolitiFact, as it offers a comprehensive repository and aligns with our limited annotation capacity. Claims phrased as questions were excluded to ensure clarity and factual precision. We used spaCy to extract relevant keywords—including nouns, verbs, adjectives, pronouns, and numerical values—from each claim. These keywords were combined into conjunctive (ANDed) queries and used to retrieve related tweets through Twitter API v2. To increase the likelihood of obtaining tweets highly relevant to the original claim, we skipped any claim that produced fewer than five extracted keywords. Only English-language tweets were included, as our work does not address multilingual models. To promote linguistic diversity and avoid re-

dundancy, we excluded retweets, replies, and quoted tweets. Tweets containing fewer than 30 characters were also filtered out, as they typically lack meaningful semantic content. To preserve temporal relevance, tweet retrieval was limited to a window spanning from one month before to three months after the publication date of each claim. This process resulted in a total of 36,154 claim-tweet pairs.

3.3 Claim-Tweet Pair Sanitization

During the claim-tweet pair collection process, we encountered various types of tweets that were not suitable for truthfulness stance detection. To reduce redundant annotation effort and ensure dataset diversity, we applied both algorithmic similarity measures and heuristic rules to filter out such tweets.

One major category of exclusion was duplication, which manifested in two forms. First, tweets that are highly similar to other tweets which were already collected during the sequential data collection process. For instance, the new tweet “Israeli female pilot Sarah Ahronot has reportedly been captured in Iran.” was almost identical to an already collected one, “An Israeli female pilot Sarah Ahronot has been captured in Iran.” The difference lies only in minor tokens such as “an” and “reportedly,” making the duplicate unnecessary to annotate. Second, tweets that closely resembled their paired factual claims. For example, a tweet stating “BREAKING NEWS: Dolly Parton announces her refusal to celebrate Pride Month in June.” is almost a verbatim repetition of the factual claim “Dolly Parton announces her refusal to celebrate Pride Month in June.” Including such cases offers limited value for enriching the stance detection dataset while requiring repetitive annotation efforts.

To identify these redundant or near-duplicate pairs, we employed the Ratcliff-Obershelp algorithm (Black, 2021), which measures similarity based on the longest contiguous matching substrings. The similarity score is computed as:

$$D_{ro} = \frac{2 \times K_m}{|S_1| + |S_2|} \quad (3.1)$$

where K_m is the total number of matching characters (found by recursively identifying the longest common subsequences), and $|S_1|$, $|S_2|$ are the lengths of the two sequences. A score of 1.0 indicates identical strings, and 0.0 indicates no overlap.

For tweet-to-tweet and tweet-to-claim comparisons, we removed hyperlinks and hashtags before computing similarity scores with the same Ratcliff–Obershelp algorithm. We excluded pairs under two conditions: (1) the tweet was highly similar to its paired factual claim, with a similarity score above 0.9; or (2) the tweet was nearly identical to another tweet from a different pair, with a similarity score above 0.8. In the latter case, we retained the tweet collected earlier and discarded the duplicate.

Another minor category of exclusion was fact-checking tweets. Specifically, the tweets posted by fact-checking organizations that directly quoted or summarized fact-check content were removed. For example, a tweet such as “PolitiFact—Wisconsin Public Service Commission has approved rate hikes, but not \$1.9 billion. <https://www.politifact.com/factchecks/2023/dec/29/americans-prosperity/wisconsin-public-service-commission-has-approved-r/>” contains the name of the fact-checker and a direct link to the original article. These tweets are trivial to annotate, as their stance clearly aligns with the fact-check verdict. We excluded such tweets using the following heuristic: a tweet was removed if it contained the claim and any of the following—(1) it was posted by a known fact-checking account (e.g., @PolitiFact, @snopes, @factcheckdotorg), (2) it included a hyperlink to a fact-check article, or (3) it mentioned the name of a fact-checking organization.

In total, we removed 30,032 pairs due to duplication or high similarity (criteria 1 and 2), and 329 pairs based on the fact-checker heuristic (criterion 3). Starting with

36,154 claim–tweet pairs, this sanitization process resulted in 5,793 remaining pairs, comprising 2,283 unique factual claims and 5,793 tweets authored by 5,227 distinct Twitter accounts. The factual claims span a temporal range from November 15, 2007, to April 11, 2022.

CHAPTER 4

THE TRUTHFULNESS STANCE DATA ANNOTATION

After collecting the factual claims and their associated tweets for TSD-CT, the next crucial step is to obtain human annotations for the development and evaluation of truthfulness stance detection models. This chapter outlines our annotation process, encompassing annotator recruitment, interface design, quality control strategy, and an evaluation of the dataset’s quality.

4.1 Annotator Recruitment

Annotators were compensated for their time and effort to encourage careful and attentive work. To recruit annotators, we distributed flyers and sent emails with announcements across our university. All annotators 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. Annotators with high-quality performance could earn up to 20 U.S. cents per claim–tweet pair. In total, 214 individuals registered as annotators.

It is worth noting that all data collection and annotation procedures for the TSD-CT dataset were approved by the Institutional Review Board (IRB) at the University of Texas at Arlington (Protocol No. 2023-0093.2), reflecting our commitment to ethical research involving human subjects. IRB approval ensures that all procedures comply with established standards for research ethics and integrity.

4.2 Annotation Interface

We developed a web-based annotation interface for truthfulness stance labeling, available at https://idir.uta.edu/stance_annotation/. The interface was adapted from ClaimBuster’s data collection platform (Arslan et al., 2020). While much of the backend was retained, we introduced several enhancements to meet the specific requirements of our project. For instance, ClaimBuster supports labeling single sentences, whereas TSD-CT involves labeling claim–tweet pairs. Additionally, our quality control strategy differs, as ClaimBuster uses a three-class schema, while TSD-CT employs five. The remainder of this section outlines the user interface and its functions.

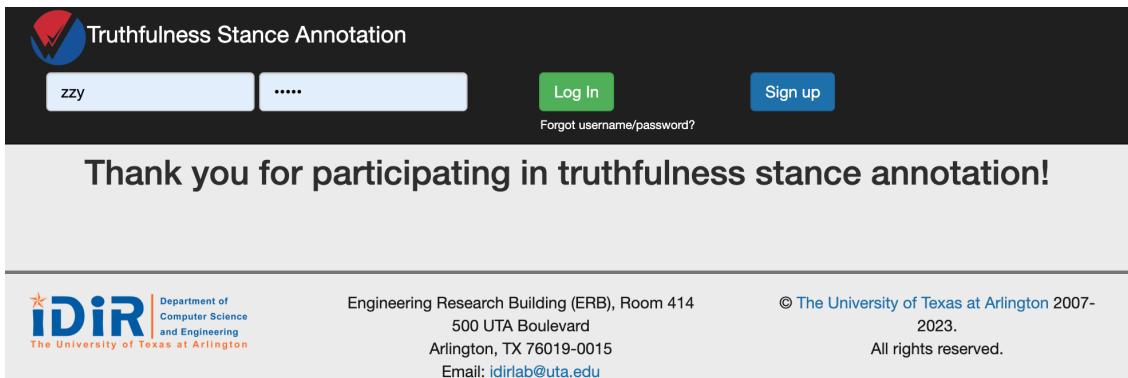


Figure 4.1: User login page of the annotation interface.

User Login and Registration. A screenshot of the user login page is shown in Figure 4.1. A new user can create an account by clicking the “Sign up” button in the top-right corner. This opens a registration form (Figure 4.2), where the user must provide their email address, username, and password. After submitting the form, a verification email is sent. The account is activated once the user clicks the verification link in the email.

The screenshot shows a modal window titled "New User Sign Up". It contains four input fields: "Email Address" (placeholder: "Enter Email Address"), "Username" (placeholder: "Enter Username"), "Password" (placeholder: "Enter Password"), and "Confirm Password" (placeholder: "Enter Password Again"). A blue "Submit" button is located at the bottom right of the form.

Figure 4.2: The registration form after clicking the “Sign up” button.

Annotation Instructions. When users log in for the first time after signing up, a consent form appears. This form provides information about the research project, including time commitment, compensation, and procedures. Users must click “Accept” to proceed. Upon acceptance, they are directed to the annotation instruction page, as shown in Figure 4.3. The instructions clearly define each stance class and include examples with explanations. The instructions also included edge cases and special scenarios. Annotators are expected to review each example carefully to understand the nuances between stance classes. The detailed examples can be found in Appendix A.1.

Before contributing to the main annotation task, each annotator was required to complete a training phase including 16 expert-annotated claim–tweet pairs. After they annotate each training pair, the interface immediately revealed the gold-standard label along with the corresponding explanation to help annotators better understand the annotation guidelines.

The screenshot shows the 'Truthfulness Stance Annotation' interface. At the top, there is a logo with a red and blue 'W' and the text 'Truthfulness Stance Annotation'. To the right are buttons for 'zzy labeled 2278 pairs', 'Leaderboard', 'Instructions', and 'Log Out'. Below this, the title 'Annotation Instructions' is displayed.

Annotation Instructions

Below is a screenshot of the annotation task. You will be shown a tweet-claim pair. In each pair, the claim on the left is fact-checked by PolitiFact, and the tweet on the right is displayed as it would appear on Twitter. Your task is to decide the truthfulness stance of the tweet towards the factual claim, i.e., whether the tweet believes the factual claim is true or false. Your five annotation options are as follows.

The screenshot shows the annotation task interface. On the left, a PolitiFact review for Rand Paul's statement is displayed. It includes the claim, claimant info, and a summary of the fact-check. On the right, a tweet from 'Augusta Chronicle' (@AUG_Chronicle) is shown. The tweet discusses COVID-19 vaccinations and boosters. Below the tweet, hyperlinks and titles are listed. At the bottom, a question asks 'What is the truthfulness stance of the tweet towards the factual claim?' followed by five annotation options (radio buttons) and a 'Submit' button.

POLITIFACT
Go to the fact-check
Rand Paul
stated on January 4, 2022 in an ad:
"No one has been hospitalized for" the omicron variant of COVID-19.

Politifact rating and ruling summary
FALSE
Since omicron was detected in late November, early indications have been that it results in milder illness than other variants of COVID-19. But not even one hospitalization? That was the claim from Sen. Rand Paul, R-Ky., in an ad that solicited contributions for his 2022 reelection campaign. "The COVID hysteria has

Claimant info
Rand Paul, a U.S. senator from Kentucky elected in 2010. He ran for president in 2016 before dropping out to seek re-election to the Senate. Paul is a medical doctor and an ophthalmologist. His father is Ron Paul, a 2008 presidential candidate and a former member of Congress from Texas.
<Retrieved from Politifact>

Augusta Chronicle @AUG_Chronicle Follow
New data shows individuals who are vaccinated and boosted are far less likely to be hospitalized by the omicron variant compared to those who've only had two shots.
augustachronicle.com
Update: Need a COVID-19 vaccination, boost... Health experts continue to encourage COVID-19 vaccinations and boosters as safeguards ...
4:04 PM - Jan 21, 2022
1 See the latest COVID-19 information on Twitter
Read 1 reply

Hyperlinks and titles in the tweet
<https://www.augustachronicle.com/story/news/coronavirus/2021/03/24/covid-19-vaccinations-where-get-one-augusta-area/4660151001/>: COVID-19 vaccinations: Where to get one in the Augusta area

What is the truthfulness stance of the tweet towards the factual claim?

- The tweet believes the factual claim is true.
- The tweet expresses a neutral or no stance towards the factual claim's truthfulness.
- The tweet believes the factual claim is false.
- The tweet and the claim discuss different topics.
- The tweet is created for sarcasm or parody only, or the tweet is problematic (e.g., hyperlink leading to "page not found" error, page content of the hyperlink behind paywall).

Submit Skip this pair Modify My Previous Responses

Truthfulness stance detection annotation example

Figure 4.3: The annotation instruction page.

Main Annotation Page. The main annotation page (Figure 4.4) consists of two primary panels. The left panel displays information about the factual claim from PolitiFact, including the claim itself, claimant details, claim verdict, and a summary of the fact-check review. The right panel displays a corresponding tweet, including the user who posted it and the timestamp. To assist annotators in verifying external references, the panel includes clickable page titles that link directly to the sources mentioned in the tweet.

The figure displays a user interface for annotation. It features two main panels side-by-side.

Left Panel (Politifact):

- POLITIFACT** logo with the tagline "Go to the fact-check".
- Rand Paul** section: A statement from Rand Paul on January 4, 2022, in an ad: "'No one has been hospitalized for' the omicron variant of COVID-19."
- PolitiFact rating and ruling summary:** A box labeled "FALSE" with the note "POLITIFACT RATING OF THIS CLAIM".
- Claimant info:** A brief biography of Rand Paul, noting he was elected in 2010, ran for president in 2016, and is a medical doctor and ophthalmologist.
- Source:** "Retrieved from Politifact»"

Right Panel (Twitter):

- Augusta Chronicle (@AUG_Chronicle) · Follow:** A tweet from the Augusta Chronicle account.
- Text of the tweet:** "New data shows individuals who are vaccinated and boosted are far less likely to be hospitalized by the omicron variant compared to those who've only had two shots." (with a link to augustachronicle.com).
- Timestamp:** 4:04 PM · Jan 21, 2022
- Engagement:** 1 like, 1 reply, 1 retweet, with a link to "See the latest COVID-19 information on Twitter".
- Hyperlinks and titles in the tweet:** A list of links including "https://www.augustachronicle.com/story/news/coronavirus/2021/03/24/covid-19-vaccinations-where-get-one-augusta-area/4660151001/" and "COVID-19 vaccinations: Where to get one in the Augusta area".

Annotation Options:

What is the truthfulness stance of the tweet towards the factual claim?

- The tweet believes the factual claim is true.
- The tweet expresses a neutral or no stance towards the factual claim's truthfulness.
- The tweet believes the factual claim is false.
- The tweet and the claim discuss different topics.
- The tweet is created for sarcasm or parody only, or the tweet is problematic (e.g., hyperlink leading to "page not found" error, page content of the hyperlink behind paywall).

Buttons at the bottom:

- Submit
- Skip this pair
- Modify My Previous Responses

Figure 4.4: The main annotation page.

Below the two panels, annotators can select one of five options: positive (\oplus), negative (\ominus), neutral/no stance (\odot), different topics (dif), and problematic pairs (prb). Specifically, the \oplus stance applies when a tweet conveys the belief that the claim is true. The \ominus stance indicates that the tweet believes the claim is false. The \odot stance signifies that the tweet either expresses uncertainty about the truthfulness of the claim (neutral) or does not explicitly take a position on the claim's truthfulness, even though both the tweet and the claim discuss the same topic (no stance). The

dif option indicates that the tweet and the claim discuss different topics, even though some of their keywords are matched. The *prb* option indicates the tweet is created for sarcasm or parody only (sarcasm detection and parody detection are out of the scope of this study), or the tweet is problematic (e.g., hyperlink leading to “page not found” error, page content of the hyperlink behind a paywall). Annotators have the option “Skip the pair” if they are uncertain about the answer. Additionally, an option “Modify My Previous Responses” allows annotators to review and correct their past annotations. Once annotators click this button, they will be directed to a new page (Figure 4.5). This feature enables annotators to correct past annotations if they find their work quality is low, thereby reinforcing their understanding of the annotation.

The screenshot shows a web interface titled "Previous Answers" with a sub-instruction "(latest responses are at the top)". At the top left is a "Go Back" button. The interface displays three separate sections, each containing a factual claim, a tweet, and a response from the annotator.

- Section 1:**
 - Factual claim:** Alert: Law enforcement and the Red Cross have advised that there are people going door to door indicating that they are authorized to do testing for Coronavirus."
 - Tweet:** Family and friends ALERT, LAW ENFORCEMENT AND THE RED CROSS HAVE ADVISED THAT THERE ARE PEOPLE GOING DOOR TO DOOR INDICATING THAT THEY ARE AUTHORIZED TO DO HOME TESTING FOR CORONAVIRUS. DON'T LET THEM IN, CALL THE POLICE. SHARE WITH ON YOUR PAGE.
 - Response:** Your Response at 2025-06-25 15:44:04: The tweet believes the factual claim is true.
- Section 2:**
 - Factual claim:** The Pfizer COVID-19 vaccine "is not approved."
 - Tweet:** So without a FDA commissioner #Pfizer gets approved!!! I guess the #Democrats are the Captains of that ship too!!! #Unbelievable
 - Response:** Your Response at 2025-06-25 15:43:58: The tweet is created for sarcasm or parody only, or the tweet is problematic (e.g., hyperlink leading to "page not found" error; page content of the hyperlink behind paywall).
- Section 3:**
 - Factual claim:** Says Doug Jones "is for full-term abortion."
 - Tweet:** Commander's Wife Kayla Moore says her husband's opponent Doug Jones supports "full term abortion". Sad when a Wife doesn't understand how pregnancy works.
 - Response:** Your Response at 2025-06-25 15:43:52: The tweet expresses a neutral or no stance towards the factual claim's truthfulness.

Figure 4.5: The page for modifying previous annotations.

In addition to the annotation interface, a leaderboard (Figure 4.6) will appear if an annotator has annotated more than 50 pairs in one session (continuously annotating 50 pairs after logging in), displaying the annotator’s quality and the total amount they have earned.

Leaderboard

USERNAME	ANSWERED	WORK QUALITY SCORE	TOTAL POINTS
	2422	13.80	334.34
	2278	12.24	278.88
	2096	13.06	273.54
	1498	15.30	229.24
	1406	10.84	152.32
780	10.60	82.70	
	908	8.10	73.54
	431	14.30	61.60
	540	10.32	55.72
	338	12.88	43.52
	266	10.06	26.78
	190	11.48	21.82

Figure 4.6: The leaderboard pop-up in the annotation interface.

Control Center. The control center enables system administrators to track annotation progress at both the overall level and the individual annotator level. Access to the control center is restricted to a predefined group of users.

To monitor overall annotation progress, the control center presents a table view showing the progress for each dataset subset (Figure 4.7). The entire set of claim-tweet pairs was randomly divided into ten subsets. These subsets were released sequentially during the annotation process, depending on the progress of each. This subset-based release strategy allowed us to concentrate annotations on specific portions of the dataset. As a result, individual subsets were more likely to reach the

stopping condition (see Section 4.4), enabling us to train preliminary models for experimental purposes. The table also includes details such as the number of pairs that have met the stopping condition (shown as “finished” in the table), the number of finished pairs for each class, and the number of annotations contributed by high-quality annotators.

Overall Annotation Progress per Subset															
Show	All	entries													
subset	total	Training	screening	todo	finished	neg	neu	pos	dif	prb	total_collected	top_quality_collected	progress		
0	579	1	33	545	509	78	88	208	22	113	3050	2637	93.39 %		
1	566	2	47	517	490	87	90	205	12	96	3021	2358	94.78 %		
2	581	2	51	528	509	97	71	227	22	92	3118	2451	96.40 %		
3	556	2	20	534	498	106	87	178	26	101	3092	2580	93.26 %		
4	577	2	25	550	521	82	64	194	35	146	3332	2741	94.73 %		
5	578	2	25	551	521	84	97	203	19	118	3244	2604	94.56 %		
6	567	2	13	552	504	79	63	195	37	130	3383	2737	91.30 %		
7	575	1	12	562	516	66	100	203	30	117	2617	2414	91.81 %		
8	576	1	17	558	508	86	97	195	25	105	2444	2316	91.04 %		
9	561	1	10	550	501	73	81	193	30	124	2410	2274	91.09 %		
TOTAL	5716	16	253	5447	5077	838	838	2001	258	1142	29711	25112	93.21 %		

Showing 1 to 11 of 11 entries

Previous 1 Next

Figure 4.7: A table showing annotation progress by subset.

To monitor individual annotator performance, the control center presents both annotator label distributions (Figure 4.8) and their annotation quality (Figure 4.9). The label distribution table reports the number and percentage of annotations each annotator assigned to each class. It also indicates whether the annotator is classified as high-quality based on our quality control strategy (see Section 4.3) and whether they are currently active (i.e., have submitted annotations within the past ten minutes).

Label Distribution														^
#	username	ANS	neg	neu	pos	dif	prb	neg(%)	neu(%)	pos(%)	dif(%)	prb(%)	top_participant*	ACTIVE
1	SCREENING	253	45	43	103	46	16	17.79	17.00	40.71	18.18	6.32	YES/NO	0
2		94	11	20	31	8	24	11.70	21.28	32.98	8.51	25.53	NO	0
3		60	7	14	23	7	9	11.67	23.33	38.33	11.67	15.00	YES	0
4		41	7	11	13	8	2	17.07	26.83	31.71	19.51	4.88	NO	0
5		1524	243	442	449	209	181	15.94	29.00	29.46	13.71	11.88	YES	0
6		431	56	85	152	99	39	12.99	19.72	35.27	22.97	9.05	YES	0
7		190	35	28	50	36	41	18.42	14.74	26.32	18.95	21.58	YES	0
8		2	0	0	2	0	0	0.00	0.00	100.00	0.00	0.00	NO	0
9		2354	322	597	879	423	133	13.68	25.36	37.34	17.97	5.65	YES	0
10		133	29	15	59	15	15	21.80	11.28	44.36	11.28	11.28	YES	0
11		50	12	12	14	3	9	24.00	24.00	28.00	6.00	18.00	NO	0
12		24	8	0	8	1	7	33.33	0.00	33.33	4.17	29.17	NO	0
13		439	97	76	227	25	14	22.10	17.31	51.71	5.69	3.19	NO	0
14		11	1	2	5	3	0	9.09	18.18	45.45	27.27	0.00	YES	0
15		151	36	38	37	24	16	23.84	25.17	24.50	15.89	10.60	NO	0
16		68	14	10	21	6	17	20.59	14.71	30.88	8.82	25.00	NO	0
17		50	17	8	19	4	2	34.00	16.00	38.00	8.00	4.00	NO	0
18		5	0	2	3	0	0	0.00	40.00	60.00	0.00	0.00	NO	0
19		100	12	22	49	7	10	12.00	22.00	49.00	7.00	10.00	YES	0
20		505	68	84	191	78	84	13.47	16.63	37.82	15.45	16.63	YES	0

Figure 4.8: A table showing label distribution of each annotator.

The work quality table (Figure 4.9) includes each annotator’s quality scores such as “RANK_W” (a weighted error score), their overall quality rating, and the percentage of skipped claim–tweet pairs. Additional information includes their responses to screening pairs (e.g., a “pos_neu” entry indicates the correct label was \oplus , but the annotator selected \odot), total compensation earned, the average length of annotated claim–tweet pairs, and so on. A detailed explanation of the work quality evaluation and the computation of error scores is provided in Section 4.3.

To monitor the progress of each claim–tweet pair, the control center provides a table (Figure 4.10) that displays annotations from high-quality annotators. Specifically, it shows the raw text of each claim–tweet pair, the number of annotations it has

Ranking Information		^															
Answered pairs >= 50																	
Show	10	Search:															
entries																	
#	USER	RANK_W	RANK_L	QUALITY	SKIP	PRT_(€)	PMT_(\$)	LEN	ANS	pos	neu	neg	dif	prb	pos_pos	pos_neu	pos_neg
1	-0.095	815.59	6.32	0.0029	12.64	660.70	262.9485	5224	2138	1306	1006	467	307	99	1	0	
2	i -0.116	676.41	7.06	0.0285	14.24	550.20	265.8152	3864	1403	717	728	445	571	95	1	0	
3	-0.097	561.93	6.40	0.0020	13.06	460.62	265.7144	3526	1222	1145	586	339	234	94	3	2	
4	-0.111	409.58	6.88	0.0137	14.76	344.58	273.1093	2333	864	597	321	418	133	91	3	0	
5	-0.063	207.23	5.20	0.0223	10.98	167.24	271.5250	1522	448	441	243	209	181	64	6	0	
6	-0.175	138.98	9.12	0.0178	18.02	111.20	262.3420	617	301	120	113	53	30	21	0	0	
7	-0.191	119.82	9.68	0.0081	19.52	96.46	264.3785	494	220	133	88	44	9	23	0	0	
8	-0.054	107.31	4.89	0.0105	9.64	82.94	261.6465	860	330	179	137	112	102	27	1	2	
9	-0.068	108.46	5.38	0.1321	10.60	82.70	269.8128	780	295	216	122	96	51	58	3	0	
10	-0.013	96.31	3.46	0.0000	8.10	73.54	284.5011	908	178	271	126	83	250	66	5	2	
#	USER	RANK_W	RANK_L	QUALITY	SKIP	PRT_(€)	PMT_(\$)	LEN	ANS	pos	neu	neg	dif	prb	pos_pos	pos_neu	pos_neg

Showing 1 to 10 of 102 entries

Previous 1 2 3 4 5 ... 11 Next

Figure 4.9: A table showing annotation quality of each annotator.

received from high-quality annotators, and the corresponding labels and annotators. The table also indicates whether the pair is a screening pair, whether it is finished, and if finished, the final assigned label.

4.3 Quality Control Strategy

To ensure high-quality annotations, we adopted the design in ClaimBuster (Arslan et al., 2020) and implemented a screening-based quality control strategy that continuously evaluates annotator performance. For 253 claim–tweet pairs, four domain experts’ annotations agreed on the same label for each pair. These form the *screening pairs*, which were randomly interleaved with regular pairs throughout the annotation process. Annotators were not informed of screening pairs’ presence to ensure genuine assessment of annotation quality. Screening pairs were drawn to reflect the class distribution of the full screening set: 43% positive (\oplus), 17% neutral/no instance (\odot), 17% negative (\ominus), 17% different topic, and 6% problematic pairs.

For each annotator, the initial probability of receiving a screening pair is 30%, which gradually decreases to 10% as they complete annotating more pairs. Specifi-

Progress per "Question"			
Considers only annotations from two participants.			
Id	suspect	claim	tweet
6144	4	-3 "I don't think anybody is questioning the legitimacy of the election. I'm not questioning the legitimacy of the presidential election." "15-year-old boy passes away from heart attack two days after Pfizer Co.'s D/F's experimental jab."	"It's important that this is respected. A Country deserves Change. I don't think anybody is questioning the legitimacy of the presidential election." https://www.washingtonpost.com/news/politics/article/2020/12/04/election-legitimacy-487912
6230	0	-3 "A 15 year old.....having a heart attack.....after being vaccinated.....yeah, right..... 15-Year Old Boy Passes Away from Heart Attack Two Days After Pfizer COVID-19 Experimental Job https://www.eurekalert.org/2020/10/04/15-year-old-boy-passes-away-from-heart-attack-two-days-after-pfizer-covid-19-experimental-job	"A 15 year old.....having a heart attack.....after being vaccinated.....yeah, right..... 15-Year Old Boy Passes Away from Heart Attack Two Days After Pfizer COVID-19 Experimental Job https://www.eurekalert.org/2020/10/04/15-year-old-boy-passes-away-from-heart-attack-two-days-after-pfizer-covid-19-experimental-job
5114	4	-3 There's a 100% chance of dying from COVID-19. "So what?"	"45% of AIs are dead; only 1.8% have had COVID. Vaxed have a better chance of surviving COVID-19. 88% of AIs's COVID-19 cases & 89% of COVID-19 deaths are unvaxed in Lars. The unvaxed are the ones becoming severely ill, lasting in hospital & dying. https://www.wsj.com/2021/11/11/unvaccinated-alabamians-make-up-majority-covid-causes-death-state/ "
4333	3	-3 The Pfizer COVID-19 vaccine is not approved for use over 16.	"So what?"
4160	0	-3 The Pfizer COVID-19 vaccine is not approved."	"COVID-19 vaccine approved to children aged 5-11 Decision comes after European Medicines Agency approves Pfizer pediatric vaccine https://www.catalinaweb.com/covid-19-vaccine-approved-for-children-aged-5-11 "
777	7	-3 "Worldwide data has confirmed the associated with developing COVID in children under 12 years old. Acquired immune Deficiency Syndrome."	"BREAKING: While you've been diagnosed by Russia-Ukraine, Worldwide data has confirmed the Vaccinated and developed COVID Vaccine induced Acquired Immune Deficiency Syndrome. https://usaforapost.substack.com/p/breaking-whilst-youve-been-diagnosed "

Figure 4.10: A table showing annotation details for claim–tweet pairs.

cially, the probability P_s that a given pair is a screening pair decreases over time as the annotator completes more annotations, according to the following dynamic formula:

$$P_s = \tau_{\max} - \left(\frac{\tau_{\max} - \tau_{\min}}{N_{\text{cutoff}}} \right) \cdot \min(N_{\text{cutoff}}, N_{\text{ann}})$$

Here, $\tau_{\max} = 0.3$ is the initial screening probability, $\tau_{\min} = 0.1$ is the minimum probability after sufficient annotations, $N_{\text{cutoff}} = 250$ is the stabilization point, N_{ann} is the number of annotations completed by a given annotator. As a result, the probability is set to a minimum of 10% when the annotator has finished 250 pairs or more.

To assess annotation quality, we calculated a weighted error score (E_w) for each annotator based on their responses to the screening pairs:

$$E_w = \frac{-0.2 \cdot N_{\text{correct}} + 0.5 \cdot N_{\text{mild}} + 1.0 \cdot N_{\text{moderate}} + 2.0 \cdot N_{\text{severe}}}{N_{\text{screen}}}$$

where N_{screen} is the total number of completed screening pairs. N_{correct} , N_{mild} , N_{moderate} , and N_{severe} respectively represent the number of correct annotations and the number of incorrect annotations at mild, moderate, and severe error levels. For example, annotating \oplus as \ominus is considered a severe error, \odot as \oplus is moderate, and prb as \odot is mild. We link annotation quality to monetary compensation through a reward function R :

$$R = 2.0 \cdot Q(E_w) \cdot \left(\frac{A_{\text{ann}}}{A_{\text{all}}} \right)^2 \cdot 0.6^{\frac{N_{\text{skip}}}{N_{\text{ann}}}} \cdot \frac{N_{\text{ann}}}{100}$$

Here, A_{ann} is the average token length of the claim–tweet pairs annotated by the annotator, and A_{all} is the average length across the whole dataset. N_{skip} refers to

the number of claim–tweet pairs the annotator chose to skip. The quality multiplier $Q(E_w)$ rewards low-error annotators and penalizes high-error ones, defined as:

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

In this way, annotators who maintain low error rates and engage with lengthier content are fairly rewarded, while discouraging careless or low-effort contributions.

4.4 Stopping Condition

The annotations from low-quality annotators are ignored in determining whether a claim–tweet pair requires more annotations. An annotator is considered high-quality if their error score falls below a defined threshold. By adjusting this threshold, we control the trade-off between annotation quality and dataset size.

A claim–tweet pair is considered finished when sufficient agreement is reached among high-quality annotators (those with $E_w < 0$ in our setting). Specifically, given a pair, the stopping condition is met if any one of the four labels, i.e, \oplus , \odot , \ominus , dif , satisfies all of the following: (1) It has been selected by at least three annotators; (2) It exceeds each of the other three labels by at least two votes; (3) The number of votes it receives is greater than or equal to half the combined total for the other three labels. Alternatively, the annotation also stops immediately if any annotator selects the *prb*. Formally, the stopping condition is met if any one of the four labels, i.e, $\oplus(L_0)$, $\odot(L_1)$, $\ominus(L_2)$, $dif(L_3)$, satisfies all of the following: (1) It has been selected by at least three annotators; (2) It exceeds each of the other three labels by at least two votes; (3) The number of votes it receives is greater than or equal to half the

combined total for the other three labels (i.e., the remaining labels excluding itself). Alternatively, the annotation also stops immediately if any high-quality annotator selects *prb* (L_4). This decision rule D is formally defined as:

$$D(L) = \begin{cases} \text{true,} & \text{if } \exists i \in \{0, 1, 2, 3\} \text{ such that} \\ & L_i \geq 3 \wedge L_i \geq L_j + 2 \forall j \neq i, j \in \{0, 1, 2, 3\} \wedge \\ & L_i \geq \left\lfloor \frac{1}{2} \sum_{\substack{j=0 \\ j \neq i}}^3 L_j \right\rfloor \\ \text{true,} & \text{if } L_4 > 0 \\ \text{false,} & \text{otherwise} \end{cases}$$

4.5 Dataset Statistics

Threshold-Based Quality Filtering and Dataset Growth. Figure 4.11 shows how dataset statistics (e.g., the count of finished pairs and high-quality annotators) change with different error score thresholds used to define high-quality annotators. The top portion of the figure presents inter-annotator agreement (IAA) scores, including Krippendorff's alpha (Krippendorff, 2011) and Gwet's AC2 (Gwet, 2008), both of which can handle missing data. IAA scores were evaluated based on the \oplus , \odot , and \ominus labels, which are treated as ordinal categories. The *dif* label is outside the task scope, and the *prb* label is mostly assigned by only one high-quality annotator, making agreement calculations inapplicable. Given the setting of our annotation setup, the most appropriate agreement scores are Krippendorff's alpha and Gwet's AC2. Other scores, such as Cohen's Kappa (Cohen, 1960) and Scott's Pi (Scott, 1955), assume exactly two fully overlapping annotators per item, and Fleiss' Kappa (Fleiss, 1971) requires a fixed number of annotators for each item. These assumptions do not match

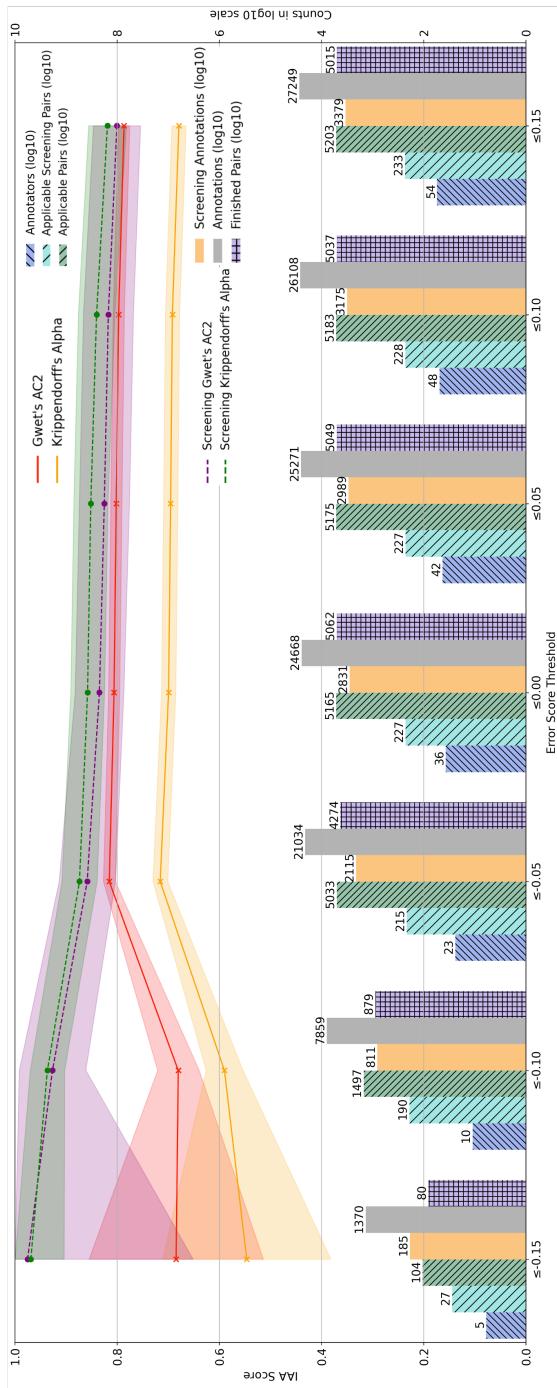


Figure 4.11: Annotation statistics and IAA scores by varying error score thresholds. Line plots show Krippendorff's alpha and Gwet's AC2 for screening and non-screening pairs with confidence intervals. Bar plots indicate counts on a logarithmic scale for annotations, screening annotations, finished pairs, applicable pairs, and annotators.

the characteristics of our annotation design. The two IAA scores were computed for both screening and non-screening (i.e., real task) annotations. The shaded areas represent 95% confidence intervals. The bottom portion of the figure presents the counts on a logarithmic scale for annotations, finished pairs (i.e., pairs that satisfied the stopping condition), screening annotations, and the number of applicable screening (i.e., those with at least two annotations) and applicable non-screening pairs.

As we relax the threshold (from ≤ -0.15 to ≤ 0.15), more high-quality annotators are included, leading to sharp increases in both the number of annotated pairs and total annotations. For instance, at a very strict threshold (≤ -0.15), only five annotators are included, producing 1,370 annotations and covering just 80 finalized claim–tweet pairs. At a more moderate threshold of ≤ 0.00 , 36 annotators contribute 24,668 annotations and 5,062 finalized pairs. However, beyond this threshold, the number of finalized pairs begins to decline, suggesting that including lower-quality annotators makes it harder to reach agreement.

Inter-Annotator Agreement Across Thresholds. The IAA scores show an inverse-U pattern. Specifically, at strict thresholds (≤ -0.15 or ≤ -0.10), although only annotators with top performance are included, the IAA scores are low due to the limited number of applicable pairs—many claim–tweet pairs have too few annotations to compute a robust agreement, resulting in wide confidence intervals. For example, at ≤ -0.15 , Krippendorff’s alpha and Gwet’s AC2 for screening pairs reach near-perfect scores due to the minimal size and high consistency, but their confidence intervals are large and less reliable. As the threshold increases to ≤ -0.05 , we observe peak agreement scores—Gwet’s AC2 reaches 0.815 and Krippendorff’s alpha reaches 0.716. Beyond this point, agreement scores slightly decline as lower-quality annotators are included, but they remain strong and stable. Notably, the number of finalized pairs

peaks at the threshold of 0.00, where Gwet’s AC2 is 0.803 and Krippendorff’s alpha is 0.699. To balance annotation quality and dataset size, we selected 0.00 as the threshold for our final dataset.

Error Analysis of Screening Pair Annotations. To better understand the mistakes annotators made, we analyzed the 2,831 annotations contributed by high-quality annotators (those with $E_w < 0$) on screening pairs. Figure 4.12 presents a confusion matrix between the gold-standard screening labels (x-axis) and the annotators’ responses (y-axis). Diagonal cells represent correct annotations. Off-diagonal cells show both the count and the percentage of incorrect responses for each screening label, where the percentage (column-wise) represents the recall for that label. Across all screening pairs, we observed 413 incorrect annotations, corresponding to an overall error rate of 14.59%. The *prb* (problematic) label had the lowest recall at 48.6%, which may be caused by the small number of *prb* screening pairs. In contrast, \oplus , \odot , and \ominus achieved high recall: 92.5%, 81.4%, and 91.4%, respectively.

The \odot stance was the most frequently chosen incorrect label across all other screening classes, both in absolute error count and percentage. For example, *dif* was mislabeled as \odot 52 times, and *prb* was mislabeled as \odot 54 times. This pattern suggests that \odot may be perceived as a “safe” choice when annotators are uncertain. This is consistent with the observations from prior stance detection studies (Joseph et al., 2017).

Final Dataset Composition. Under the final threshold, TSD-CT contains 5,331 finalized claim–tweet pairs (including 269 screening and training pairs) covering 2,201 unique factual claims. The label distribution is shown in Table 4.1: 2,104 (39.47%) are labeled as \oplus , 882 (16.54%) as \odot , 883 (16.56%) as \ominus , 309 (5.80%) as *dif*, and 1,153 (21.63%) as *prb*. The claim veracity is also diverse, with 722 claims labeled

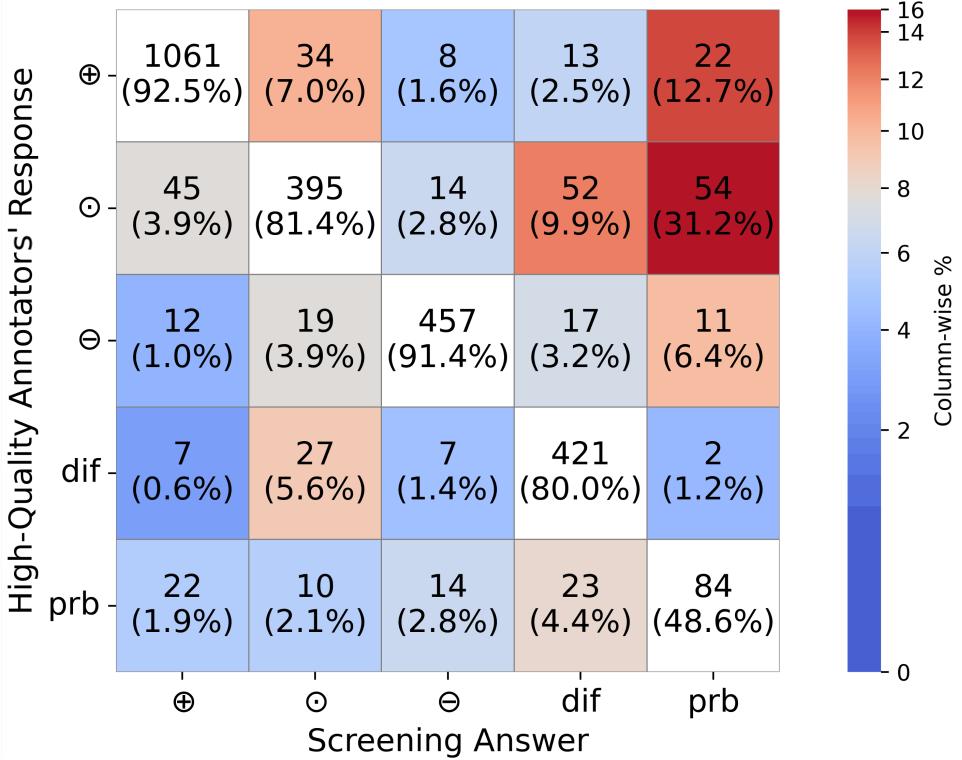


Figure 4.12: Confusion matrix for annotations on screening pairs by high-quality annotators. Diagonal values indicate correct answers; off-diagonal values indicate errors with both raw count and recall (column-wise percentage).

as false (32.80%), 353 as pants on fire (16.04%), 340 as barely true (15.45%), 292 as half true (13.27%), 287 as mostly true (13.04%), and 207 as true (9.40%). On average, each claim–tweet pair contains 34.61 tokens and 261.76 characters. Additionally, 2,169 pairs (40.71%) include at least one hyperlink, indicating external reference or contextual support. Among all the 845 topics, the dataset is dominated by discussions on 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%) and so on. This diverse and richly annotated dataset supports fine-grained modeling of stance, belief dy-

namics, and misinformation-related behaviors across topics and veracity types. To promote transparency and reusability, the dataset is published in accordance with the FAIR principles (Wilkinson et al., 2016). It is released under a CC BY 4.0 license and made available in standardized formats with detailed documentation on Zenodo (<https://doi.org/10.5281/zenodo.15620262>). In compliance with X’s content sharing policy, the dataset does not include raw tweet text; instead, it provides tweet IDs and associated metadata.

Label	\oplus	\odot	\ominus	dif	prb
Pair Count	1,832	689	736	257	1,068
Percentage	40.64%	14.52%	16.32%	6.96%	21.55%

Table 4.1: Distribution of stance labels in the finalized TSD-CT dataset.

CHAPTER 5

EXPLORING TRUTHFULNESS STANCE DETECTION APPROACHES

This chapter presents a range of approaches for detecting stance (including different types of stance as discussed in Chapter 2). Those approaches span from early rule-based techniques to modern retrieval-augmented neural models. We organize the discussion by model architecture and functionality, highlighting how each generation of methods builds on the last to better handle the challenges of stance detection in social media. Section 5.1 reviews traditional approaches (lexicon-based and classical machine learning) and their limitations. Section 5.2 covers the rise of neural network models: from Recurrent Neural Networks (RNNs) and Convolutional Neutral Networks (CNNs) to attention mechanisms and graph-based models. Section 5.3 surveys the transformer-era and beyond, including large pre-trained language models and retrieval-augmented frameworks. Throughout, we note key trends, representative benchmark tasks, and the persistent challenges (e.g., interpretability, context sparsity, noisy text) that motivated newer approaches.

5.1 Early Approaches

In the early stages of stance detection research, many models relied on hand-crafted rules and lexicons. Researchers manually encoded explicit cues such as stance-indicative keywords, negation patterns, and sentiment lexicons to infer a post’s stance. Such rule-based approaches were easy to interpret but brittle: they often failed when a speaker expressed the same stance using different phrasing or an ironic tone. By

the 1990s and 2000s, the field had shifted to supervised learning on annotated data. Classic machine learning models such as support vector machines, logistic regression, Naïve Bayes, and decision trees became common (Pranckevičius and Marcinkevičius, 2017). These still needed extensive feature engineering (n-grams, part-of-speech (POS) tags, sentiment scores, etc.) but generalized better than rigid rules. For example, a prior study (Ferreira and Vlachos, 2016) trained a logistic regression classifier to predict whether a news headline supported or refuted a claim. Even so, these models struggled with context and data sparsity. Early benchmarks highlighted the shortcomings: participants in the SemEval-2016 stance detection task on tweets, for instance, found that simple word-overlap or sentiment features alone were insufficient for high accuracy (Mohammad et al., 2016). In short, while rule-based and classical machine learning methods laid the groundwork, they lacked deep language understanding, motivating the search for more powerful approaches.

5.2 Neutral Network Models

With the deep learning revolution in the mid-2010s, researchers began adopting neural networks to learn informative features for stance detection automatically. CNNs (Lai et al., 2015) and RNNs (Hochreiter and Schmidhuber, 1997) quickly became prevalent. These models could capture word order and composition, allowing them to detect patterns such as negation or emphasis. In tasks with explicit targets (e.g., stance toward “Donald Trump” or “climate change”), models were augmented with target-specific attention: the network learns to focus on words related to the given topic (Du et al., 2017). Researchers also extended context beyond a single post. In rumor stance classification, some methods represent the reply-thread as a graph or tree. For example, a Hierarchical Graph Convolutional Network-Recurrent Neural Network (HGCRN-RNN) model encodes a Twitter thread with graph convolutional

networks plus an RNN, while a Recursive NN (RvNN) processes the propagation tree with a tree-structured RNN (Patel et al., 2022). These architectures can learn how earlier posts influence the stance of a reply. Overall, neural models achieved much richer text representations and better performance. However, they introduced new challenges: deep models are largely opaque and require large data, a problem since stance datasets remain relatively small. Indeed, recent work notes the scarcity of labeled data and the difficulty of explaining deep models as key hurdles (Hou et al., 2022). By 2017–2018, pure neural approaches were state-of-the-art on many stance benchmarks. For example, at SemEval-2017 RumourEval (classifying whether replies support, deny, or question a rumor), the top system used a Bidirectional Long Short-Term Memory (Bi-LSTM) to model conversation threads, achieving the highest accuracy in Subtask A (Kochkina et al., 2017). These successes set the stage for leveraging even more powerful pre-trained models.

5.3 Pre-Trained Language Models

The next leap in stance detection came with the advent of large pre-trained language models, particularly the transformer-based architectures. BERT (Bidirectional Encoder Representations from Transformers) and its variants brought a “pre-train then fine-tune” paradigm that rapidly became dominant after 2018. Instead of training a stance model from scratch, researchers now fine-tune models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), or XLNet (Yang et al., 2019b), which are already imbued with general language understanding from huge text corpora, on stance datasets. This shift dramatically improved accuracy across many natural language processing (NLP) tasks, including stance detection, because the models came with a rich grasp of semantics, synonyms, and even some factual knowledge. In stance detection benchmarks, the impact was evident: for instance, the

winning system in SemEval-2019’s rumor stance task (BUT-FIT) (Fajcik et al., 2019) simply fine-tuned a BERT-based classifier and achieved top performance. Another high-performing entry (BLCU_NLP) (Yang et al., 2019a) employed a GPT 2-based model with an “inference chain” approach, essentially leveraging a generative transformer to reason about the claim and replies. These transformer-based approaches significantly outperformed previous RNN/CNN models, especially in capturing subtle semantic relations between a claim and a statement. By leveraging transfer learning and large-scale training, transformers set new state-of-the-art results on tasks such as FEVER (Thorne et al., 2018), where the goal is to verify claims against evidence, and on emerging stance datasets such as COVIDLies. A notable trend in this era was the focus on domain adaptation: researchers developed domain-specific variants (e.g. SciBERT (Beltagy et al., 2019), COVID-Twitter-BERT (Müller et al., 2023)) and training strategies to handle unique vocabulary and topics in health, climate, or political domains. Despite their superior accuracy, transformer models inherited the typical deep model issues of data hunger and interpretability. In addition, they still struggled with some nuances of social media text. For example, tweets often contain idioms, sarcasm, or context that even BERT isn’t explicitly trained on. Moreover, stance detection in the wild sometimes requires knowledge of real-world events or facts (to know what a claim refers to), beyond what is encoded in the model’s parameters. These shortcomings prompted researchers to enrich transformer-based systems with more context and to make their predictions more explainable.

5.4 The RATSD Framework

The most recent developments marry powerful language models with external knowledge retrieval and generation capabilities. The guiding idea is that, to judge a tweet’s stance about a claim (especially a truthfulness stance), a model may benefit

from retrieving related information (e.g. verified facts, previous fact-checks) and from producing human-readable rationales. Our proposed RATSD is an example of this new breed.

The design of the RATSD framework hinges on two key challenges associated with the data: 1) Both claim c and tweet t are standalone sentences that often lack sufficient context, making it difficult for a classification model to make an informed decision. 2) Tweets frequently contain acronyms, hashtags, and slang, which pose challenges for the classification model to interpret accurately.

RATSD counters these challenges with two innovative data augmentation ideas, both leveraging LLMs' abilities. One is to employ RAG (retrieval augmented generation) to retrieve relevant contextual information from external knowledge corpora to compensate for the inherent lack of context. The other is to synthesize an analysis of the tweet t based on the retrieved context. The tweet analysis directly incorporates LLM's perspective on t 's truthfulness stance toward c . Additionally, it helps mitigate the challenges posed by the aforementioned informal language. Recent advancements have demonstrated the effectiveness of RAG in knowledge retrieval (Lewis et al., 2020; Wang et al., 2023) and LLMs' success in text analysis (Tang et al., 2023).

Reflecting this design, the RATSD framework as depicted in Figure 5.1 comprises three main components: the construction of external knowledge corpora (dashed blue box), the LLM-enabled data augmentation which includes RAG and tweet analysis generation (dotted orange box), and the fine-tuning of truthfulness stance classification model (dash-dotted yellow box). The rest of this section discusses these components in detail.

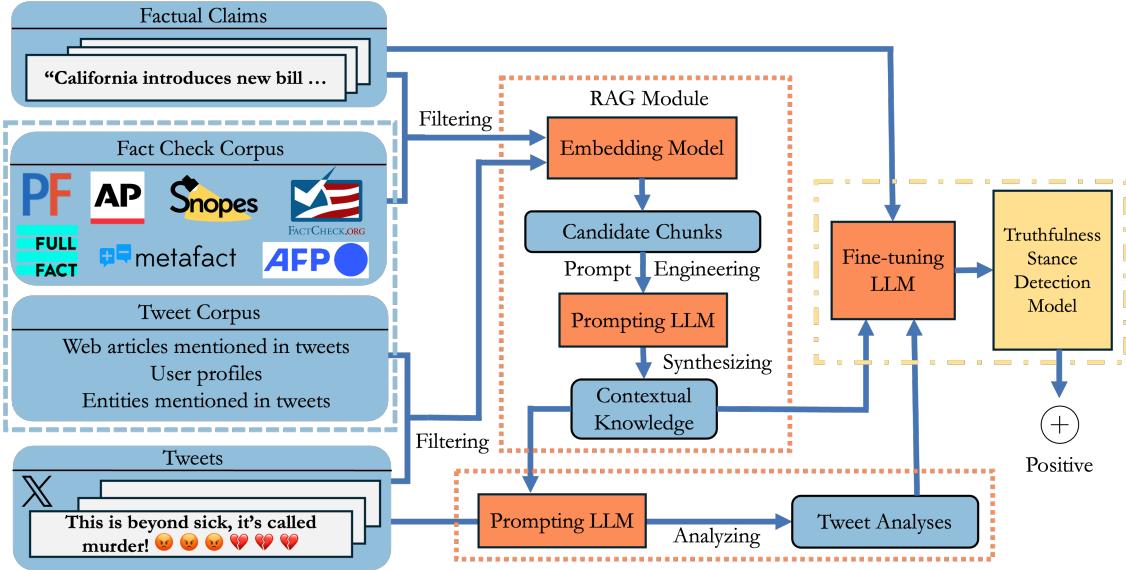


Figure 5.1: The RATSD framework.

5.4.1 Knowledge Corpora Construction

Two knowledge corpora were constructed to provide contextual knowledge for other components in RATSD, one for claims and the other for tweets.

The first knowledge corpus, denoted \mathcal{D}_C , encompasses 52,596 synthesized documents for factual claims. It is worth noting that, although the claim-tweet pairs in the TSD-CT dataset include claims from PolitiFact only, the knowledge corpus incorporates claims and corresponding fact-checks published by seven fact-checking websites from 1995 to 2023. Additionally, some claims were fact-checked by multiple websites. Given a claim c , the corresponding synthesized document d_c was constructed by concatenating excerpts from fact-checks (i.e., articles) on the claim. Each excerpt includes the following information: the claim c itself, the name of the claimant and their profile description from the fact-checking website, the date and location of the claim, the publication date of the fact-check article, the summary of

the fact-checking ruling provided in the article, and the main body of the article. The resulting d_c typically ranges from 10,000 to 30,000 words in length.

The second knowledge corpus, \mathcal{D}_T , consists of 8,236 synthesized documents for tweets posted from 2010 to 2023. Given a tweet t , the corresponding document d_t was constructed by concatenating the following information: the raw HTML content of all web pages linked in the tweet, the profile description (retrieved using Twitter API) of the account that posted the tweet, and information (name and description, from Twitter API) about the entities mentioned in t .

5.4.2 Contextual Knowledge Generation

Using the constructed knowledge corpora \mathcal{D}_C and \mathcal{D}_T , RATSD generates contextual knowledge in the form of a document e_c for c and a document e_t for t , given a claim-tweet pair (c, t) . Note that the set of claims from the claim-tweet pairs in TSD-CT, i.e., the set of c for which e_c was generated (let us call it C_1), is not identical to the set of c for which d_c was constructed in forming \mathcal{D}_C (call it C_2). Specifically, C_2 is a much larger superset of C_1 , as each $c \in C_1$ is sourced from PolitiFact while each $c \in C_2$ can be from any of the seven fact-checking websites. The rationale was that useful contextual knowledge for a claim can come from not only the claim itself but also other relevant claims. Similarly, the set of tweets from TSD-CT is the annotated subset of tweets in d_t (see Section 3.2 for the tweet collection process of d_t).

The e_c and e_t are critical for accurate truthfulness stance detection. Particularly, such contextual information is instrumental in mitigating LLM-generated hallucination (Ji et al., 2023; Yao et al., 2023; Tonmoy et al., 2024). The generation process of e_c and e_t follows four steps: 1) document preprocessing, 2) relevant document selection, 3) relevant chunk retrieval, and 4) prompting LLM.

Document Preprocessing. All the documents in both \mathcal{D}_C and \mathcal{D}_T were segmented into smaller chunks (i.e., continuous sequence of tokens), each with a token size of 512. For each chunk, we used the BAAI general embedding (BGE) model (Xiao et al., 2024) to generate its text embeddings. The BGE model, being a lightweight, pre-trained model, has demonstrated strong performance in the text embedding leaderboard (Muennighoff et al., 2023).

Relevant Document Selection. We used a keyword-based approach to select relevant documents for c from the claim knowledge corpus \mathcal{D}_C . Nouns, verbs, and adjectives were extracted from c . Jaccard similarity between the extracted words and each document $d_{c_i} \in \mathcal{D}_C$ was calculated. Top 10 documents based on the similarity scores were selected as relevant documents for c . The same approach was used to select the 10 most similar documents for t from the tweet knowledge corpus \mathcal{D}_T . This step excludes irrelevant documents from consideration and thus reduces noise in the next step. Furthermore, it also helps reduce the computational cost of LLM retrieval by limiting it to a smaller set of documents.

Relevant Chunk Retrieval. Not all the chunks of the selected top documents are relevant to c and t . Given each c and t , the top 10 most relevant chunks were retrieved. For retrieving relevant chunks, we used the BGE embeddings and applied cosine similarity to measure the semantic alignment between each chunk and a text query based on c (or t). The query is essentially the same prompt instruction used in prompting the LLM, as follows.

Prompting the LLM. To generate high-quality e_c and e_t , we designed a prompt. It includes both c and t , along with the specific instruction to generate relevant contextual knowledge. The top portion of Figure A.1 in the Appendix shows an example prompt. As described above, this prompt was used to find relevant chunks

based on their vector embeddings. Once the most relevant chunks have been retrieved, they are fed into the LLM along with the same prompt to generate the contextual knowledge e_c and e_t for the factual claim c and the tweet t . Examples of e_c and e_t are shown in Figures A.2 and A.3 in the Appendix, respectively.

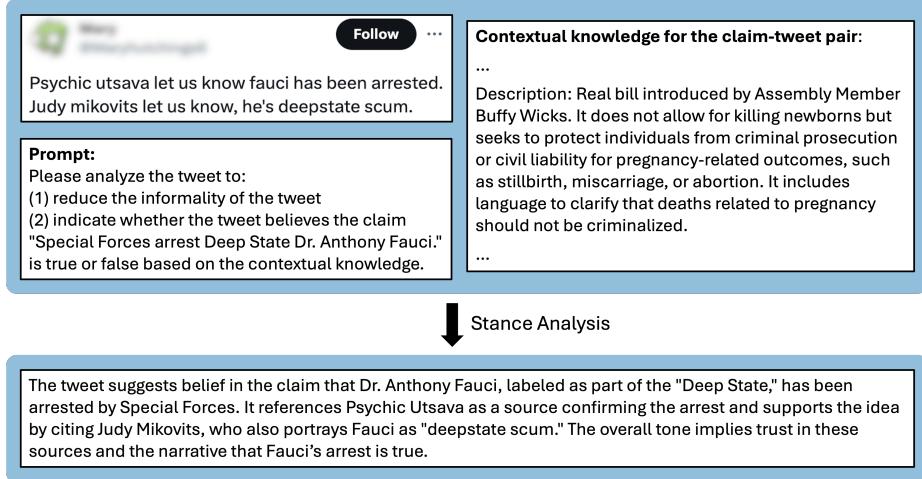


Figure 5.2: An example of stance analysis.

Stance Analysis. Utilizing the contextual knowledge described above, RATSD generates the stance analysis for each claim-tweet pair (c, t) . Specifically, an LLM is prompted using c, t, e_c and e_t as the input to generate a narrative of t 's truthfulness stance regarding c . We use a to denote the generated stance analysis. The prompt instruction and an example input can be found in Figure 5.2.

When training the stance detection model, a will replace t in the input claim-tweet pair, as detailed in Section 5.4.3. This approach is helpful for producing the final stance classification model in three ways. First, it leverages the power of LLMs to analyze the tweet's stance and the analysis is directly included in training the detection model. (Chapter 6 reports experiment results comparing our approach

with directly prompting LLMs.) Second, the analysis incorporates additional context from e_c and e_t which is not in the original t . Finally, this approach helps reduce the informality in tweet content (e.g., acronyms, hashtags, slang, and nickname references of entities) which otherwise presents a challenge in training the model.

5.4.3 Classification Model

RATSD 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 $h_i = ([CLS], a_i, [SEP], c_i, [SEP], e_{ti}, e_{ci})$. The [CLS] token’s final hidden state is utilized as the aggregate representation of the entire input sequence. The vector is fed into a single fully connected layer and a softmax layer to produce the probability distribution of stance orientation labels $\{\hat{s}_i^\oplus, \hat{s}_i^\odot, \hat{s}_i^\ominus\} = \text{softmax}(Wh_i + b)$ where W and b are trainable parameters. The LLM is optimized by a cross-entropy loss $\min_{\Theta} \mathcal{L} = -\sum_i \sum_{o \in \{\oplus, \odot, \ominus\}} s_i^o \log(\hat{s}_i^o) + \lambda \|\Theta\|^2$ where s_i^o and \hat{s}_i^o are the gold-standard probability and predicted probability for stance orientation o of the i -th input, Θ denotes all trainable parameters of the model, and λ represents the coefficient of L_2 -regularization. The model parameters are fine-tuned during training and optimized using the Adam optimizer (Kingma and Ba, 2015). The fine-tuning process involves minimizing the cross-entropy loss between the predicted stance distribution and the gold-standard stance distribution.

CHAPTER 6

EVALUATION

6.1 Experiment Datasets

As noted in Chapter 2, several benchmark datasets are available for stance detection and only a few of these datasets closely align with our concepts. Therefore, we selected the three most similar benchmark datasets—SemEval-2019, WT-WT, and COVIDLies—for performance comparison, along with our own TSD-CT dataset. However, the stance and class categories are defined and named differently in these datasets. Thus, merging and renaming labels is necessary to ensure a fair comparison of model performance. Specifically, we renamed “support” as “positive” and “deny” as “negative” for the SemEval-2019 dataset. Additionally, we merged the “comment” and “query” classes into a single “neutral/no stance” class because the “query” class is included in the “neutral” class in our conceptual framework. For the WT-WT dataset, we renamed “support” as “positive,” “refute” as “negative,” and removed the “unrelated” category as “unrelated” category falls outside the scope of our research. For the COVIDLies dataset, we renamed “agree” as “positive,” “disagree” as “negative” and “no stance” as “neutral/no stance.” The label distributions of SemEval-2019, WT-WT, and COVIDLies are shown in Table 6.1.

6.2 Implementation Details

All experiments were conducted using 1 NVIDIA A100 80GB GPU. Due to our limited GPU memory, we applied 8-bit quantization for LLM fine-tuning. Due to the

Dataset	(⊕)	(⊙)	(⊖)	Total
SemEval-2019	1,184 (13.8%)	6,784 (79.1%)	606 (7.1%)	8,574
WT-WT	6,663 (21.0%)	20,864 (65.7%)	4,224 (13.3%)	31,751
COVIDLies	670 (9.9%)	5,748 (85.1%)	340 (5.0%)	6,758
TSD-CT	1,262 (56.9%)	451 (20.3%)	507 (22.8%)	2,220

Table 6.1: Label distribution of SemEval-2019, WT-WT, COVIDLies and TSD-CT datasets.

resource-intensive nature of LLMs, fine-tuning these models directly on these GPUs was not feasible. So we turn to quantization, which reduces the precision of numerical representations used in model calculations, typically from 32-bit floating-point numbers to 8-bit or even lower. This significantly reduces the model’s size and speeds up computations while maintaining good performance. Notably, Banner et al. (2019) indicates that up to 4-bit quantization can be applied to LLMs without substantially compromising their accuracy. The classification model in RATSD was fine-tuned using selected hyperparameters. The learning rate was set to 5e-5, balancing convergence speed and stability. We utilized a batch size of 8 for both training and evaluation. The models were trained for three epochs. We applied a weight decay of 0.01. We used GPT-3.5, with a temperature of 0.1 and a maximum output token length of 4,096, for contextual knowledge generation and stance analysis in RATSD.

6.3 Experiment Results

We evaluated the performance of two types of stance detection models: LM-based and LLM-based. The evaluation was conducted in two different settings: (1) fine-tuning the models for stance detection, and (2) applying them directly in a zero-shot setting.

Evaluation Metric. Consistent with previous studies, we used the F1 score for each class—denoted as F_{\oplus} (support), F_{\odot} (neutral), and F_{\ominus} (refute)—to measure per-class performance. To obtain an overall assessment that treats all classes equally regardless of their frequency, we computed the **Macro F1 score** (F_M), which is the unweighted average of the per-class F1 scores. Formally, for each class $c \in \{\oplus, \odot, \ominus\}$, the F1 score is computed as:

$$F_c = \frac{2 \cdot P_c \cdot R_c}{P_c + R_c} \quad (6.1)$$

where P_c and R_c denote the precision and recall for class c , respectively. The Macro F1 score is then defined as:

$$F_M = \frac{1}{3} (F_{\oplus} + F_{\odot} + F_{\ominus}) \quad (6.2)$$

Fine-tuned Model Performance.

We evaluated the performance of RATSD by comparing it to several state-of-the-art stance detection models, including fine-tuned LMs such as pre-trained model (BUT-FIT (Fajcik et al., 2019)), generative pre-trained model (BLCU_NLP (Yang et al., 2019a)), domain-adaptive pre-trained model (BERTSCORE+NLI (Hossain et al., 2020), BART+NLI (Reddy et al., 2022) and TESTED (Arakelyan et al., 2023)). In RATSD, we utilize two fine-tuned LLMs: the open-source model Zephyr (Tunstall et al., 2023) and the proprietary model GPT-3.5.

As shown in Table 6.2, 6.3, 6.4 and 6.5, RATSD demonstrates strong performance across all datasets compared to other stance detection models. On the TSD-CT dataset (Table 6.2), RATSD_{GPT-3.5} achieves the highest scores across all metrics. For the SemEval-2019 dataset (Table 6.3), RATSD_{GPT-3.5} surpasses other models in F_{\odot} score and achieves the highest macro F1 score. RATSD_{Zephyr} demonstrates its

strength on the WT-WT dataset (Table 6.4), where it secures the highest performance across all metrics. While on the COVIDLies dataset (Table 6.5), BERTSCORE+NLI and BART+NLI slightly outperform RATSD, RATSD still delivers competitive results. These results suggest that while RATSD models, especially RATSD_{GPT-3.5}, demonstrate strong performance. However, different fine-tuned LLM in RATSD may excel in specific datasets or stance categories, highlighting the importance of model selection based on the specific task and dataset characteristics.

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 _{Zephyr}	88.67	77.38	80.28	82.10
RATSD _{GPT-3.5}	93.27	80.24	87.90	87.13

Table 6.2: Performance comparison on the TSD-CT dataset.

Model	F_{\oplus}	F_{\odot}	F_{\ominus}	F_M
BUT-FIT	49.09	50.98	92.01	64.03
BLCU_NLP	70.15	40.00	88.12	66.09
BERTSCORE+NLI	46.96	60.67	91.32	66.32
BART+NLI	47.96	51.71	91.90	63.86
TESTED	46.43	58.04	92.08	65.52
RATSD _{Zephyr}	41.71	55.42	91.80	62.97
RATSD _{GPT-3.5}	56.12	63.79	83.67	67.86

Table 6.3: Performance comparison on the SemEval-2019 dataset.

Model	F_{\oplus}	F_{\odot}	F_{\ominus}	F_M
BUT-FIT	81.29	94.73	79.29	85.10
BLCU_NLP	81.02	94.74	77.09	84.28
BERTSCORE+NLI	82.02	95.06	79.11	85.39
BART+NLI	82.82	95.52	81.75	86.70
TESTED	81.75	94.98	78.00	85.91
RATSD _{Zephyr}	83.85	95.72	82.66	87.44
RATSD _{GPT-3.5}	75.78	92.98	75.07	81.27

Table 6.4: Performance comparison on the WT-WT dataset.

Model	F_{\oplus}	F_{\odot}	F_{\ominus}	F_M
BUT-FIT	47.62	97.82	23.53	56.32
BLCU_NLP	52.38	97.71	45.46	65.18
BERTSCORE+NLI	57.14	98.20	58.33	71.22
BART+NLI	50.00	98.00	60.87	69.62
TESTED	40.00	97.12	51.85	62.99
RATSD _{Zephyr}	51.42	97.63	54.55	67.87
RATSD _{GPT-3.5}	51.16	98.06	52.63	67.30

Table 6.5: Performance comparison on the COVIDLies dataset.

Zero-shot Performance on TSD-CT Dataset. To assess the model’s ability to generalize its learning to unseen classes without any prior examples. We conducted zero-shot performance evaluation on TSD-CT dataset, as shown in Table 6.6. Among the models, RATSD_{Zephyr_{zero}} achieves the highest overall performance, with the F_M of 36.55. This suggests that RATSD_{Zephyr_{zero}} is a strong framework for truthfulness stance detection in the zero-shot setting. RATSD_{Zephyr_{zero}} outperforms RATSD_{GPT-3.5_{zero}} across most metrics. The results suggest that Zephyr is better suited for zero-shot scenarios, potentially due to its model architecture or the nature of its fine-tuning, which might be better at generalizing to new tasks without task-specific training. Notably, GPT-3.5_{zero} uses direct prompting, as described in Section 5.4.2, which demonstrates its difficulty in achieving strong performance without fine-tuning.

Model	F_{\oplus}	F_{\odot}	F_{\ominus}	F_M
BUT-FIT _{zero}	12.82	0.00	33.88	15.56
BLCU_NLP _{zero}	27.05	0.00	32.81	19.95
BERTSCORE+NLI _{zero}	6.82	41.71	17.65	22.06
BART+NLI _{zero}	33.55	40.58	3.96	26.03
TESTED _{zero}	55.84	38.91	4.04	32.93
GPT-3.5 _{zero}	34.04	16.81	39.74	30.20
RATSD _{Zephyr zero}	49.74	32.14	27.78	36.55
RATSD _{GPT-3.5 zero}	28.76	29.71	33.46	30.64
RATSD _{Zephyr}	88.67	77.38	80.28	82.10
w/o analysis	87.85	74.39	81.01	81.08
w/o context & analysis	87.16	75.15	78.01	80.11

Table 6.6: The zero-shot model performance comparison and ablation study on the TSD-CT dataset.

6.4 Ablation Study

To assess the effectiveness of contextual knowledge generation and stance analysis, we conducted an ablation study with two model variations on the TSD-CT dataset: RATSD_{Zephyr} without stance analysis (w/o analysis) and RATSD_{Zephyr} without contextual knowledge generation and stance analysis (w/o context & analysis). The results in the bottom three rows of Table 6.6 reveal the impact of key components on the performance of RATSD_{Zephyr}. When stance analysis is removed, both the F_{\oplus} and F_{\odot} decline, indicating that stance analysis provides useful additional context for both positive and neutral pairs, although it slightly reduces the performance for the negative class. Further removing contextual knowledge generation results in a drop in performance across all F1 categories. The decline in the F_{\odot} and F_{\ominus} indicates that contextual knowledge generation is crucial in handling neutral or negative pairs. The decrease in F_{\oplus} , although smaller, still highlights the contextual knowledge's contribution to detecting positive class.

CHAPTER 7

APPLICATIONS: MAPPING TRUTHFULNESS STANCE OF SOCIAL MEDIA POSTS ON FACTUAL CLAIMS

This chapter extends the foundational work presented earlier by demonstrating three real-world applications. The first, **TrustMap**, is a comprehensive application that visualizes public stances on factual claims across geographical regions. The second, **ClimateStance**, is a focused case study on climate change that applies the truthfulness stance detection to analyze public perceptions within this domain. The third, the COVID-19 misinformation dashboard, is an application that aims to understand the surveillance of, impact of, and effective interventions against the COVID-19 misinfo-demic. Together, these applications demonstrate the flexibility and practical value of truthfulness stance detection in understanding and addressing misinformation.

TrustMap and **ClimateStance** share the same core technical workflow but differ in certain implementation details, reflecting their distinct objectives. The following sections outline the components they have in common and describe where their implementations diverge.

7.1 Overview of the Applications

In the digital era, the spread of factual claims, public narratives, and misinformation has intensified across multiple domains, including health (Suarez-Lledo and Alvarez-Galvez, 2021), the environment (Treen et al., 2020), and politics (Tucker et al., 2018). Social media platforms such as X (formerly Twitter) play an impor-

tant role in online discourse by enabling the broad spread of content (Bakshy et al., 2012; Zhang et al., 2024b; Shi et al., 2014). However, these platforms also facilitate the proliferation of misleading claims and conspiracy theories, which can significantly influence public opinions and decision-making processes (Allcott et al., 2019; Ausat, 2023; Yan et al., 2025). In addition, these online discussions not only reflect public sentiment but also affect perceptions of the truthfulness of factual claims, critically shaping awareness and engagement with societal issues.

Social media users frequently respond to factual claims by endorsing their veracity, disputing their validity, or expressing uncertainty. Understanding these responses, or truthfulness stances, is essential for analyzing how misinformation and public messaging shape discussions across different topics. While previous research on stance detection has explored misinformation in health (Hossain et al., 2020), news (Reddy et al., 2022), and other controversial topics (Zhang et al., 2024a), many studies have not incorporated geolocation-based stance analysis across diverse subject areas. The ability to analyze stance perceptions at regional and local levels is crucial for assessing the influence of narratives on different communities.

This chapter first demonstrates the **truthfulness stance map** (**TrustMap**), a system that visually presents stance distributions toward factual claims across geographical regions in the United States. **TrustMap** is built upon 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. Each claim-tweet pair is classified using RATSD (Zhu et al., 2025b). This model leverages fine-tuned LLMs and RAG (Lewis et al., 2020). It aggregates social media users' perceptions of truthfulness and classifies them into three stance categories—positive, negative, and neutral/no stance, based on the conceptual framework from RATSD (Zhu et al., 2025b). A positive stance indicates that a

post conveys support for a claim’s veracity, a negative stance suggests that the post disputes the claim as false, and a neutral/no stance signifies either uncertainty or a lack of explicit endorsement or refutation. The truthfulness stance detection results are rendered on **TrustMap**. In **TrustMap**, users can select one or multiple topics, choose factual claims within the selected topic(s), explore tweets related to the selected claims based on their truthfulness stance with respect to the selected claims, and visualize stance distributions across geographical areas and over time. The result analysis suggests that social media users often believe claims are true regardless of their actual veracity. From a topical perspective, users show the poorest judgment on claims related to the environment. Geographically, users in Florida demonstrate the greatest difficulty in distinguishing between true and false claims.

To the best of our knowledge, **TrustMap** is the first publicly available map application that visualizes truthfulness stances across different topics. We developed an interactive truthfulness stance map (<https://idir.uta.edu/trustmap>) that enables users to explore how stance distributions vary by topic and geographic region, down to the level of individual cities. A video demonstration is available at <https://vimeo.com/1094263767>. To support robust analysis, we also evaluated the underlying stance detection model and examined stance trends across topics and regions. Furthermore, we released the full codebase and dataset at <https://github.com/idirlab/trustmap> to encourage reproducibility and to facilitate future research on geographic patterns of truthfulness stance.

ClimateStance is a case study on climate change that applies the **TrustMap** workflow to analyze public perceptions on social media. It structures the assessments of truthfulness stance for climate-related factual claims across multiple sub-categories. The analysis highlights persistent challenges in distinguishing factual statements from

misinformation and reveals a general tendency to accept claims as true regardless of veracity.

The dashboard for mitigating the COVID-19 misinfodemic applies truthfulness stance detection to track and analyze misinformation during the COVID-19 pandemic. Unlike TrustMap and ClimateStance, it employs a distinct technical workflow, separate from the framework introduced below. The dashboard supports two core objectives: (1) tracking how misinformation spreads and (2) examining its relationship to pandemic trends. An interactive online dashboard (<https://idir.uta.edu/covid-19/>) enables users to explore case statistics along with prevailing facts and debunks, view official communications from local authorities, and access a chatbot designed to address COVID-19-related questions.

7.2 Related System

Although many existing systems focus on online narrative monitoring and analytics, most of them emphasize general data collection and analytics (Borra and Rieder, 2014; Zhang et al., 2024c) or popular domains such as sentiment analysis (Lien et al., 2022; Agarwal et al., 2018). The incorporation of truthfulness stance into interactive maps or dashboards remains a relatively new concept. Existing recent studies primarily explore stance types different from those examined in our research. StanceVis Prime (Kucher et al., 2020) is designed for the analysis and visualization of sentiment and stance in temporal text data from various social media sources. 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 is based on identifying seven different modifiers (Skeppstedt et al., 2017) in a given text, which differs from our definition of truthfulness stance. A discourse analysis dashboard for monitoring and analyzing online narratives was designed in

a recent study (Liew et al., 2024). This dashboard focuses on *sentiment*-oriented stance toward *general* topics such as vaccine side effects, which is also distinct from our focus on *truthfulness*-oriented stance toward *individual* claims.

Existing systems focusing on truthfulness stance detection primarily stem from our previous work. We developed a dashboard for the COVID-19 misinfodemic (Zhu et al., 2021), which identified stances toward COVID-19 related facts. However, stance detection was not the primary focus of that system, and the underlying method was less advanced than RATSD. In contrast, TrustMap supports more sophisticated stance detection and covers a broader range of factual claims, including both verified facts and misinformation across multiple topics beyond COVID-19. We also built a framework to understand social media users’ truthfulness stance toward claims across climate change-related topics (Zhang et al., 2024a), and we developed novel truthfulness stance detection methodologies (Zhu et al., 2025b). These previous works contribute to the foundation of TrustMap.

7.3 Overview of the Framework

Our framework for truthfulness stance detection applications follows a four-stage pipeline in its design: (1) **data ingestion**, (2) **taxonomy generation**, (3) **truthfulness stance classification**, and (4) **data exploration**. In the data ingestion stage, we collect factual claims from fact-checking websites and form claim-tweet pairs by querying X’s API for related tweets. In the taxonomy generation stage, we use an LLM with human-in-the-loop refinement to automatically construct taxonomies for collected factual claims. In the truthfulness stance classification stage, we apply RATSD to classify the truthfulness stance of each claim-tweet pair. Each claim-tweet pair is classified as either positive, negative, or neutral/no stance. In the data exploration stage, pairs with available geolocation information are grouped by

location, and stance distributions are displayed on an interactive map and charts. To enhance interpretability, the system also provides LLM-generated explanations for individual tweets and their associated stance labels.

The framework supports both general applications and domain-specific case studies. While TrustMap and ClimateStance share the same stance detection workflow and modeling backend, they differ in:

- **Domain Coverage:** TrustMap covers multiple domains (e.g., health, politics, education), whereas ClimateStance focuses exclusively on climate change.
- **Claim Source Variety:** TrustMap relies solely on PolitiFact and allows user-submitted claims for real-time exploration, while ClimateStance incorporates five fact-checking sources and offers a static claim set.
- **Research Emphasis:** TrustMap highlights regional variation and provides users with stance explanations; ClimateStance investigates topic-based accuracy in user judgments.

7.4 Framework Design

7.4.1 Data Ingestion

Factual Claim Collection. For TrustMap, we collected factual claims, denoted as \mathcal{C} , from Politifact using an in-house fact-check collection tool. Note that only PolitiFact claims are included for user exploration within TrustMap due to X API rate limits and interface constraints. PolitiFact is chosen because it provides the most comprehensive coverage, as explained in Chapter 3. For ClimateStance, \mathcal{C} was sourced from five fact-checking websites: PolitiFact (<https://politifact.com>), Snopes (<https://snopes.com>), Full Fact (<https://fullfact.org>), Metafact (<https://metafact.io>), and AP News (<https://apnews.com>). 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.5). We consider a claim c to be climate change-related if any of the keywords appear in c itself, its fact-checking article’s tags (i.e., topical labels such as “Coronavirus,” “Donald Trump,” or “Abortion”), 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.6). After removing duplicates, we obtained 1,409 unique climate change-related factual claims spanning from November 2007 to May 2024. To ensure \mathcal{C} is related to climate, the collected \mathcal{C} are manually reviewed. We find that claims collected using scenarios (1) and (2) are more directly related to climate compared to those collected using scenario (3), which are often indirectly related. For example, for the claim “AI is an existential threat to humanity,” the keyword “global warming” is present in an article that mentions AI proposing to regulate global warming by creating a nuclear winter scenario.

Tweet Collection. In TrustMap, we retrieved tweets T by constructing queries using a keyword-based retrieval strategy to extract key terms (nouns, verbs, and numbers) from the claims. Tweets T were collected over a defined time window: one month before and up to one year after the publication date of the fact-check corresponding to each claim c . Tweets shorter than 30 characters were excluded to reduce noise.

We used two strategies to collect geolocation information from the tweets. This information is key to rendering the 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]”. However, this strategy only returned limited data as X users rarely disclose their location in tweets. The second strategy involved collecting tweets and their metadata and then searching for geolocation information within the profiles of the X users who posted the tweets. The geolocation information was normalized using Geopy (GeoPy, 2023), which converts unstructured location text into structured fields such as city, county, state, and country. In total, 136,040 claim-related tweets were collected for 2,216 unique claims, of which 24,262 tweets include usable geolocation data. Figure 7.1 shows a heatmap illustrating the geographic distribution of collected tweets across U.S. states. Darker shades represent higher tweet volumes. States such as Texas, Florida, California, and New York show the highest levels of tweet activity. In contrast, several central and mountain states, including the Dakotas, Wyoming, and Montana, show relatively sparse activity.

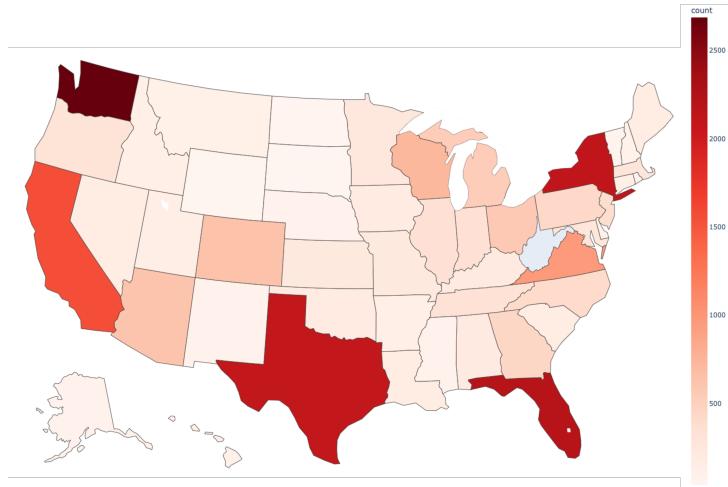


Figure 7.1: Geographic distribution of collected tweets across U.S. states.

`ClimateStance` applies the same claim–tweet pairing process as `TrustMap`, but the tweet collection is focused specifically on the curated set of climate-related factual claims. This process yielded a total of 13,050 tweets for 729 out of 1,409 climate-change-related claims. Among these 729 claims, 294 claims had more than 10 tweets.

7.4.2 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 factual claims. To minimize the manual effort, we prompt LLM, specifically Zephyr (Tunstall et al., 2023), to generate 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 a human-in-the-loop approach 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 $\{c_1, c_2, \dots, c_n\} \subset \mathcal{C}$ ($n = 7$ 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 A.4 in Appendix A.7.

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.

7.4.3 Truthfulness Stance Detection Model

For **TrustMap**, we applied RATSD to classify the stance of each tweet toward its associated factual claim. The technical details are provided in Chapter 5. For **ClimateStance**, we apply supervised fine-tuning on an LLM to build a classifier, leveraging Zephyr (Tunstall et al., 2023) as the underlying backbone LLM. The supervised fine-tuning mainly involves customizing Zephyr from a general LLM for text generation to a task-specific LLM for text classification.

7.4.4 Data Exploration

To bridge the gap between truthfulness stance detection results and user needs, **TrustMap** provides an interactive, user-oriented interface. As shown in Figure 7.2, the interface visualizes stance results on a map that supports filtering and exploration

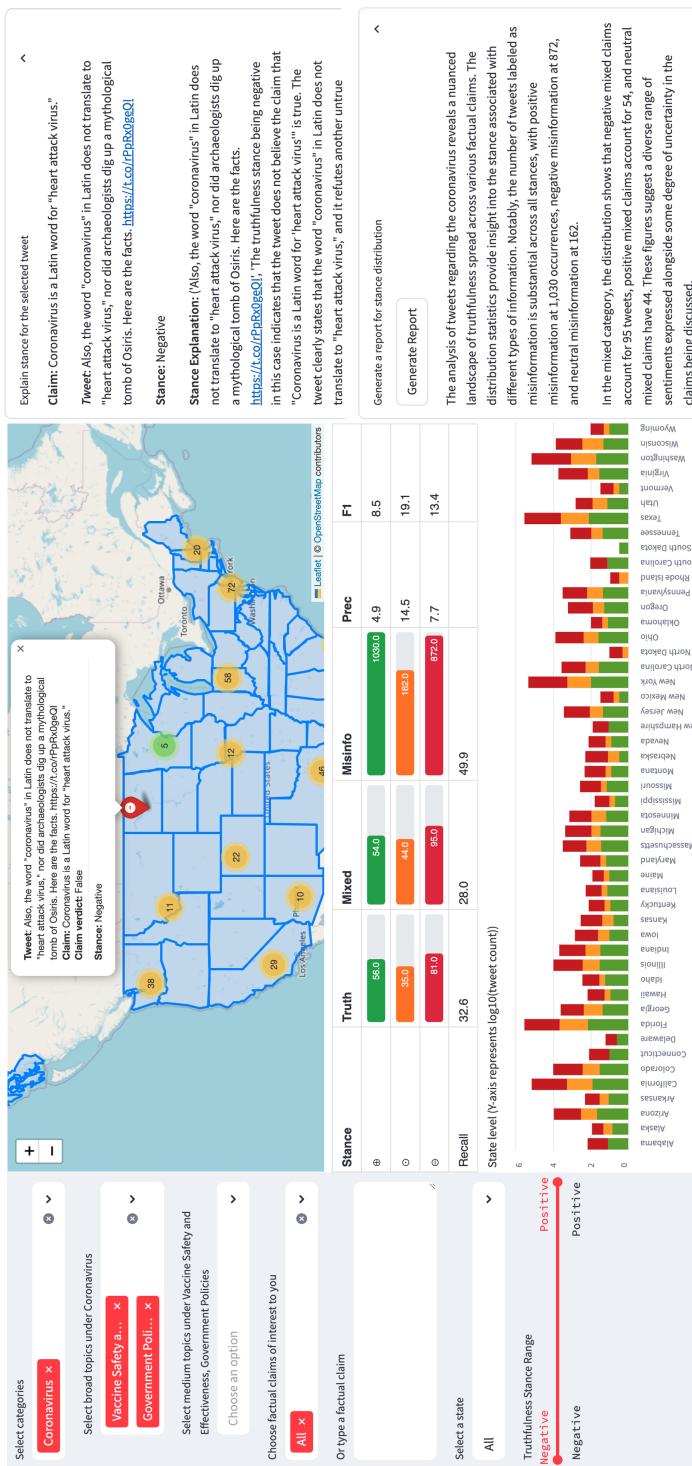


Figure 7.2: The overall user interface of TrustMap.

by region, topic, and claim. Built with Streamlit (<https://streamlit.io/>), it comprises four main components: a control panel, an interactive map view, statistical charts, and an LLM-based explanation panel. Each component is described below in terms of its user-facing functions and the underlying data and model logic. The following paragraphs provide detailed descriptions of each component.

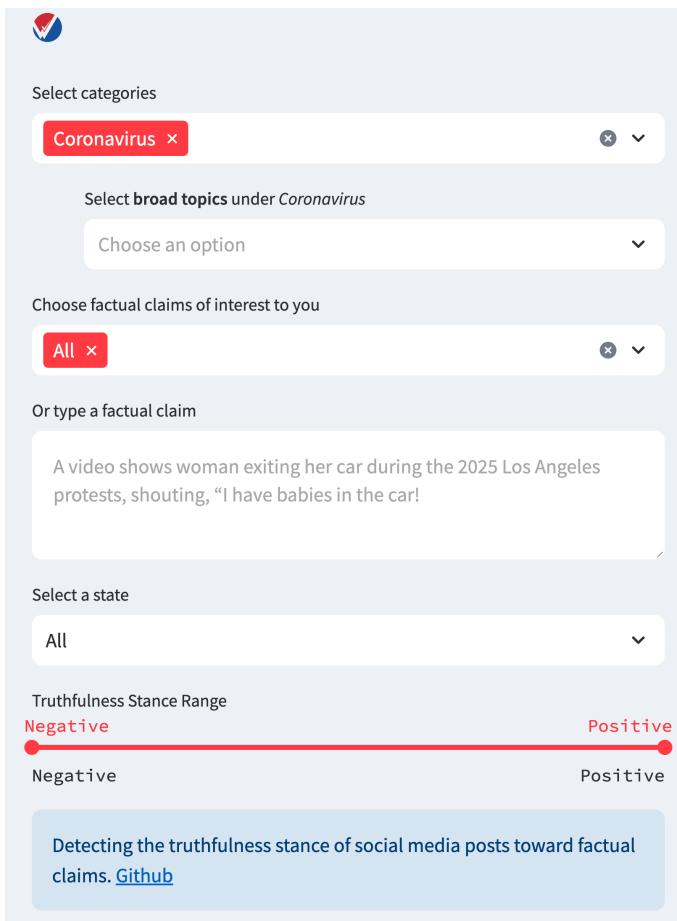


Figure 7.3: The control panel in TrustMap.

Control Panel. The control panel (Figure 7.3), located on the left side of the interface, enables users to select claims and topics for exploration. It supports hierarchical filtering: users first choose one or more high-level categories (e.g., topic domains), fol-

lowed by finer-grained subtopics generated using the taxonomy construction method described in Section 7.4.2. After a topic is selected, a dynamic multi-select dropdown lists all associated factual claims. Users can either select claims from these lists or manually enter new claims through a text input field. Whenever the selection changes, the system behaves differently depending on the action. Selecting topics or claims by checking boxes displays results from pre-collected, static data. In contrast, entering a new claim into the text field triggers a real-time query to X’s (Twitter’s) API, retrieves relevant tweets, and classifies each using the RATSD model. In both cases, the updated selections drive the backend pipeline to refresh the displayed results, including the map and charts.

Map View. The map view (Figure 7.4) plots tweets geographically on a U.S. map using latitude and longitude information when available. To reduce visual clutter, nearby tweets are aggregated into clusters that separate into individual markers as the user zooms in. The map is synchronized with the control panel: selecting a U.S. state (via dropdown or by clicking on the map) automatically zooms into that region and filters the displayed results accordingly. Marker colors or icons indicate the stance labels assigned by RATSD. For selections based on pre-collected, static data, the map instantly refreshes to display the corresponding subset. For claims entered in real time, the map is updated after the live query, tweet retrieval, and stance classification process are completed. Users can also filter displayed tweets by stance category (e.g., Truth- \oplus , Misinformation- \ominus) using a stance filter slider at the bottom of the map.

Tweet Pop-Up. Clicking an individual tweet marker opens a pop-up window (Figure 7.5) displaying detailed information about the claim–tweet pair. The pop-up contains the full text of both the claim and the tweet, along with metadata such

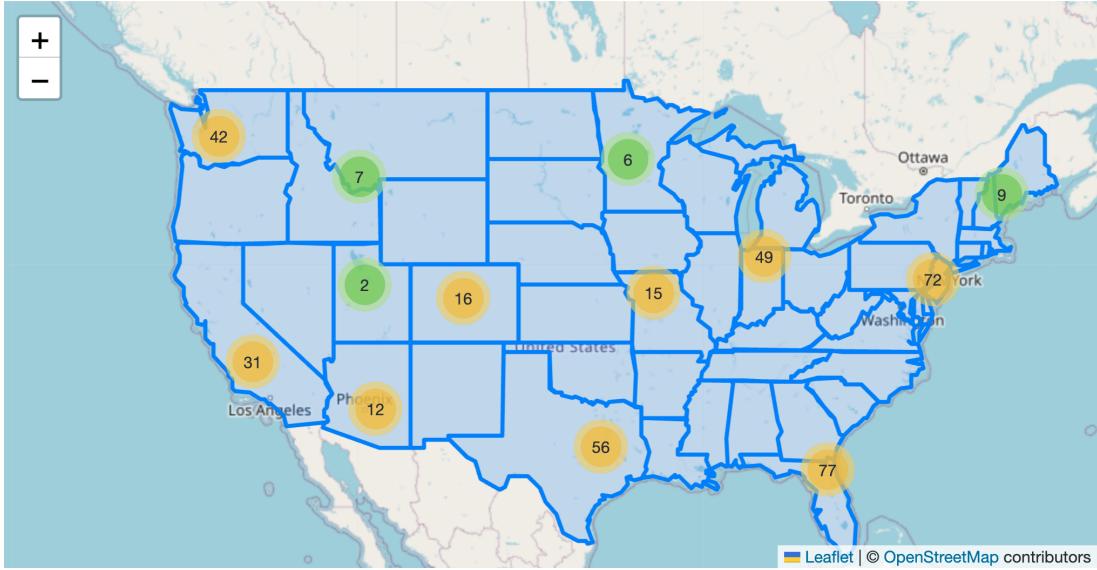


Figure 7.4: The map view in TrustMap.

as the tweet’s user-reported location (if available), the PolitiFact veracity label for the claim (e.g., “True,” “False,” “Pants on Fire”), and the stance label assigned by RATSD. For selections from static data, these details are retrieved directly from the precomputed dataset. For real-time queries, the pop-up displays the newly retrieved tweet text and metadata after stance classification is completed. This feature allows users to inspect individual examples in context without additional processing delays for static data.

Statistical Charts. Beneath the map, TrustMap presents two interactive statistical charts (Figure 7.6) summarizing stance results. The first is a bar chart showing the number of claim–tweet pairs in each stance category (Truth- \oplus , Truth- \ominus , Misinformation- \oplus , Misinformation- \ominus). It also reports the accuracy of social media users’ judgments—measured by precision, recall, and F1 score—by comparing their stances with the corresponding PolitiFact verdicts. This chart offers a concise view of both stance distribution and the degree to which users’ judgments align with those of

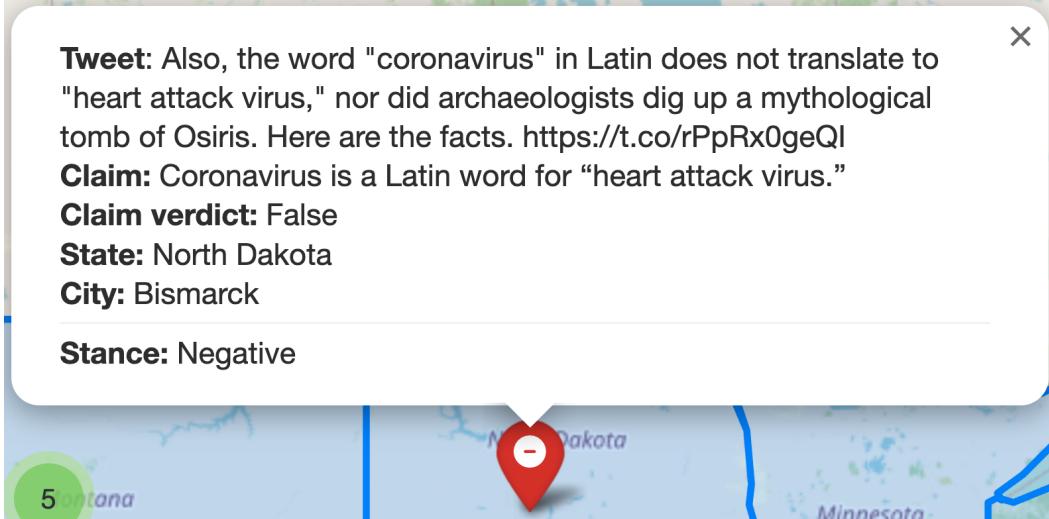


Figure 7.5: An example of a claim-tweet pair and its truthfulness stance displayed on TrustMap.

fact-checkers. The second chart visualizes the distribution of stance categories across geographic regions (such as all U.S. states). This helps users identify regional trends or polarization. Both charts are interactive: they update automatically when filters change (for example, if the user switches to a different claim or state, the counts and metrics will recompute).

Stance Explanation Panel. To support interpretation of model outputs, TrustMap includes an LLM-powered explanation panel with two complementary functions. First, for individual claim-tweet pairs, the user can click an “Generate Explanation” button located on the right-hand side of the application. This sends the claim, tweet, and stance result to a GPT-4o-mini model (OpenAI, 2024), which returns a natural-language rationale for the stance label. The generated rationale (Figure 7.7) then appears in a text area below the tweet content. This stance explanation panel, therefore, shows the reasoning behind each classification in natural language.

Stance	Truth	Mixed	Misinfo	Prec	F1
⊕	36.0	50.0	767.0	4.2	7.5
○	31.0	43.0	99.0	17.9	21.4
⊖	46.0	95.0	654.0	5.8	10.4
Recall	31.9	26.6	50.5		

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

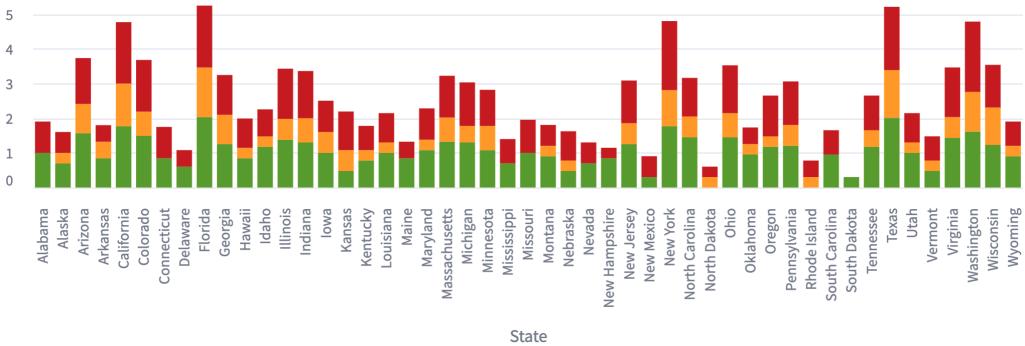


Figure 7.6: The statistical chart in TrustMap.

Second, TrustMap can generate higher-level narrative summaries (Figure 7.8) through a “Generate Report” button, the system collects overall statistics (such as the total number of tweets per stance, top topics discussed, and any notable regional differences) and feeds this information to the same GPT-4o-mini model. The model then composes a narrative report summarizing the public’s stance trends on the factual claim(s). This report is presented as a block of text in the interface. Generating the report involves no new classification; it only uses aggregate results from the RATSD outputs. The LLM serves to turn raw numbers and topics into a readable summary of what the data indicates.

Generate Explanation

Claim: "When President Biden took office ... there was no vaccine available."

Tweet: The White House is getting blasted over a blatantly false tweet stating "there was no vaccine available" when President Joe Biden took office. <https://t.co/oiH8u2wmhS>

Stance: Negative

Stance Explanation: ('The White House is getting blasted over a blatantly false tweet stating "there was no vaccine available" when President Joe Biden took office. <https://t.co/oiH8u2wmhS>', 'The tweet expresses a negative stance toward the claim that "there was no vaccine available" when President Biden took office. This implies that the tweet believes the claim is false.\n\nThe reasoning for this stance is based on historical facts regarding the COVID-19 vaccine. The first emergency use authorization for a COVID-19 vaccine was granted by the U.S. Food and Drug Administration (FDA) for the Pfizer-BioNTech vaccine on December 11, 2020. President Biden was inaugurated on January 20, 2021, meaning that vaccines were already available by that time. \n\nThus, the assertion made in the claim is inaccurate because vaccines were already being distributed and administered to the public prior to Biden's

Figure 7.7: The stance explanation for an individual tweet in TrustMap.

Generate a report for stance distribution

Generate Report

The analysis of the data related to the Coronavirus reveals a nuanced landscape of online discussions. The distribution of truthfulness stances across various factual claims illustrates a significant prevalence of misinformation. Specifically, the data indicates that there are 767 positive claims categorized as misinformation, 654 negative claims, and 99 neutral claims. This suggests a considerable amount of content that, while framed positively, does not adhere to factual accuracy.

In contrast, the categories of mixed claims show a lesser prominence, with 95 negative mixed claims and 50 positive mixed claims, indicating a reasonable degree of ambiguity in these assertions. Neutral mixed claims are relatively few, counting just 43. The totality of mixed assertions reveals the complexities faced by audiences in distinguishing between truth and misleading narratives.

When examining truthful claims, the trends indicate that there are 36 positive claims, 46 negative claims, and only 31 neutral claims. This presents an interesting paradox, where negativity about factual correctness is somewhat

Figure 7.8: The stance distribution explanation in TrustMap.

7.5 Results and Analysis

7.5.1 Results of Truthfulness Stance Detection in TrustMap

The RATSD model was trained on the early version of TSD-CT which consists of 2,220 labeled claim-tweet pairs. The dataset includes 1,262 positive stance examples (stance category denoted as \oplus), 451 neutral/no stance examples (\odot), and 507 negative stance examples (\ominus). We evaluated the performance of RATSD by comparing it to several state-of-the-art stance detection models, including fine-tuned LMs such as pre-trained models (e.g., BUT-FIT (Fajcik et al., 2019)), generative pre-trained models

Stance	Truth	Mixed	Misinfo	Precision	F1
\oplus	6,754	5,094	64,643	9.0	15.6
\odot	1,398	1,350	9,677	10.9	11.7
\ominus	3,494	4,453	39,177	83.1	48.8
Recall	58.0	12.4	34.6	-	-

Table 7.1: Distribution of X users’ truthfulness stances toward true, mixed, and false claims, along with the precision, recall, and F1-score for each stance category in TrustMap.

Topic	Truth- \oplus	Truth- \ominus	Misinfo- \oplus	Misinfo- \ominus	Accuracy	Macro F1
Public Health	77.2% (413)	22.8% (122)	53.8% (4,190)	46.2% (3,603)	61.7	39.3
Elections	40.0% (8)	60.0% (12)	58.9% (4,582)	41.1% (3,201)	40.6	29.3
Immigration	54.1% (53)	45.9% (45)	60.6% (3,023)	39.4% (1,969)	46.8	29.8
Economy	69.5% (534)	30.5% (234)	57.4% (1,174)	42.6% (872)	56.1	49.2
Abortion	79.9% (446)	20.1% (112)	44.3% (759)	55.7% (953)	67.8	59.6
Education	57.0% (231)	43.0% (174)	50.7% (849)	49.3% (827)	53.2	46.4
Crime	78.2% (223)	21.8% (62)	79.5% (1,030)	20.5% (265)	49.4	30.8
Environment	32.1% (131)	67.9% (277)	41.0% (661)	59.0% (953)	30.0	28.6
All	66.0% (6,754)	34.0% (3,494)	62.3% (64,643)	37.7% (39,177)	51.8	35.0

Table 7.2: Truthfulness stance distribution towards **Truth** and **Misinformation** across PolitiFact’s topics. Truth- \oplus and Truth- \ominus denote positive and negative stances towards **Truth**, respectively. Misinfo- \oplus and Misinfo- \ominus denote positive and negative stances towards **Misinformation**, respectively. Note that “All” includes every topic, not just the selected eight, so the total count for the eight topics does not equal the count for “All.”

(e.g., BLCU_NLP (Yang et al., 2019a)), and domain-adaptive pre-trained models (e.g., BERTSCORE+NLI (Hossain et al., 2020), BART+NLI (Reddy et al., 2022), and TESTED (Arakelyan et al., 2023)). In RATSD, we utilize two fine-tuned LLMs as alternative choices: the open-source model Zephyr (Tunstall et al., 2023) and the proprietary model GPT-3.5 (Brown et al., 2020). We used F1 scores for each class—denoted as F_{\oplus} , F_{\odot} , and F_{\ominus} —and the macro F1 score (F_M) as evaluation metrics. The evaluation results in Table 6.2 from Chapter 6 show that RATSD achieves strong performance across all stance categories.

In TrustMap, the RATSD model was applied to the collected claim-tweet pairs to predict stance categories. Given the model’s strong accuracy in stance detection, we rely on the predicted stance categories to evaluate the accuracy of X users’ judgments on claims. More specifically, given each claim-tweet pair, we compared PolitiFact’s veracity verdict for the claim with the stance expressed in the tweet regarding the claim’s veracity. PolitiFact assigns one of six verdicts to each claim: “True,” “Mostly True,” “Half True,” “Mostly False,” “False,” and “Pants on Fire.” For simplicity and clarity, as well as for avoiding overly small fragments of data, we mapped the first two to “Truth,” “Half True” to “Mixed,” and the latter three to “Misinformation.” A claim-tweet pair’s stance is considered accurate if the stance aligns with the claim’s veracity—e.g., a positive stance toward truth, a negative stance toward misinformation, or a neutral/no stance toward a mixed-veracity claim. Following this rule, we computed the precision and F1 score of users’ judgments for each stance category, as well as the recall for each verdict category of the claims (Table 7.1). These scores illustrate the difficulty X users face in distinguishing between truth and misinformation.

The results show a strong tendency for X users to believe claims are true, regardless of their actual veracity. These findings align with recent research (Moravec et al., 2019; Zhang et al., 2024a). More specifically, nearly 57.0% ($\frac{64,643}{64,643+9,677+39,177}$) of the misinformation is believed to be true by the users, and the recall for Misinfo is only 0.346. Furthermore, users show significant skepticism even toward true claims, as only 58.0% of the tweets about true claims express a positive stance, and over a third believe they are false. Mixed-veracity claims (those labeled “Half True”) reveal another concerning pattern. Users rarely express neutral or uncertain stances toward such claims. Instead, they are often polarized, leaning toward either accepting them as true or rejecting them as false.

7.5.2 Results of Truthfulness Stance Detection in ClimateStance

	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 7.3: Performance of truthfulness stance classifier on the annotated dataset.

Truth- \oplus	Neutral- \odot	Misinfo- \ominus	Total
8,003 (61.33%)	2,668 (20.44%)	2,379 (18.23%)	13,050

Table 7.4: Distribution of tweets by truthfulness stance toward claims.

As shown in Table 7.3, 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 7.4 reveals that the majority (8,003 out of 13,050 tweets) believe that the claims are true.

7.5.3 Result of Taxonomy Generation in ClimateStance

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

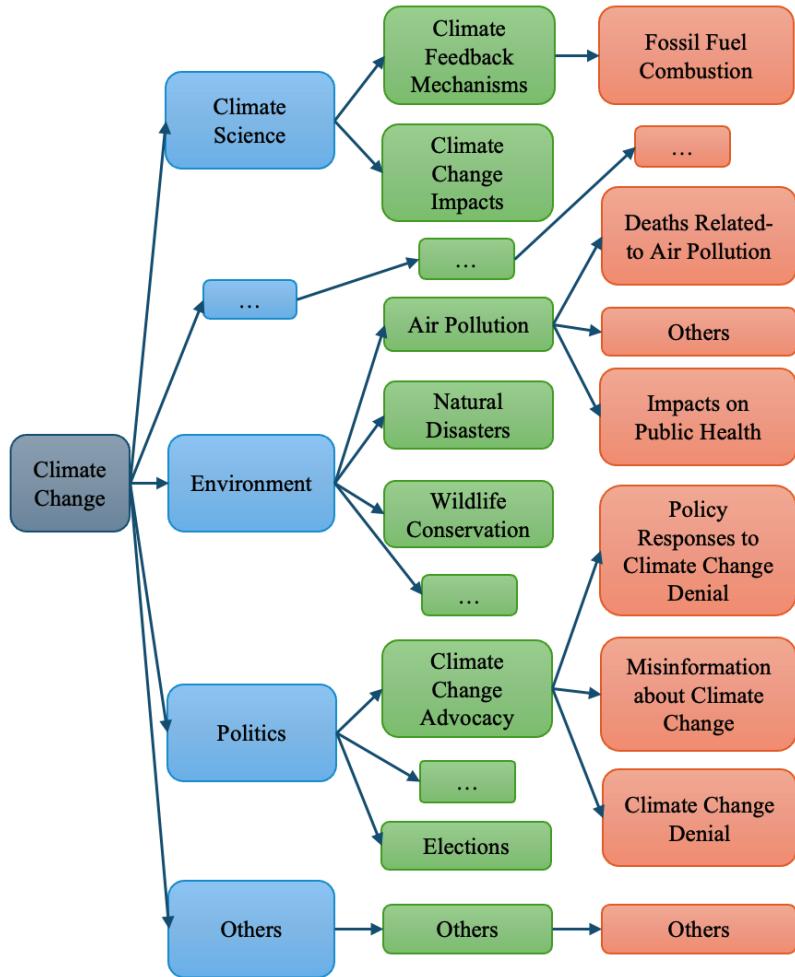


Figure 7.9: A fragment of climate change taxonomy.

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 Pollution*” 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 7.9.

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.

7.5.4 Detailed Analysis for TrustMap

TrustMap can be useful in identifying topics and regions where X users' perceptions toward true and false claims diverge from the veracity of claims themselves. Such insights can help guide more targeted fact-checking and public awareness efforts. Therefore, we conducted analyses from both topic and geographical perspectives. For these analyses, we excluded a claim-tweet pair if the claim has a “Mixed” verdict and/or the tweet is classified as \odot . This enables us to derive clearer insights without accounting for the less frequent and more ambiguous stance and verdict categories. Moreover, we produced results separately for each of the top eight states in terms of number of claim-tweet pairs. Similarly, we produced results for top eight topics. Note that while topics such as “National,” “Space,” and “Federal Budget” are populous, they were not analyzed individually because they are either highly imbalanced (e.g., containing only misinformation with no truth) or do not truly constitute distinct topics.

Region	Truth-⊕	Truth-⊖	Misinfo-⊕	Misinfo-⊖	Accuracy	Macro 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
United States	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 7.5: Truthfulness stance distribution toward **Truth** and **Misinformation** across U.S. states.

Topic-level analysis. Table 7.2 shows how X users perceive factual claims across different topics. For each topic, we report the proportions of positive and negative

Political Leaning	Truth-⊕	Truth-⊖	Misinfo-⊕	Misinfo-⊖	Accuracy	Macro F1
Red States	67.9% (527)	32.1% (249)	63.7% (4,344)	36.3% (2,476)	52.1	35.3
Blue States	70.8% (659)	29.2% (272)	55.6% (4,259)	44.4% (3,402)	57.6	41.3
Swing States	73.4% (257)	26.6% (93)	62.5% (1,513)	37.5% (908)	55.5	38.7

Table 7.6: Truthfulness stance distribution by political leaning.

stances toward both true and false claims, as well as the accuracy and macro F1 score to assess alignment between user stance and claim veracity. The results show that in the topic of “*Crime*,” users express high levels of belief in both true and false claims, with 78.2% for Truth-⊕ and even the higher 79.5% for Misinfo-⊕. Conversely, the topic “*Environment*” shows the highest rate of skepticism, with 67.9% and 59.0% of tweets rejecting truth and misinformation, respectively.

Across topics, accuracy and macro F1 scores vary, but most fall below 0.5. “*Abortion*” has the highest scores (Accuracy: 67.8, F1: 59.6), indicating relatively better user judgment in this domain. In contrast, “*Environment*” and “*Elections*” exhibit the lowest F1 scores (28.6 and 29.3), suggesting limited user ability to differentiate between true and false claims related to these topics.

Geographical-level analysis. Similarly, we analyzed X users’ stance by geographical location. As shown in Table 7.5, both tweet count and percentage for Misinfo-⊕ reach the highest (66.6%) in Florida, meaning that misinformation is widely spread in this region. In addition, Florida X users exhibit 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-⊕ (42.7%), indicating in Washington relatively more X users push back against misinformation. Even in higher-performing states such as Washington, however, F1 scores remain modest. This suggests that, while some regional variation exists, users overall struggle to discern the truthfulness of claims. This observation is reinforced by the national-

level (“United States”) accuracy at 55.0 and macro F1 at just 38.3. When comparing “United States” with “All”, which includes claim-tweet pairs from within the U.S., outside the U.S., and those without geolocation, we find that users in the United States perform slightly better at distinguishing between true and false claims, as reflected in their higher accuracy and macro F1.

We also uncovered an interesting finding regarding political leaning. Similar to Table 7.5, Table 7.6 presents the distributions of truthfulness stances across “red” (i.e., leaning toward Republicans), “blue” (i.e., leaning toward Democrats), and “swing” (i.e., no leaning) states. The classification of red, blue, and swing states is based on information from Wikipedia (https://en.wikipedia.org/wiki/Red_states_and_blue_states). The table shows X users in blue states are better at distinguishing between true and false information compared to those in red states. This result is consistent with the findings from prior research in political science and psychology (Garrett and Bond, 2021; Dobbs et al., 2023; Spampatti et al., 2024; Zhu and Pechmann, 2025).

Our analyses highlight a consistent pattern: X users frequently believe claims are true regardless of their actual veracity. These findings align with recent research (Moravec et al., 2019; Zhang et al., 2024a) showing that individuals often fail to distinguish between real and fake content. In addition, in both topic- and state-level evaluations, we found that X users’ judgments of factual claims vary across topics and regions.

7.5.5 Detailed Analysis for Climate Stance

In the final results, as indicated in Table 7.7, each (c, p) pair is associated with a stance, a broad topic, 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 7.8. 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 \ominus , as they provide less meaningful insights.

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.	\oplus	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.	\odot	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.	\ominus	Climate Science	Climate Feedback Mechanisms	Misconceptions

Table 7.7: Examples of truthfulness stance detection and their corresponding topics in the taxonomy.

The high percentage of both Truth- \oplus and Misi- \oplus 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- \oplus (88.7%) and Misi- \oplus (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 7.8), 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 A.1 in Appendix A.8. 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 in a social science study (Moravec et al., 2019), which suggests that social media users have difficulty detecting fake news and that most users would make more accurate judgments by simply flipping a coin.

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

7.6 Limitations and Ethics Statement of TrustMap

While TrustMap offers a novel interface to visualize truthfulness stance toward factual claims across the U.S., it has several limitations. The reliance on X's API introduces data access constraints. Due to API rate limits and restrictions on histor-

ical tweet access, the dataset may not comprehensively capture all relevant tweets for every claim. Additionally, many tweets lack precise geolocation data, which will lead to potential sampling bias. The geolocation inference method based on user profiles introduces further noise, as profile locations are user-defined and may not reflect actual user locations. Additionally, while **TrustMap** presents static snapshots of public discourse, it does not support real-time data updates, which would be necessary for early misinformation detection and timely interventions.

This application raises ethical considerations related to user privacy and content sensitivity. Although all analyzed data is publicly available from X, user tweets may contain sensitive or personally identifiable content. We take steps to minimize exposure of private information by displaying only essential tweet content and omitting usernames, user profiles, and direct user identifiers. Additionally, stance classification results are generated automatically using LLMs. These predictions may not always reflect the intent of the original X users and could misrepresent X users' views. Users of the system should interpret the results as aggregated patterns, not definitive assessments of individual posts. Finally, while **TrustMap** is intended to support fact-checking and public awareness efforts, there is a risk of misuse if the system is interpreted as providing a definitive judgment on the truthfulness of claims. Our tool is designed to support transparency and encourage informed evaluation, not to police online discourse.

7.7 A Dashboard for Mitigating the COVID-19 Misinfodemic

This section introduces a preliminary application of truthfulness stance detection, focusing on the COVID-19 infodemic. Unlike the applications discussed in the previous sections, this application employs a distinct technical workflow and serves

a different conceptual purpose. Therefore, it is presented separately. The following sections describe the details of the dashboard for mitigating the COVID-19 misinfodemic.

7.7.1 Background

Alongside the COVID-19 pandemic, there is a raging global misinfodemic (Mian and Khan, 2020; Roozenbeek et al., 2020) just as deadly. As fear grows, false information related to the pandemic goes viral on social media and threatens to affect an overwhelmed population. Such misinformation misleads the public on how the virus is transmitted, how authorities and people are responding to the pandemic, as well as its symptoms, treatments, and so on. This onslaught exacerbates the vicious impact of the virus, as the misinformation drowns out credible information, interferes with measures to contain the outbreak, depletes resources needed by those at risk, and overloads the health care system. Although health misinformation is not new (Oyeyemi et al., 2014), such a dangerous interplay between a pandemic and a misinfodemic is unprecedented. It calls for studying not only the outbreak but also its related misinformation; the fight on these two fronts must go hand-in-hand.

This demo case study describes the current milestones achieved in our ongoing project that aims to understand the surveillance of, impact of, and effective interventions against the COVID-19 misinfodemic. 1) For *surveillance*, we seek to discover the patterns by which different types of COVID-19 misinformation spread. 2) To understand the *impact* of misinformation, we aim to compare the spreading of the SARS-CoV-2 virus and misinformation and derive their correlations. 3) To understand what types of *interventions* are effective in containing misinformation, we will contrast the spreading of misinformation before and after debunking efforts. 4) To understand whether the outcomes related to 1), 2) and 3) differ by geographical lo-

cations and demographic groups, we will study the variability of misinformation and debunking efforts across geographical and demographic groups.

While we continue to pursue these directions, we have built an online dashboard at <https://idir.uta.edu/covid-19/> to directly benefit the public. A screencast video of the dashboard is at bit.ly/3c6v5xf. The dashboard provides a map, a navigation panel, and timeline charts for looking up numbers of cases, deaths, and recoveries, similar to a number of COVID-19 tracking dashboards.¹²³ However, our dashboard also provides several features not found in other places. 1) It displays the most prevalent factual information among Twitter users in any user-selected U.S. geographic region. 2) The “factual information” comes from a catalog that we manually curated. It includes statements from authoritative organizations, verdicts, debunks, and explanations of (potentially false) factual claims from fact-checking websites, and FAQs from credible sources. The catalog’s entries are further organized into a taxonomy. For simplicity, we refer to it as the *catalog and taxonomy of COVID-19 facts* or just *facts* in ensuing discussion. 3) The dashboard displays COVID-19 related tweets from local authorities of user-selected geographic regions. 4) It embeds a chatbot built specifically for COVID-19 related questions. 5) It shows case-statistics from several popular sources which sometimes differ.

The codebase of the dashboard’s frontend, backend, and data collection tools are open-sourced at <https://github.com/idirlab/covid19>. All collected data are available at <https://github.com/idirlab/covid19data>. Particularly, the catalog and taxonomy of facts are also available through a SPARQL endpoint hosted on the lab server at <https://cokn.org/deliverables/7-covid19-kg/> and the corresponding RDF dataset can be requested there.

¹<https://www.covid19-trials.com/>

²<https://coronavirus.jhu.edu/map.html>

³<https://www.cdc.gov/covid-data-tracker/index.html>

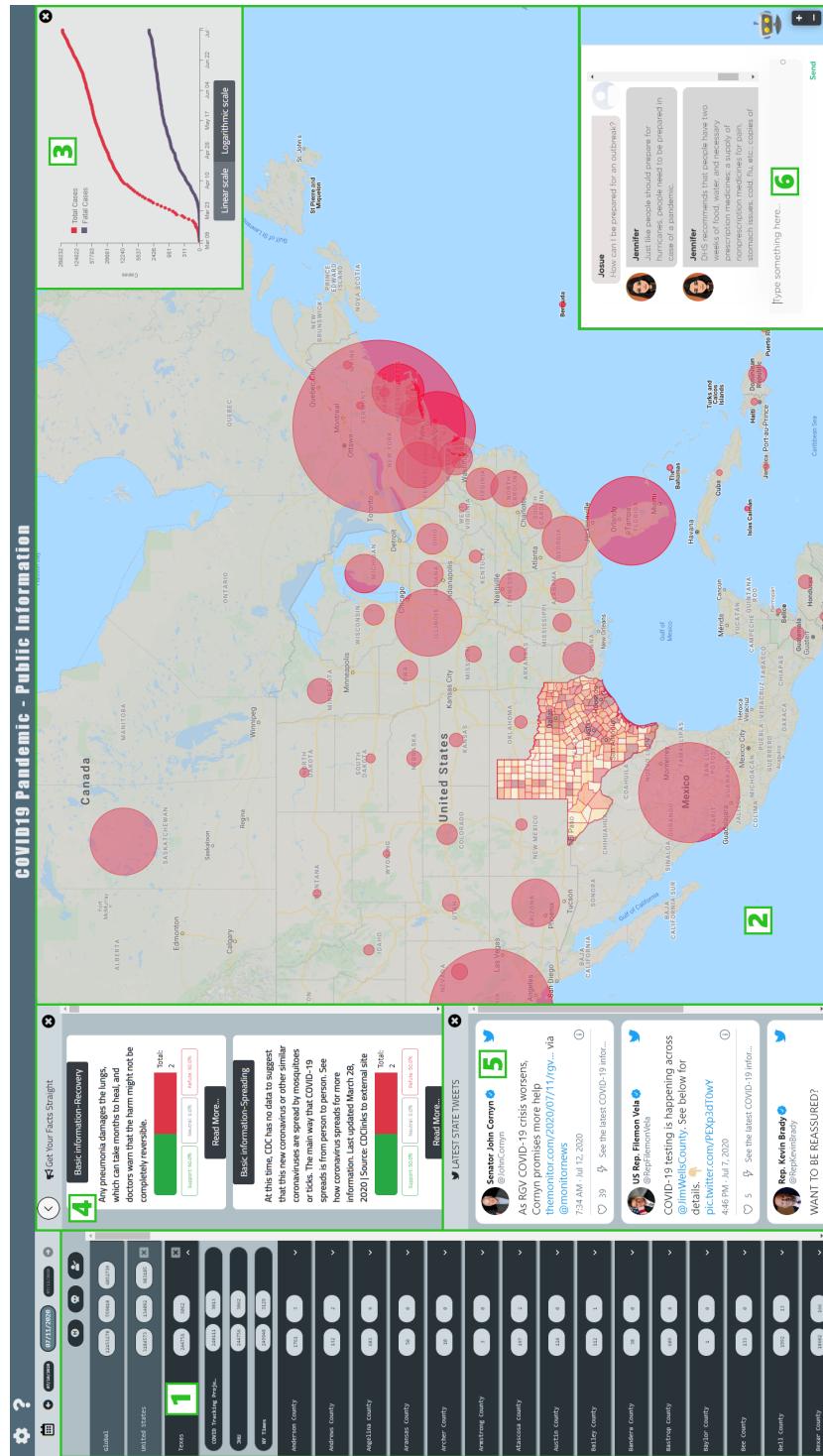


Figure 7.10: The user interface of the dashboard for mitigating the COVID-19 misinfodemic.

What is particularly worth noting about the underlying implementation of the dashboard is the adaptation of state-of-the-art textual semantic similarity and stance detection models. Tweets are first passed through a **claim-matching** model, which selects the tweets that semantically match the facts in our catalog. Then, the **stance detection** model determines whether the tweets agree with, disagree with, or merely discuss these facts. This enables us to pinpoint pieces of misinformation (i.e., tweets that disagree with known facts) and analyze their spread.

A few studies analyzed and quantified the spread of COVID-19 misinformation on Twitter (Kouzy et al., 2020; Memon and Carley, 2020; Al-Rakhami and Al-Amri, 2020) and other social media platforms (Brennen et al., 2020). However, these studies conducted mostly manual inspection of small datasets, while our system automatically sifts through millions of tweets and matches tweets with our catalog of facts.

7.7.2 The Dashboard

Figure 7.10 shows the dashboard’s user interface, with its components highlighted.

Geographic region selection panel (Component 1). A user can select a specific country, a U.S. state, or a U.S. county by using this panel or the interactive map (Component 2). Once a region is selected, the panel shows the counts of confirmed cases, deaths and recovered cases for the region in collapsed or expanded modes. When a region is expanded by the user, counts from all available sources are displayed; on the other hand, if it is collapsed, only counts from the default (which the user can customize) data source are displayed. These sources do not provide identical numbers.

Interactive map (Component 2). On each country and each U.S. state, a red circle is displayed, with an area size proportional to its number of confirmed cases.

When a state is selected, the circle is replaced with its counties' polygons in different shades of red, proportional to the counties' confirmed cases.

Timeline chart (Component 3). It plots the counts of the selected region over time and can be viewed in linear or logarithmic scale.

Panel of facts (Component 4). For the selected region, this panel displays facts from our catalog, and the distribution of people discussing, agreeing, or disagreeing with them on Twitter. A large number of people refuting these facts would indicate wide spread of misinformation. To avoid repeating misconceptions, the dashboard displays facts from authoritative sources only.

Government tweets (Component 5). It displays COVID-19 related tweets in the past seven days from officials of the user-selected geographic region. These tweets are from a curated list of 3,744 Twitter handles that belong to governments, officials, and public health authorities at U.S. federal and state levels.

Chatbot (Component 6). This component embeds the *Jennifer Chatbot* built by the New Voices project of the National Academies of Sciences, Engineering and Medicine (Li et al., 2020), which was built specifically for answering COVID-19 related questions. As part of the collaborative team behind this chatbot, we are expanding it using the aforementioned catalog.

7.7.3 The Datasets

The dashboard uses the following three datasets.

1) Counts of confirmed cases, deaths, and recoveries.

We collected these counts daily from Johns Hopkins University (<https://github.com/CSSEGISandData/COVID-19>), the New York Times (NYT) (<https://www.nytimes.com/2020/03/12/us/coronavirus-us-cases.html>).

[//github.com/nytimes/covid-19-data](https://github.com/nytimes/covid-19-data)), and the COVID Tracking Project (<https://covidtracking.com/>). These sources provide statistics at various geographic granularities (country, state, county).

2) Tweets.

We are using a collection of approximately 250 million COVID-19 related tweets from January 1st, 2020 to May 16th, 2020, obtained from (Banda et al., 2020) (version 10.0). We removed tweets and Twitter handles (and their tweets) that do not have location information, resulting in 34.6 million remaining tweets. We then randomly selected 10.4% of each month's tweets, leading to 3.6 million remaining tweets. We used the OpenStreetMap (Quinon et al., 2020) API to map the locations of Twitter accounts from user-entered free text to U.S. county names. We used the ArcGIS API (<https://developers.arcgis.com/python/guide/reverse-geocoding/>) to map the locations of tweets from longitude/latitude to counties.

3) A catalog and a taxonomy of COVID-19 related facts. The manually curated catalog currently has 9,512 entries from 21 credible websites, including statements from authoritative organizations (e.g., WHO, CDC), verdicts, debunks, and explanations of factual claims (of which the truthfulness varies) from fact-checking websites (e.g., the IFCN CoronaVirusFacts Alliance Database (<https://www.poynter.org/ifcn-covid-19-misinformation/>), PolitiFact), and FAQs both from credible sources (e.g., FDA, NYT) and a dataset curated by (Wei et al., 2020).

We organized the entries in this catalog into a taxonomy of categories, by integrating and consolidating the available categories from a number of source websites, placing entries from other websites into these categories or creating new categories, and organizing the categories into a hierarchical structure based on their inclusion relationship. The taxonomy is as follows, in the format of {level-1 cat-

egories [*level-2 categories* (*level-3 categories*)]:⁴ {Animals, Basic Information [*Causes, Definition, Disease Alongside, Recovery, Spreading, Symptoms, Testing*], Cases, Contribution, Diplomacy, Economics/Finance [*Crisis, Grants/Stimulus, Tax, Unemployment*], Family Preparation, Funeral, Government Control [*Administration (Lockdown, Reopen, Staff), Law, Medical Support, Military*], Mental Health, Prevention [*Actions to Prevent (Hand Hygiene, Isolation, Masks, Social Distancing), Medication, Vaccines*], Religion, Schools/Universities, Travel, Treatment [*Medication, Minor Symptom, Severe Symptom*], Violence/Crime}.

We also stored the catalog and the taxonomy as an RDF dataset, in which each entry of the catalog is identified by a unique resource identifier (URI). It is connected to a mediator node that represents the multiary relation associated with the entry. For example, Figure 7.12 shows a question about COVID-19, its answer and source, and the lowest-level taxonomy nodes that the entry belongs to, all connected to a mediator node. This RDF dataset, with 12 relations and 78,495 triples, is published in four popular RDF formats—N-Triples, Turtle, N3, and RDF/XML. Furthermore, we have set up a SPARQL query endpoint at <https://cokn.org/deliverables/7-covid19-kg/> using OpenLink Virtuoso (<https://virtuoso.openlinksw.com/>).

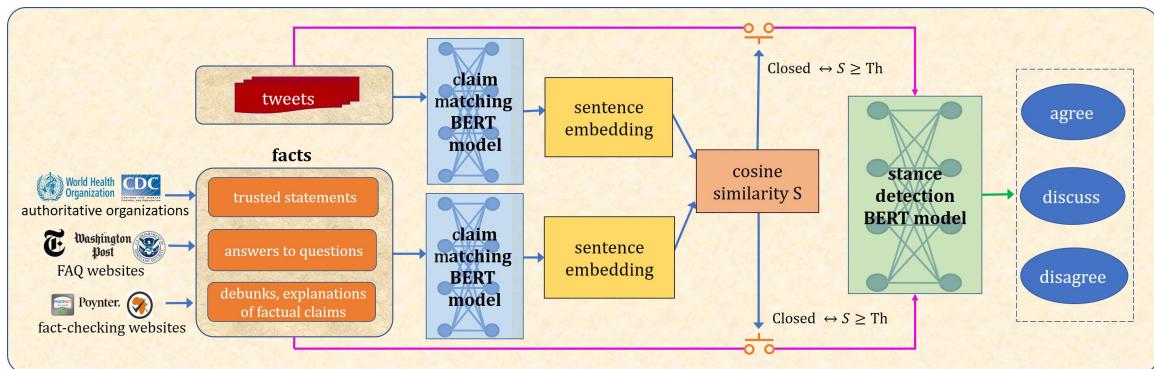


Figure 7.11: Matching tweets with facts and stance detection.

⁴Not every level-1 or level-2 category has subcategories.

Tweet	Fact	Taxonomy Categories	Similarity	Stance
Coronavirus cannot be passed by dogs or cats but they can test positive.	There has been no evidence that pets such as dogs or cats can spread the coronavirus.	Animals, Spreading	0.817	agree
More people die from the flu in the U.S. in 1 day than have died of the Coronavirus across the world ever.	Right now, it appears that COVID-19, the disease caused by the new coronavirus, causes more cases of severe disease and more deaths than the seasonal flu.	Cases	0.816	disagree

Table 7.9: Example results of matching tweets with facts and stance detection.

7.7.4 Matching Tweets with Facts and Stance Detection

Given the catalog of COVID-19 related facts F and the tweets T , we first employ **claim-matching** to locate a set of tweets $\mathbf{t}^f \in T$ that discuss each fact $f \in F$. Next, we apply **stance detection** on pairs $\mathbf{p}^f = \{(t, f) \mid t \in \mathbf{t}^f\}$ to determine whether each t is agreeing with, disagreeing with, or neutrally discussing f . Finally, aggregate results are displayed on Component 4 of the dashboard to summarize the public's view on each fact. Figure 7.11 depicts the overall claim-matching and stance detection pipeline. For both tasks, we employed Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019). Table 7.9 shows some example results of claim matching and stance detection.

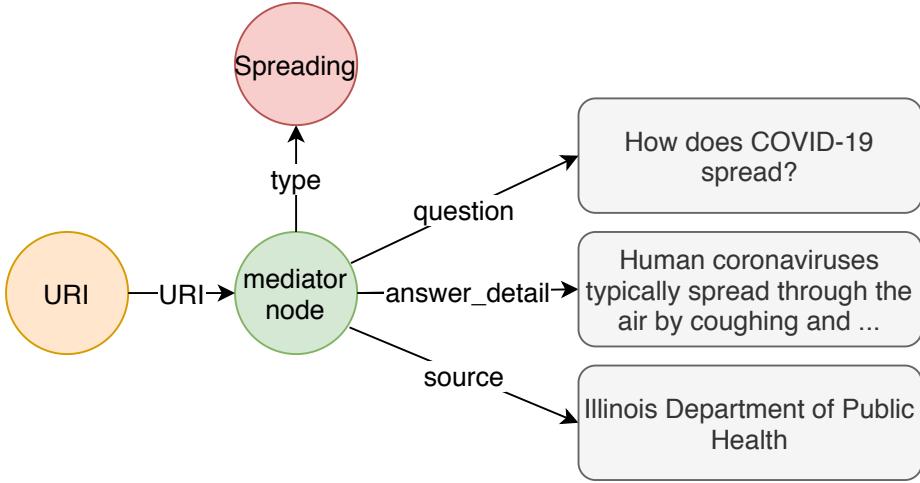


Figure 7.12: An entry of the catalog stored in RDF.

Claim Matching. We generate sentence embeddings \mathbf{s}^t and \mathbf{s}^f , for t and f respectively, using the mean-tokens pooling strategy in Sentence-BERT (Reimers and Gurevych, 2019). The relevance between t and f is then calculated as:

$$R^{t,f} = \frac{\mathbf{s}^t \cdot \mathbf{s}^f}{\|\mathbf{s}^t\| \times \|\mathbf{s}^f\|} \quad (7.1)$$

Given $R^{t,f}$, we model claim-matching as a ranking task on the relevance between facts and tweets. Thus, the output of this stage is $\mathbf{t}^f = \{t \in T \mid R^{t,f} \geq \theta\}$ for each fact $f \in F$, where the threshold θ is 0.8 in our implementation.

Stance Detection. Given \mathbf{t}^f , we detect the stance that each tweet t takes toward fact f . There could be 3 classes of stance: agree (t supports f), discuss (t neutrally discusses f), and disagree (t refutes f). For this task, we obtained a pre-trained BERT_{Base} model (<https://github.com/google-research/bert>) and trained it on the Fake-News Challenge Stage 1 (FNC-1) dataset (<http://www.fakenewschallenge.org/>). We denote this model Stance-BERT.

We first pre-process \mathbf{p}^f to conform with BERT input conventions by 1) applying $W(\cdot)$, the WordPiece tokenizer (Wu et al., 2016), 2) applying $C(a_1, a_2, \dots, a_n)$, a function that concatenates arguments in appearance order, and 3) inserting specialized BERT tokens [CLS] and [SEP]. Since BERT has a maximum input length of $M = 512$ and some facts can exceed this limit, we propose a sliding-window approach inspired by (Devlin et al., 2019) to form input \mathbf{x}^f :

$$\mathbf{x}^f = \left\{ \left\{ C([\text{CLS}], W(t), [\text{SEP}], W(f)_{[i*S, i*S+L]}, [\text{SEP}]) \mid 0 \leq i < \left\lceil \frac{|W(f)|}{S} \right\rceil \right\} \mid (t, f) \in \mathbf{p}^f \right\} \quad (7.2)$$

where S defines the distance between successive windows and $L = M - (|W(t)| + 3)$ is the sequence length available for each fact. If $i * S + L$ is an out-of-bounds index for $W(f)$, the extra space is padded using null tokens.

Each element $\mathbf{w} \in \mathbf{x}^f$ contains a set of windows representing a tweet-fact pair. Each window $w_i \in \mathbf{w}$ is passed into Stance-BERT, which returns probability distributions (each containing 3 entries, 1 for each class) $\hat{\mathbf{y}}_{w_i}^f$ for each window.

Stance aggregation. For each fact f , the stance detection results are accumulated to generate scores S_C^f , where $C \in \{\text{agree}, \text{discuss}, \text{disagree}\}$ that denote the percentage of tweets that agree, discuss, and disagree with f :⁵

$$S_C^f = \frac{\sum_{\mathbf{w} \in \mathbf{x}^f} [\text{argmax } \sigma(\{\hat{\mathbf{y}}_{w_i}^f \mid w_i \in \mathbf{w}\}) = C]}{|\mathbf{x}^f|} \quad (7.3)$$

⁵We use the Iverson bracket: $[P] = 1$ if P is true, else 0

where $\sigma(\cdot)$ is a function that averages the model’s output scores for each class across all windows of the tweet-fact pair. The three final stance scores are passed to the dashboard’s panel of facts (Component 4) for display.

7.7.5 Evaluation and Results

Performance of Claim Matching. To evaluate the performance of the claim matching component, we first created a Cartesian product of the 3.6 million tweets with 500 “facts” from the catalog (see Section 7.7.3 for description of datasets), followed by randomly selecting 800 tweet-fact pairs from the Cartesian product. To retain a balanced dataset, 400 pairs were drawn from those pairs scored over 0.8 by the claim matching component, and another 400 pairs were drawn from the rest. To obtain the gold-standard labels on these 800 pairs, we used three human annotators. 183 pairs were labeled “matched” (i.e., the tweet and the fact have matching topics) and 617 pairs “unmatched.” Table 7.10 shows the claim matching component’s performance on these 800 pairs, measured by precision@ k and nDCG@ k (normalized Discounted Cumulative Gain at k). Both precision@ k and nDCG@ k are metrics of ranking widely used in classification problems; the order of the top k predictions is considered in nDCG@ k but not in precision@ k .

Metric	@5	@10	@20	@50	@100
Precision	0.80	0.80	0.70	0.56	0.52
nDCG	0.62	0.72	0.78	0.81	0.83

Table 7.10: Performance of claim matching on the 800 tweet-fact pairs.

Performance of Stance-BERT.

Model	F1 score			
	agree	discuss	disagree	macro
Stance-BERT _{window} (FNC-1)	0.65	0.45	0.84	0.65
Stance-BERT _{trunc} (FNC-1) (Xu et al., 2018) (FNC-1)	0.66	0.41	0.82	0.63
Stance-BERT _{window} (COVID-19)	0.55	0.15	0.73	0.48
Stance-BERT _{window} (COVID-19)	0.75	0.03	0.58	0.45

Table 7.11: Performance of Stance-BERT on the FNC-1 test dataset and 200 matched tweet-fact pairs.

Table 7.11 shows Stance-BERT’s performance on the FNC-1 competition test dataset and our tweet-fact pairs, using F1 scores for all 3 classes as well as macro-F1. On FNC-1, we tested 2 variations of the same model: Stance-BERT_{window}, which uses the sliding-window approach (Section 7.7.4), and Stance-BERT_{trunc}, a model that truncates/discards all inputs after M tokens but is otherwise identical to Stance-BERT_{window}. Both variants significantly outperformed the method used in (Xu et al., 2018), one of the recent competitive methods on FNC-1. Because the sliding-window approach avoids discarding information from long inputs, it outperforms the truncation method.

Note that FNC-1 also includes a fourth “unrelated” class that we discarded, since we already have a claim-matching component. Because other recent stance detection methods (Mohtarami et al., 2018; Fang et al., 2019) only reported macro-F1 scores calculated using all four classes, including “unrelated,” we cannot report a direct comparison with their methods. However, we argue that our macro-F1 of 0.65 remains highly competitive. The model of (Xu et al., 2018) achieved a 0.98 F1 score on “unrelated,” which suggests that “unrelated” (i.e., separating related and unrelated

pairs) is far easier than the other 3 classes (i.e., discerning between different classes of related pairs). Given that Stance-BERT significantly outperformed (Xu et al., 2018) on all three other classes, it is plausible that Stance-BERT will remain a top performer under all four classes.

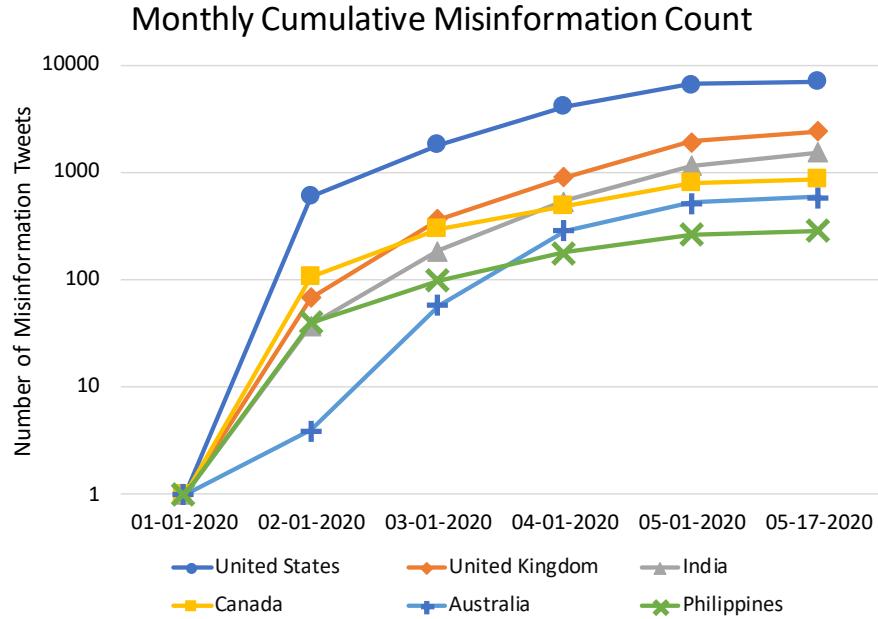


Figure 7.13: Six countries with the most misinformation tweets.

To evaluate Stance-BERT’s performance on our tweet-fact pairs, the three human annotators produced gold-standard labels on another set of 481 randomly selected tweet-fact pairs. 200 pairs are labeled as “matched.” These 200 pairs are further labeled as “agree”/“discuss”/“disagree,” in a distribution of 110/73/17 tweet-fact pairs. Ultimately, we discovered that Stance-BERT performs remarkably well on “agree” and “disagree” classes but falters on “discuss.” We hypothesize that this is caused by the significant class imbalance in FNC’s training set. Thus, we claim that the macro-F1 scores of 4-class evaluations are inflated w.r.t our 3-class results via

the inclusion of the unrelated class. With regard to inference speed, Stance-BERT processes 15 claim-evidence pairs per second on FNC-1 using 1 Nvidia GeForce GTX 1080 Ti GPU.

Misinformation Analysis.

Figure 7.13 is the cumulative timeline for the top-6 countries with the most COVID-19 misinformation tweets in the dataset. “Misinformation tweets” refer to tweets that go against known facts as judged by our stance detection model.

We also conducted a study on the correlation between misinformation tweet counts and COVID-19 case counts. We looked at the percentage of cases relative to a country’s population size, and the percentage of misinformation tweets relative to the total number of tweets from a country. The Pearson correlation coefficients between them are in Table 7.12. We find that the number of misinformation tweets most positively correlates with the number of confirmed cases. In contrast, its correlation with the number of recovered cases is weaker.

Country	Confirm	Death	Recover
United States	0.763	0.738	0.712
United Kingdom	0.862	0.833	-
India	0.794	0.798	0.755
Canada	0.706	0.667	0.663
Australia	0.954	0.922	0.887
Philippines	0.720	0.696	0.618

Table 7.12: Correlation between the percentage of confirmed/deceased/recovered cases and the percentage of misinformation tweets. The number of recovered cases in U.K. after April 13th is missing from the data source.

Finally, we manually categorized the misinformation tweets based on the taxonomy (Section 7.7.3). Table 7.13 lists the five most frequent categories of misinformation tweets: *Definition* (discussions about the nature of COVID-19), *Spreading*

Category	Count	Percentage
Definition	2503	15.1
Spreading	2118	12.7
Other	1450	8.7
Testing	1301	7.8
Disease Alongside	936	5.6
Total	8308	49.9

Table 7.13: Most frequent categories of misinformation tweets.

(descriptions of virus transmission), *Other* (tweets not fitting into predefined categories), *Testing* (tweets about methods, reliability, or aspects of COVID-19 testing), and *Disease Alongside* (mentions of diseases caused by or similar to COVID-19). We observe that the *Definition* category is the misinformation tweet category that is most frequently spread. These five categories make up 49.9% of all misinformation tweets, with the other 50.1% being spread out over the other 33 categories. The analysis of correlation and category is especially important for analyzing Twitter trends and enabling researchers to predict potential case surges in the future.

CHAPTER 8

CONCLUSIONS AND FUTURE WORK

This dissertation presented a comprehensive study on understanding and combating misinformation on social media through the lens of **truthfulness stance**—the stance a social media post takes regarding the truth or falsehood of a factual claim. We began by establishing a conceptual framework defining truthfulness stance as whether an utterance believes a claim to be true, false, or expresses neutral/no-stance about the claim’s veracity. Building on this foundation, a key contribution of this work is the creation of the TSD-CT dataset. This dataset is the first of its kind to capture social media stances on the truthfulness of general factual claims. Building on this dataset, we proposed the RATSD framework to automatically determine a social media post’s stance toward a given claim. RATSD is a novel stance detection architecture that integrates LLMs with retrieval-based context augmentation. Our experiments demonstrated that RATSD significantly outperforms prior state-of-the-art stance detection models. In summary, the TSD-CT dataset and the RATSD framework together advance the state of the art in truthfulness stance detection and lay important groundwork for future misinformation analysis research. In addition to methodological advances, this dissertation demonstrated the practical utility of truthfulness stance detection through three applications. TrustMap visualizes stance distributions toward factual claims across U.S. regions, enabling exploration of geographic patterns in belief and misinformation. The COVID-19 misinformation dashboard monitors stance toward verified facts and debunkings related to COVID-19, revealing regional and temporal trends to support public health responses. The ClimateStance applies the framework

to climate-related claims, identifying widespread over-acceptance of misinformation and limited discernment between truth and falsehood. Together, these applications show how stance detection can translate into actionable insights for monitoring public opinion, tracking misinformation dynamics, and informing targeted interventions.

8.1 Societal Impact

The research in this dissertation has broad implications for combating misinformation and strengthening an informed digital society. Truthfulness stance detection enables automated identification of whether social media content endorses or rejects factual claims to support timely and targeted countermeasures. Fact-checkers and content moderators can use stance outputs to prioritize emerging false narratives, while public health officials, policymakers, and social scientists can pinpoint communities most vulnerable to specific misinformation, as demonstrated in the COVID-19 and climate change case studies.

Beyond immediate interventions, stance detection offers a “map” of the misinformation landscape—revealing geographic and topical hotspots where false beliefs persist. This insight can guide educational campaigns, policy responses, and communication strategies tailored to specific audiences. Public-facing visualizations, such as TrustMap, could further promote digital literacy by making patterns of belief and disbelief transparent, encouraging individuals to critically evaluate widely held but inaccurate claims.

As automated systems mature, they can perform large-scale monitoring, freeing human experts to focus on the most consequential or complex falsehoods and fostering collaboration among technology providers, journalists, educators, and policymakers. Ethical deployment is essential: stance labels should be transparent, explainable, and privacy-conscious, with active management of potential biases. Done responsibly,

truthfulness stance detection can be a cornerstone of a healthier information ecosystem—one in which misinformation is rapidly identified and addressed, and the public is empowered to navigate information critically.

8.2 Future Work

While this dissertation advances truthfulness stance detection and its applications, several promising research directions remain.

Enhancing Stance Detection Models. Future work should aim to improve the accuracy, robustness, and interpretability of stance detection. Building on RATSD’s retrieval-augmented LLM approach, models could incorporate more sophisticated retrieval methods, domain-specific LMs, and real-time knowledge updates from news or fact-checking databases to address newly emerging claims. Handling nuanced content (such as sarcasm, humor, memes, or highly context-dependent language) may require integrating pragmatic cues such as emoji semantics, multimodal features, or conversational history. For real-world deployment, reliability and fairness are essential; adversarial testing and bias assessments should ensure consistent performance across diverse topics, user groups, and linguistic styles. In addition, providing clear, human-interpretable explanations for stance predictions remains an open challenge. Building on our use of LLM-generated analyses, future systems could produce concise rationales or highlight the evidence that informed a stance classification, increasing transparency, user trust, and decision-making value.

Expansion to Multilingual and Multimodal Data. Another natural extension is to broaden the scope of truthfulness stance detection beyond English text-based social media. Misinformation is a global phenomenon, spreading in many languages and through diverse media formats. An important future direction is to create **mul-**

tilingual stance detection systems that can handle content in different languages, or even mixed-code posts. This might involve collecting and annotating truthfulness stance data in other major languages, or employing cross-lingual transfer learning so that models trained on English (such as TSD-CT) can generalize to low-resource languages using machine translation or multilingual embeddings. In parallel, stance detection should be extended to **multimodal** content. Many viral misinformation narratives are conveyed not just through text, but via images, videos, and memes. Future research could explore detecting the stance of an image meme or a short video towards a claim (for example, a political cartoon implying a claim is false). This will require multimodal models that combine computer vision and NLP techniques. Some initial steps might include pairing our stance detection with image verification systems or analyzing the text in images along with captions. By expanding to multilingual and multimodal domains, truthfulness stance detection could provide a comprehensive coverage of misinformation in the wild, keeping pace with the evolving ways in which deceptive content is disseminated.

Real-Time Misinformation Tracking. Our current systems, such as the COVID-19 dashboard and TrustMap, analyze misinformation and stances in near-real-time, but there is room to make these capabilities more fully real-time and autonomous. One future work direction is to develop streaming stance detection algorithms that continuously ingest social media data (e.g., a Twitter stream or other platforms' feeds) and instantly tag posts with stance labels relative to a set of emerging factual claims. Realizing this requires not only efficient models but also an infrastructure for identifying and updating the set of claims of interest on the fly. Future research could leverage streaming topic detection or rumor detection methods to first discover trending claims or questions, and then apply stance detection in real-time to map how

those claims are being accepted or refuted by the public. The ultimate goal would be a system that functions as an early-warning radar for misinformation outbreaks: as soon as a dubious claim starts trending, the system would signal how people are reacting to it (believing or not) and where. Achieving real-time performance will involve addressing scalability (processing large volumes of data with low latency) and reliability (maintaining accuracy with minimal human oversight). Nonetheless, the benefit of such systems would be substantial – they would allow countermeasures (such as official corrections, platform interventions, or public awareness messages) to be deployed at the earliest stages of a misinformation spread, potentially preventing falsehoods from going viral.

Integration with Fact-Checking Ecosystems. Finally, a crucial direction for future work is to integrate truthfulness stance detection more deeply into the broader fact-checking and misinformation mitigation ecosystem. Currently, our models and applications operate mostly as stand-alone tools. In the future, they could be combined with professional fact-checking workflows and community-driven verification efforts. One idea is to use stance detection to automatically route information to fact-checkers: for example, if our system identifies a claim that is gaining traction and is widely believed to be true by the public but has no verified fact-check yet, it could alert fact-checking organizations to investigate that claim. Conversely, given a newly published fact-check article, an integrated system could immediately retrieve and label social media posts that discuss the claim in question, showing how the fact-check’s message is (or isn’t) percolating through the public sphere. This kind of symbiosis would greatly speed up the loop from debunking a false claim to ensuring that the debunking reaches the affected communities. Another integration point is with social media platforms themselves: stance detection could enhance platform

content moderation or context-providing features (imagine a feature that, beneath a post about a controversial claim, displays a summary such as “X% of related posts express doubt about this claim’s truthfulness”). This would not declare something true or false outright (thus avoiding censorship), but it gives readers a sense of the collective stance and may nudge them to verify information. Additionally, our approach of using external evidence through retrieval aligns well with automated fact-checking research; future systems might merge stance detection with claim verification models to both flag a post’s stance and provide evidence-backed verdicts on the claim’s accuracy. By blending into the fact-checking ecosystem, truthfulness stance detection can act as a force multiplier for human fact-checkers and a scalable mechanism to inoculate the public against misinformation.

In conclusion, the work presented in this dissertation opens several paths forward. Improving core stance detection technology, extending its reach across languages and media, operating in real-time, and teaming up with fact-checking efforts will all be critical to fully realize the promise of truthfulness stance detection. The continuing evolution of social media and misinformation poses new challenges, but it also highlights the importance of the foundations laid here. By building on these foundations, future researchers and practitioners can develop smarter, faster, and more inclusive tools to understand and curb misinformation. Ultimately, such advancements will contribute to a healthier information environment — one where citizens are better informed, misinformation is promptly checked, and the truth has a louder and more persuasive voice in the social media landscape.

APPENDIX A
IMPLEMENTATION DETAILS

A.1 Examples Included on the Instruction Page

Below are the examples presented on the instruction page:

The tweet believes the factual claim is true.

Example 1

Factual claim: ‘The largest bust in U.S. history’ 412 Muslims arrested from Michigan!

Tweet: 412 Michigan Muslims Arrested In Fed’s ‘LARGEST BUST IN U.S. HISTORY’ After Uncovering Deadly Hidden Secret <https://leadpatriot.com/412-michigan-arrested-in-feds-largest-bust-in-u-s-history-after-uncovering-deadly-hidden-secret/206/>

Explanation: The tweet rephrases the claim and capitalizes various words. It also contains a link to a webpage that provides further details on the claim. Based on these we can deduce that the tweet believes the factual claim is true.

Example 2

Factual claim: Ohio student suspended for staying in class during National Walkout Day.

Tweet: Patriot attorney needed. Reach out to this family. Ohio student suspended for staying in class during National Walkout Day [https://truepundit.com/ohio-student-suspended-staying-class-national-walkout-day/...](https://truepundit.com/ohio-student-suspended-staying-class-national-walkout-day/)

Explanation: The tweet calls for supporting the student who was allegedly suspended based on the claim. This suggests the tweet believes the claim is true.

Example 3

Factual claim: KAMALA HARRIS Says Schools in Berkeley Weren’t Integrated When She Was a Kid — But Yearbook Pictures Prove She’s Lying.

Tweet: It is like these morons don't know how the Internet works. We can look up your lies! Unreal... KAMALA HARRIS Says Schools in Berkeley Weren't Integrated When She Was a Kid – But Yearbook Pictures Prove She's Lying.

Explanation: The tweet believes Kamala Harris lied and that her alleged lie could be easily refuted by looking up online. For that reason, the tweet also calls Harris a "moron". All these signals indicate that the tweet believes the claim is true.

The tweet expresses a neutral or no stance toward the factual claim's truthfulness.

Example 1

Factual claim: NASA confirms Earth will experience 15 days of darkness In November 2017.

Tweet: How true is this..NASA Confirms Earth Will Experience 15 Days Of Complete Darkness in November 2015 [https://newswatch33.com/science/n](https://newswatch33.com/science/nasa-confirms-earth-will-experience-15-days-of-complete-darkness-in-november-2015/)

Explanation: The URL in the tweet suggests that the linked page discusses the same claim. Further reading the page content verifies that. The tweet inquires about the claim's veracity and it appears to be a genuine inquiry. Hence, it seems the tweet is not certain about the claim's truthfulness. We bundle "neutral stance" and "no stance" together as one answer option. Still, they are not the same. This example is a case of neutral stance (rather than no stance) since the tweet is directly asking about the claim's truthfulness.

Example 2

Factual claim: The media distorted what happened with a baby at his rally.

Tweet: Trump says 'dishonest' media distorted his 'baby joke': He referred to

the event at a rally in Des Moines, Io... <http://bit.ly/2aVwNUx>

Explanation: The tweet only stated that former U.S. President Donald Trump made the claim. It expresses no stance regarding the claim's veracity.

Example 3

Factual claim: Paul Ryan has blocked all action to strengthen our gun laws.

Tweet: Wisconsin students are marching 50 miles to Paul Ryan's hometown for action on gun laws <https://buff.ly/2up74Pn>

Explanation: The tweet and the claim are topic related since they both mention Paul Ryan's actions (or lack thereof) on gun laws. However, the tweet does not express any stance on whether Paul Ryan has blocked such actions. It only says the students are pressing him for actions.

Example 4

Factual claim: Is the Red Cross 'Not Helping California Wildfire Victims'?

Tweet: Don't forget tomorrow is HAT DAY! Your \$2 donation to the Red Cross will help California wildfire victims!

Explanation: The tweet and the claim are related because they both concern the topic of the Red Cross's aid to California wildfire victims. The tweet advocates for people to donate so that the Red Cross can better help the victims. Therefore, if the tweet user is to express their stance toward the claim's truthfulness, they will say they believe the Red Cross was helping. However, the tweet itself does not express any stance toward the claim. That's why this is an example of "no stance". Note that the fact-check article presents the claim in the form of a question. The background is that there was a claim spreading online that said the Red Cross was not helping the wildfire victims. Regardless of whether the

claim said the Red Cross was helping or not, the tweet expresses no stance toward the claim, although the tweet user would have clear stance if they express the stance.

The tweet believes the factual claim is false.

Example 1

Factual claim: We got rid of the Johnson Amendment.

Tweet: Donald Trump: “We got rid of the Johnson Amendment.” Rated Four Pinocchios by Washington Post – Via FactStream #BillionDollarLoser <https://factstream.co/factcheck/20452>

Explanation: The Washington Post uses “Pinocchios” as a measure of veracity of claims in their fact-checks. Four Pinocchios is their way of saying something is outright false. The tweet refers to WaPo’s “four Pinocchios” rating of this claim from the former U.S. President Donald Trump. It shows the tweet believes the claim is false.

Example 2

Factual claim: Rick Scott won and he won by a lot.

Tweet: Rick Scott is in office today because he won his election in that large state, Florida, by a mere 10,033 votes. He’s in pretty far over his skis, ego-wise, as a member of the Just Lucky To Be Here Caucus.

Explanation: The tweet states that Scott only won “by a mere 10,033 votes” and also considers him “just lucky”. This contradicts with the claim that he “won by a lot”. We can conclude that the tweet did not believe the factual claim was true.

Example 3

Factual claim: Jennifer Lawrence links 9/11 to Trump's election.

Tweet: A New Conspiracy Theory Falsely Claims Jennifer Lawrence Blamed Trump for 9/11. A meme with over 11,000 shares quotes Lawrence as blaming Trump for September 11 because he stole the election, but it's totally bogus.

Explanation: The tweet mentioned "conspiracy theory", "falsely claims" and "totally bogus". It is apparent the tweet believes the claim is false.

The tweet and the claim discuss different topics.

Example 1

Factual claim: New Jersey "will be out of gas for a week."

Tweet: I will not be at my High School reunion this weekend in New Jersey. We ran into too many commitments. Have a nice time Scotch Plains- Fanwood

Explanation: The tweet and the claim discuss apparently unrelated topics.

Example 2

Factual claim: Joe Biden and Democrats "have not legitimately won" the presidency.

Tweet: Biden fumes over Cuomo's DNC speech, book claims "Every four years, Democrats asked themselves the same question about the NY Gov: 'How is Andrew Cuomo going to f- -k us this time?' write the authors of 'Lucky: How Joe Biden Barely Won the Presidency.' <https://nypost.com/2021/03/02/book-says-biden-camp-fumed-over-cuomos-dnc-speech/>

Explanation: The tweet and the claim address different topics. The tweet focuses on Biden's and Democrats' attitude or concern regarding Cuomo, but the claim focuses on the legitimacy of Joe Biden's presidential win.

Example 3

Factual claim: There is a “0.05% chance of dying from COVID.”

Tweet: “My mom is not the only person COVID has killed this week,” said Thompson. “She is not a statistic. She is a human being. She was supposed to make a comeback, but she never got a chance.”

Explanation: The claim and the tweet are related in that they both discuss COVID deaths. However, they largely discuss different topics. The claim is about death rate, while the tweet is a reflection on family loss to the virus.

The tweet is created for sarcasm or parody only, or the tweet is problematic (e.g., hyperlink leading to “page not found” error, page content of the hyperlink behind paywall).

Example 1

Factual claim: Bill Gates talked about using vaccines to control population growth in an unedited 2010 TED Talk video.

Tweet: Thats his middle name all the Way to Hell! Idiot. Its been said he has been saying this. <https://share.newsbreak.com/1u38cjfg>

Explanation: The tweet’s hyperlink leads to a 404 (page not found) error.

Example 2

Factual claim: Nick Freitas supports a plan letting insurance companies deny coverage for preexisting conditions like asthma or diabetes.

Tweet: Checking the facts on Virginia Del. and Congressional Candidate Nick Freitas who “supports a plan letting insurance companies deny coverage for pre-existing conditions like asthma or diabetes.” https://starexponent.com/news/checking-the-facts-on-freitas-and-health-care/article_67f429c1-8

3a9-517f-ad5d-6d66e3671775.html?utm_medium=social&utm_source=twitter&utm_campaign=user-share via @culpeperse *Explanation:* The page content at the hyperlink is behind paywall.

Example 3

Factual claim: If we stopped testing right now, we'd have very few cases, if any.

Tweet: "If we stopped testing right now, we'd have very few cases." - Donald Trump He's right! LOL

Explanation: The tweet is created for sarcasm.

A.2 Explanations for the Challenging Cases

Below are detailed explanations for the six challenging truthfulness stance detection examples shown in Figure 2.2:

(1) This tweet implicitly affirms the claim without directly restating it. By calling for a “patriot attorney” and encouraging support for the student, the author signals agreement with the assertion that the suspension was unjust. The stance is labeled as \oplus , though it requires interpretation of sentiment and implied endorsement.

(2) The tweet merely asks whether Paul Ryan blocked all action to strengthen gun laws, without asserting a belief or providing evidence. Since it conveys uncertainty without supporting or opposing the claim, the stance is labeled **neutral**.

(3) Although the tweet is topically relevant—mentioning Paul Ryan, gun laws, and “action”—it does not convey a judgment on the truthfulness of the claim. It simply states a related event, and thus is best labeled as **no stance**.

(4) This tweet provides a nuanced rebuttal, noting that vaccinated individuals can still transmit the virus, albeit less frequently. It directly contradicts the claim that vaccinated people “cannot spread it to you,” and is therefore labeled **negative**.

However, recognizing this contradiction requires contextual understanding of public health discourse.

(5) The content shifts away from evaluating the claim itself and instead uses it as part of a sarcastic commentary. It neither confirms nor denies the claim, and instead engages with broader misinformation themes. The stance is labeled as *diff* due to this divergence.

(6) While the tweet quotes the claim verbatim, the response “He’s right” introduces ambiguity. The laughing emoji may imply sarcasm or mockery, making the author’s stance unclear. Because of this interpretive difficulty, the stance is labeled **problematic**.

A.3 Data Schema of the Collected Fact-Checks

For the fact-check collection, we adopted the data schema from ClaimReview. The detailed description of each field is as follows:

- **Publisher:** The organization or individual responsible for publishing the fact-check.
- **Claim Reviewed:** The specific factual claim that is being evaluated.
- **Review Summary:** A brief summary of the entire fact-check article. For instance, the “IF YOUR TIME IS SHORT” section from Politifact.
- **Review:** The detailed content of the fact-check article analyzing the veracity of the claim.
- **Verdict:** A label representing the assessed veracity of the claim, such as “Mostly True,” “False,” or “Pants on Fire.”
- **Author:** The individual who made the original factual claim.
- **Claim Published Date:** The date when the factual claim was originally made.

- **Factcheck Published Date:** The date when the fact-check was published by the publisher.
- **Thumbnail URL:** The URL to an image associated with the fact-check article.
- **URL:** The direct URL of the fact-check article.
- **Tags:** Topics or categories related to the factual claim, as assigned by the fact-checkers.

A.4 Prompt and Example for Contextual Knowledge Generation

For RATSD presented in Chapter 5, Figure A.1 shows the prompt used to generate contextual knowledge for each claim–tweet pair. Given a claim and a tweet, the prompt instructs the LLM to generate two summaries: one capturing factual background about the claim (Figure A.2) and another providing context about the tweet (Figure A.3). These summaries are constructed using retrieved knowledge chunks and are critical for improving stance classification by grounding the model in relevant, verifiable information.

Consider this factual claim: {California introduces new bill that would allow mothers to kill their babies up to 7 days after birth}. The claimant is {Facebook Posts}. Please provide the details about the claimant, enumerate entities along with their information, and provide related information regarding those entities.

Consider this tweet: {California introduces new bill that would allow mothers to kill their babies up to 7 days after birth. A newly proposed bill would legalize the murder of children up to nine months gestation and in the week(s) after birth} . The tweet user profile is {The last non-communist in California. Bible-believing Christian. Editor: The Liberty Daily and America First Report}. Please enumerate entities along with their information and provide related information regarding those entities. Please generate the context knowledge for the factual claim and tweet separately in a structured way.

Figure A.1: The prompt used to generate contextual knowledge for each claim–tweet pair.

A.5 Keywords for Collecting Climate Change-Related Factual Claims

In Chapter 7, we curated a list of keywords related to climate change from the glossary of the Environmental Protection Agency (EPA) to collect factual claims from

Factual Claim: "California introduces new bill that would allow mothers to kill their babies up to 7 days after birth."

Claimant: Facebook Posts

Source: A viral Facebook post dated March 31, 2022.

Description: The post contained a headline that falsely claimed that California lawmakers introduced a bill allowing mothers to kill newborns up to seven days after birth. The post was flagged as part of Facebook's efforts to combat misinformation.

Key Entities: California Assembly Bill 2223 (AB 2223)

Description: Real bill introduced by Assembly Member Buffy Wicks. It does not allow for killing newborns but seeks to protect individuals from criminal prosecution or civil liability for pregnancy-related outcomes, such as stillbirth, miscarriage, or abortion. It includes language to clarify that deaths related to pregnancy should not be criminalized.

Purpose: To prevent parents from being prosecuted or investigated for pregnancy loss or newborn deaths due to pregnancy-related complications.

Affiliation: Democrat, California Assembly Member, representing Oakland.

Role: Author of AB 2223, Wicks emphasizes the bill's intent to protect reproductive rights and prevent wrongful prosecutions linked to pregnancy loss or newborn death due to pregnancy-related causes.

Action: The claim was flagged by Facebook's team as part of its broader effort to prevent the spread of false news. The post was identified as misleading.

Description: Two California women who were prosecuted after delivering stillborn babies. Their cases highlighted the legal issues surrounding pregnancy loss, which Wicks' bill aims to address.

Additional Information: The false claim stemmed from a misunderstanding or deliberate misrepresentation of the term "perinatal" in the bill, which refers to a period following pregnancy. The bill would not protect individuals who intentionally harm newborns.

PolitiFact Rating: The claim that the bill would allow for the killing of babies is categorically false. The bill aims to prevent investigations or prosecutions for pregnancy-related outcomes, not to legalize murder.

Figure A.2: The contextual knowledge about the claim.

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.6 Fact check verdicts and their categories

For ClimateStance illustrated in Chapter 7, the factual claims' verdicts below are categorized into "Truth," "Uncertain," "Misinformation."

- "Truth": True, Correct Attribution, No-Flip, Mostly True, Likely, Near certain.
- "Uncertain": Uncertain, Half True, Research In Progress, Mixture, Unknown, Half-flip, Missing context.

Tweet: "California introduces new bill that would allow mothers to kill their babies up to 7 days after birth. A newly proposed bill would legalize the murder of children up to nine months gestation and in the week(s) after birth."

Tweet User Profile: "The last non-communist in California. Bible-believing Christian. Editor: The Liberty Daily and America First Report."

Affiliation: Describes themselves as a conservative, Christian editor for The Liberty Daily and America First Report, likely reflecting a far-right political stance.

Focus: The profile suggests a conservative, possibly religious-based opposition to legislation like AB 2223, aligning with anti-abortion or anti-reproductive-rights views.

Key Entities: The Liberty Daily

Description: A conservative online news outlet known for its far-right perspectives on political and social issues. The platform is often critical of progressive policies, particularly those related to reproductive rights.

Content Focus: Frequently publishes articles that critique liberal or progressive initiatives, including those related to abortion or reproductive health laws.

Description: Another far-right media outlet with a focus on populist and nationalist views, advocating for America-centric policies and conservative social values.

Editor: The user of the tweet serves as the editor, indicating their significant role in shaping the outlet's editorial stance on various issues.

Additional Information: The tweet amplifies a false claim by conflating the legal language of the bill with an inflammatory accusation about legalizing infanticide. The bill (AB 2223) does not permit infanticide; rather, it protects individuals from prosecution in cases of pregnancy-related deaths, clarifying that prosecution should not be pursued in instances like miscarriage or stillbirth due to pregnancy complications.

Figure A.3: The contextual knowledge about the tweet.

- “Misinformation”: False, Pants on Fire, Fake, Full Flop, Labeled Satire, Mostly False, Barely True, False, Unlikely, Extremely Unlikely, Miscaptioned.

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 A.1: 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 A.2: Examples of relatively accurate medium topics in the public’s judgments.



Figure A.4: Prompt used to generate topics for each claim.

A.7 Prompt for Topic Generation

In Chapter 7, the prompt for topic generation includes 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 A.4. 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 applied to all the factual claims to generate topics for them.

A.8 Truthfulness Stance Distribution across Medium Topics

Tables A.1 and A.2 show examples of medium topics where the public's judgments of truth and misinformation are relatively inaccurate and accurate, respectively. In Table A.2, 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. This indicates a relatively balanced and correct perception of truth and misinformation within these topics. In contrast, Table A.1 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*.” This suggests a need for better information dissemination and public education in these areas to improve judgment accuracy.

Bibliography

- J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- A. Agarwal, R. Singh, and D. Toshniwal. Geospatial sentiment analysis using Twitter data for UK-EU referendum. *Journal of Information and Optimization Sciences*, 39(1):303–317, 2018.
- J. Al Qundus, A. Paschke, S. Gupta, A. M. Alzouby, and M. Yousef. Exploring the impact of short-text complexity and structure on its quality in social media. *Journal of Enterprise Information Management*, 33(6):1443–1466, 2020.
- M. S. Al-Rakhami and A. M. Al-Amri. Lies kill, facts save: Detecting COVID-19 misinformation in Twitter. *IEEE Access*, 8:155961–155970, 2020.
- A. Aleksandric, H. I. Anderson, A. Dangal, G. M. Wilson, and S. Nilizadeh. 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*, volume 18, pages 15–28, 2024.
- E. Allaway and K. R. McKeown. Zero-shot stance detection: A dataset and model using generalized topic representations. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 8913–8931, 2020.
- H. Allcott, M. Gentzkow, and C. Yu. Trends in the diffusion of misinformation on social media. *Research and politics*, 6(2):2053168019848554, 2019.

- N. S. Alturayeif, H. Luqman, and M. A. Ahmed. A systematic review of machine learning techniques for stance detection and its applications. *Neural Comput. Appl.*, 35(7):5113–5144, 2023.
- E. Arakelyan, A. Arora, and I. Augenstein. Topic-guided sampling for data-efficient multi-domain stance detection. In A. Rogers, J. L. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 13448–13464, 2023.
- F. Arslan, N. Hassan, C. Li, and M. Tremayne. A benchmark dataset of check-worthy factual claims. In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media (ICWSM)*, pages 821–829, 2020.
- P. Atanasova. Generating fact checking explanations. In *Accountable and Explainable Methods for Complex Reasoning over Text*, pages 83–103. Springer, 2024.
- A. M. A. Ausat. The role of social media in shaping public opinion and its influence on economic decisions. *Technology and Society Perspectives (TACIT)*, 1(1):35–44, 2023.
- E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic. The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*, pages 519–528, 2012.
- J. M. Banda, R. Tekumalla, G. Wang, J. Yu, T. Liu, Y. Ding, and G. Chowell. A large-scale COVID-19 twitter chatter dataset for open scientific research - an international collaboration. *CoRR*, abs/2004.03688, 2020.
- R. Banner, Y. Nahshan, and D. Soudry. Post training 4-bit quantization of convolutional networks for rapid-deployment. In H. M. Wallach, H. Larochelle, A. Beygelz-

imer, F. d’Alché-Buc, E. B. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7948–7956, 2019.

I. Beltagy, K. Lo, and A. Cohan. SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

P. E. Black. Ratcliff/Obershelp pattern recognition, 2021.

E. Borra and B. Rieder. Programmed method: Developing a toolset for capturing and analyzing tweets. *Aslib journal of information management*, 66(3):262–278, 2014.

J. S. Brennen, F. M. Simon, P. N. Howard, and R. K. Nielsen. Types, sources, and claims of COVID-19 misinformation. *Reuters Institute*, 2020.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, H. Zhang, B. Zhu, M. Jordan, J. E. Gonzalez, and I. Stoica. Chatbot arena: An open platform for evaluating llms by human preference, 2024.

J. Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

C. Conforti, J. Berndt, M. T. Pilehvar, C. Giannitsarou, F. Toxvaerd, and N. Collier. Will-They-Won’t-They: A very large dataset for stance detection on Twitter. In D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, editors, *Proceedings of the*

58th Annual Meeting of the Association for Computational Linguistics, pages 1715–1724, 2020.

A. Das, H. Liu, V. Kovatchev, and M. Lease. The state of human-centered nlp technology for fact-checking. *Information processing & management*, 60(2):103219, 2023.

L. Derczynski, K. Bontcheva, M. Liakata, R. Procter, G. Wong Sak Hoi, and A. Zubaga. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, pages 69–76, Aug. 2017.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186, 2019.

M. Diab, L. Levin, T. Mitamura, O. Rambow, V. Prabhakaran, and W. Guo. Committed belief annotation and tagging. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 68–73, 2009.

D. V. Dimitrova and J. Matthes. Social media in political campaigning around the world: Theoretical and methodological challenges, 2018.

P. Dizikes. False news travels faster than true stories on Twitter.
<https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308>, Mar. 2018.

I. J. B. Do Nascimento, A. B. Pizarro, J. M. Almeida, N. Azzopardi-Muscat, M. A. Gonçalves, M. Björklund, and D. Novillo-Ortiz. Infodemics and health misinformation

tion: A systematic review of reviews. *Bulletin of the World Health Organization*, 100(9):544, 2022.

M. Dobbs, J. DeGutis, J. Morales, K. Joseph, and B. Swire-Thompson. Democrats are better than republicans at discerning true and false news but do not have better metacognitive awareness. *Communications Psychology*, 1(1):46, 2023.

F. Dretske. The stance stance. *Behavioral and Brain Sciences*, 11(3):511–512, 1988.

J. Du, R. Xu, Y. He, and L. Gui. Stance classification with target-specific neural attention networks. In *26th International Joint Conference on Artificial Intelligence, IJCAI 2017*, pages 3988–3994. International Joint Conferences on Artificial Intelligence, 2017.

Y. K. Dwivedi, E. Ismagilova, D. L. Hughes, J. Carlson, R. Filieri, J. Jacobson, V. Jain, H. Karjaluoto, H. Kefi, A. S. Krishen, et al. Setting the future of digital and social media marketing research: Perspectives and research propositions. *International journal of information management*, 59:102168, 2021.

U. K. Ecker, S. Lewandowsky, J. Cook, P. Schmid, L. K. Fazio, N. Brashier, P. Kendeou, E. K. Vraga, and M. A. Amazeen. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1):13–29, 2022.

M. Fajcik, P. Smrz, and L. Burget. BUT-FIT at SemEval-2019 task 7: Determining the rumour stance with pre-trained deep bidirectional transformers. In J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, and S. M. Mohammad, editors, *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1097–1104, 2019.

W. Fang, M. Nadeem, M. Mohtarami, and J. Glass. Neural multi-task learning for stance prediction. In *EMNLP Workshop on Fact Extraction and Verification*, pages 13–19, 2019.

W. Ferreira and A. Vlachos. Emergent: A novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1163–1168, June 2016.

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

M. H. Gad-Elrab, D. Stepanova, J. Urbani, and G. Weikum. ExFaKT: A framework for explaining facts over knowledge graphs and text. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 87–95, 2019.

R. K. Garrett and R. M. Bond. Conservatives’ susceptibility to political misperceptions. *Science Advances*, 7(23):eabf1234, 2021.

GeoPy. GeoPy: Geocoding library for python, 2023. Includes support for Nominatim (OpenStreetMap) geocoding services.

W. H. Gomaa, A. A. Fahmy, et al. A survey of text similarity approaches. *International Journal of Computer Applications*, 68(13):13–18, 2013.

G. Gorrell, E. Kochkina, M. Liakata, A. Aker, A. Zubiaga, K. Bontcheva, and L. Derczynski. 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, 2019.

D. Graves. Understanding the promise and limits of automated fact-checking. *Reuters Institute for the Study of Journalism*, 2018.

- L. Graves. Anatomy of a fact check: Objective practice and the contested epistemology of fact checking. *Communication, culture & critique*, 10(3):518–537, 2017.
- L. Grimminger and R. Klinger. Hate towards the political opponent: A twitter corpus study of the 2020 US elections on the basis of offensive speech and stance detection. In O. D. Clercq, A. Balahur, J. Sedoc, V. Barrière, S. Tafreshi, S. Buechel, and V. Hoste, editors, *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 171–180, 2021.
- Z. Guo, M. Schlichtkrull, and A. Vlachos. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206, 2022.
- K. L. Gwet. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48, 2008.
- M. Hardalov, A. Arora, P. Nakov, and I. Augenstein. A survey on stance detection for mis- and disinformation identification. In M. Carpuat, M. de Marneffe, and I. V. M. Ruiz, editors, *Findings of the Association for Computational Linguistics*, pages 1259–1277, 2022a.
- M. Hardalov, A. Arora, P. Nakov, and I. Augenstein. A survey on stance detection for mis- and disinformation identification. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1259–1277, Seattle, United States, July 2022b.
- N. Hassan, C. Li, and M. Tremayne. Detecting check-worthy factual claims in presidential debates. In *Proceedings of the 24th ACM conference on Information and knowledge management (CIKM)*, pages 1835–1838, 2015.

- N. Hassan, F. Arslan, C. Li, and M. Tremayne. 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)*, pages 1803–1812, 2017.
- S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- T. Hossain, R. L. Logan IV, A. Ugarte, Y. Matsubara, S. Young, and S. Singh. COVIDLies: Detecting COVID-19 misinformation on social media. In *Proceedings of the 1st Workshop on NLP for COVID-19*, Dec. 2020.
- Y. Hou, P. van der Putten, and S. Verberne. The COVMis-stance dataset: Stance detection on Twitter for COVID-19 misinformation. *arXiv preprint arXiv:2204.02000*, 2022.
- Z. Ji, T. Yu, Y. Xu, N. Lee, E. Ishii, and P. Fung. Towards mitigating LLM hallucination via self reflection. In *Findings of the Association for Computational Linguistics*, pages 1827–1843, 2023.
- K. Joseph, L. Friedland, W. Hobbs, O. Tsur, and D. Lazer. ConStance: Modeling annotation contexts to improve stance classification. In *EMNLP*, 2017.
- D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations*, 2015.
- E. Kochkina, M. Liakata, and I. Augenstein. Turing at SemEval-2017 Task 8: Sequential approach to rumour stance classification with branch-LSTM. *arXiv preprint arXiv:1704.07221*, 2017.

- L. Konstantinovskiy, O. Price, M. Babakar, and A. Zubiaga. Toward automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection. *Digital threats: research and practice*, 2(2):1–16, 2021.
- M. Koppel and J. Schler. The importance of neutral examples for learning sentiment. *Computational Intelligence*, 22(2):100–109, 2006.
- Z. Kou, L. Shang, Y. Zhang, C. Youn, and D. Wang. FakeSens: A social sensing approach to COVID-19 misinformation detection on social media. In *2021 17th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, pages 140–147. IEEE, 2021.
- R. Kouzy, J. Abi Jaoude, A. Kraitem, M. B. El Alam, B. Karam, E. Adib, J. Zarka, C. Traboulsi, E. W. Akl, and K. Baddour. Coronavirus goes viral: Quantifying the COVID-19 misinformation epidemic on Twitter. *Cureus*, 12(3), 2020.
- K. Krippendorff. Computing krippendorff’s alpha-reliability, 2011.
- K. Kucher, R. M. Martins, C. Paradis, and A. Kerren. StanceVis Prime: Visual analysis of sentiment and stance in social media texts. *Journal of Visualization*, 23(6):1015–1034, 2020.
- D. Küçük and F. Can. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37, 2020.
- S. Lai, L. Xu, K. Liu, and J. Zhao. Recurrent convolutional neural networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

- D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, et al. The science of fake news. *Science*, 359(6380):1094–1096, 2018.
- S. Lewandowsky, U. K. Ecker, C. M. Seifert, N. Schwarz, and J. Cook. Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest*, 13(3):106–131, 2012.
- P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.
- Y. Li, T. Grandison, P. Silveyra, A. Douraghy, X. Guan, T. Kieselbach, C. Li, and H. Zhang. Jennifer for COVID-19: An nlp-powered chatbot built for the people and by the people to combat misinformation. In *ACL Workshop on Natural Language Processing for COVID-19*, pages 1–9, 2020.
- Y. Li, T. Sosea, A. Sawant, A. J. Nair, D. Inkpen, and C. Caragea. P-Stance: A large dataset for stance detection in political domain. In *Findings of the Association for Computational Linguistics*, pages 2355–2365, 2021.
- A. K. Lien, L. M. Randem, H. P. F. Taralrud, and M. Edalati. OSN dashboard tool for sentiment analysis. *arXiv preprint arXiv:2206.06935*, 2022.
- X. Y. Liew, N. Hameed, J. Clos, and J. E. Fischer. Designing and evaluating a discourse analysis dashboard. In *Proceedings of the Second International Symposium on Trustworthy Autonomous Systems*, pages 1–5, 2024.

- Z. C. Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.
- Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- S. Loomba, A. De Figueiredo, S. J. Piatek, K. De Graaf, and H. J. Larson. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature human behaviour*, 5(3):337–348, 2021.
- W. Mansour, T. Elsayed, and A. Al-Ali. This is not new! Spotting previously-verified claims over Twitter. *Information Processing & Management*, 60(4):103414, 2023.
- S. A. Memon and K. M. Carley. Characterizing COVID-19 misinformation communities using a novel Twitter dataset. *arXiv preprint arXiv:2008.00791*, 2020.
- A. Mian and S. Khan. Coronavirus: The spread of misinformation. *BMC medicine*, 18(1):1–2, 2020.
- S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, pages 31–41, 2016.
- M. Mohtarami, R. Baly, J. Glass, P. Nakov, L. Márquez, and A. Moschitti. Automatic stance detection using end-to-end memory networks. In *NAACL*, pages 767–776, 2018.
- P. L. Moravec, R. K. Minas, and A. R. Dennis. Fake news on social media: People believe what they want to believe when it makes no sense at all. *MIS Q.*, 43(4), 2019.

N. Muennighoff, N. Tazi, L. Magne, and N. Reimers. MTEB: Massive text embedding benchmark. In A. Vlachos and I. Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2006–2029, 2023.

M. Müller, M. Salathé, and P. E. Kummervold. COVID-Twitter-BERT: A natural language processing model to analyse COVID-19 content on Twitter. *Frontiers in artificial intelligence*, 6:1023281, 2023.

P. Nakov, A. Barrón-Cedeno, T. Elsayed, R. Suwaileh, L. Màrquez, W. Zaghouani, P. Atanasova, S. Kyuchukov, and G. Da San Martino. Overview of the CLEF-2018 CheckThat! Lab on automatic identification and verification of political claims. In *International conference of the cross-language evaluation forum for european languages*, pages 372–387. Springer, 2018.

J. Ni, M. Shi, D. Stammbach, M. Sachan, E. Ash, and M. Leippold. AFaCTA: Assisting the annotation of factual claim detection with reliable LLM annotators. *arXiv preprint arXiv:2402.11073*, 2024.

K. Ognyanova, D. Lazer, R. E. Robertson, and C. Wilson. 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*, 2020.

OpenAI. GPT-4o-mini: Optimized mini version of GPT-4, 2024.

S. O. Oyeyemi, E. Gabarron, and R. Wynn. Ebola, Twitter, and misinformation: a dangerous combination? *BMJ*, 349:g6178, 2014.

- S. Patel, P. Bansal, and P. Kaur. Rumour detection using graph neural network and oversampling in benchmark Twitter dataset. *arXiv preprint arXiv:2212.10080*, 2022.
- D. Pomerleau and D. Rao. Fake News Challenge Stage 1 (FNC-1): Stance Detection, 2017.
- K. Popat, S. Mukherjee, A. Yates, and G. Weikum. DeClarE: Debunking fake news and false claims using evidence-aware deep learning. *arXiv preprint arXiv:1809.06416*, 2018.
- T. Pranckevičius and V. Marcinkevičius. Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Baltic Journal of Modern Computing*, 5(2):221, 2017.
- B. Quinion, S. Hoffmann, and M. T. Metten. Nominatim: A search engine for open-streetmap data, 2020.
- R. G. Reddy, S. Chetan, Z. Wang, Y. R. Fung, K. Conger, A. Elsayed, M. Palmer, P. Nakov, E. Hovy, K. Small, et al. 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*, pages 6002–6018, 2022.
- N. Reimers and I. Gurevych. Sentence-BERT: Sentence embeddings using siamese bert-networks. In *EMNLP-IJCNLP*, pages 3973–3983, 2019.
- Reuters. Global audiences suspicious of ai-powered newsrooms, report finds. *Reuters News*, 2024. Highlights public concern about fake news and AI in sensitive domains.

- J. Roozenbeek, C. R. Schneider, S. Dryhurst, J. Kerr, A. L. Freeman, G. Recchia, A. M. Van Der Bles, and S. Van Der Linden. Susceptibility to misinformation about COVID-19 around the world. *Royal Society open science*, 7(10):201199, 2020.
- M. Saeed, N. Traub, M. Nicolas, G. Demartini, and P. Papotti. Crowdsourced fact-checking at Twitter: How does the crowd compare with experts? In *Proceedings of the 31st ACM international conference on information & knowledge management*, pages 1736–1746, 2022.
- R. Saurí and J. Pustejovsky. FactBank: A corpus annotated with event factuality. *Language resources and evaluation*, 43:227–268, 2009.
- W. A. Scott. Reliability of content analysis: The case of nominal scale coding. *Public opinion quarterly*, pages 321–325, 1955.
- G. K. Shahi and D. Nandini. FakeCovid: A multilingual cross-domain fact check news dataset for COVID-19. *arXiv preprint arXiv:2006.11343*, 2020.
- S. Shi, K. Qiao, J. Chen, S. Yang, J. Yang, B. Song, L. Wang, and B. Yan. MGTAB: A multi-relational graph-based Twitter account detection benchmark. *arXiv preprint arXiv:2301.01123*, 2023.
- Z. Shi, H. Rui, and A. B. Whinston. Content sharing in a social broadcasting environment: Evidence from Twitter. *MIS quarterly*, 38(1):123–142, 2014.
- M. Skeppstedt, V. Simaki, C. Paradis, and A. Kerren. Detection of stance and sentiment modifiers in political blogs. In *International conference on speech and computer*, pages 302–311, 2017.

- I. Smirnov. The digital flynn effect: Complexity of posts on social media increases over time. In *Social Informatics: 9th International Conference*, pages 24–30. Springer, 2017.
- P. Sobkowicz, M. Kaschesky, and G. Bouchard. Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web. *Government Information Quarterly*, 29(4):470–479, 2012.
- T. Spampatti, U. J. Hahnel, and T. Brosch. Conservatives are less accurate than liberals at recognizing false climate statements, and disinformation makes conservatives less discerning: Evidence from 12 countries. *Harvard Kennedy School Misinformation Review*, 2024.
- V. Suarez-Lledo and J. Alvarez-Galvez. Prevalence of health misinformation on social media: systematic review. *Journal of medical Internet research*, 23(1):e17187, 2021.
- Talker Research. Majority of americans trust what’s online less than ever before. *New York Post*, 2025.
- T. Tang, H. Lu, Y. E. Jiang, H. Huang, D. Zhang, W. X. Zhao, and F. Wei. Not all metrics are guilty: Improving nlg evaluation with LLM paraphrasing. *arXiv preprint arXiv:2305.15067*, 2023.
- G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, et al. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal. FEVER: a large-scale dataset for fact extraction and verification. In M. A. Walker, H. Ji, and

- A. Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 809–819, 2018.
- S. Tonmoy, S. Zaman, V. Jain, A. Rani, V. Rawte, A. Chadha, and A. Das. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*, 2024.
- H. Touvron, L. Martin, K. Lu, O. Ruwase, S. Bhosale, J. Jiang, A. Joulin, M. Ott, and Y. LeCun. Llama 3: Open foundation and instruction models, 2024.
- K. M. d. Treen, H. T. Williams, and S. J. O'Neill. Online misinformation about climate change. *Wiley Interdisciplinary Reviews: Climate Change*, 11(5):e665, 2020.
- J. A. Tucker, A. Guess, P. Barberá, C. Vaccari, A. Siegel, S. Sanovich, D. Stukal, and B. Nyhan. Social media, political polarization, and political disinformation: A review of the scientific literature. *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)*, 2018.
- L. Tunstall, E. Beeching, N. Lambert, N. Rajani, K. Rasul, Y. Belkada, S. Huang, L. von Werra, C. Fourrier, N. Habib, et al. Zephyr: Direct distillation of LM alignment. *arXiv preprint arXiv:2310.16944*, 2023.
- S. Vosoughi, D. Roy, and S. Aral. The spread of true and false news online. *science*, 359(6380):1146–1151, 2018.
- C. Wang, X. Liu, Y. Yue, X. Tang, T. Zhang, C. Jiayang, Y. Yao, W. Gao, X. Hu, Z. Qi, et al. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv preprint arXiv:2310.07521*, 2023.

J. Wang and Y. Dong. Measurement of text similarity: a survey. *Information*, 11(9):421, 2020.

W. Y. Wang. “Liar, Liar Pants on Fire”: A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*, 2017.

D. J. Watts, D. M. Rothschild, and M. Mobius. Measuring the news and its impact on democracy. *Proceedings of the National Academy of Sciences*, 118(15):e1912443118, 2021.

J. Wei, C. Huang, S. Vosoughi, and J. Wei. What are people asking about COVID-19? A question classification dataset. *arXiv preprint arXiv:2005.12522*, 2020.

M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016.

T. Willaert, P. Van Eecke, K. Beuls, and L. Steels. Building social media observatories for monitoring online opinion dynamics. *Social Media+Society*, 6(2):2056305119898778, 2020.

Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al. Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

S. Xiao, Z. Liu, P. Zhang, N. Muennighoff, D. Lian, and J.-Y. Nie. C-Pack: Packed resources for general Chinese embeddings. In *Proceedings of the 47th interna-*

tional ACM SIGIR conference on research and development in information retrieval, pages 641–649, 2024.

B. Xu, M. Mohtarami, and J. Glass. Adversarial domain adaptation for stance detection. In *NeurIPS*, 2018.

F. Xu, V. S. Sheng, and M. Wang. A unified perspective for disinformation detection and truth discovery in social sensing: a survey. *ACM Computing Surveys (CSUR)*, 55(1):1–33, 2021.

H. Y. Yan, G. Morrow, K.-C. Yang, and J. Wilbey. The origin of public concerns over ai supercharging misinformation in the 2024 us presidential election. *Harvard Kennedy School Misinformation Review*, 2025.

R. Yang, W. Xie, C. Liu, and D. Yu. 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*, pages 1090–1096, 2019a.

Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. XLNet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019b.

J.-Y. Yao, K.-P. Ning, Z.-H. Liu, M.-N. Ning, and L. Yuan. LLM lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469*, 2023.

R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi. Defending against neural fake news. *Advances in neural information processing systems*, 32, 2019.

- X. Zeng, A. S. Abumansour, and A. Zubiaga. Automated fact-checking: A survey. *Language and Linguistics Compass*, 15(10):e12438, 2021.
- D. Y. Zhang, J. Badilla, Y. Zhang, and D. Wang. 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, 2018.
- H. Zhang, Z. Zhu, Z. Zhang, J. Devasier, and C. Li. 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, 2024a.
- Z. Zhang, Z. Zhu, H. Zhang, and C. Li. Exploring behavioral tendencies on social media: A perspective through claim check-worthiness. In *International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 373–390. Springer, 2024b.
- Z. Zhang, Z. Zhu, H. Zhang, F. Patel, J. Caraballo, P. Hennecke, and C. Li. Wildfire: A Twitter social sensing platform for layperson. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM), demonstration description*, pages 1106–1109, 2024c. doi: 10.1145/3616855.3635704.
- X. Zhu and C. Pechmann. Political polarization triggers conservatives’ misinformation spread to attain ingroup dominance. *Journal of Marketing*, 89(1):39–55, 2025.
- Z. Zhu, K. Meng, J. Caraballo, I. Jaradat, X. Shi, Z. Zhang, F. Akrami, H. Liao, F. Arslan, D. Jimenez, M. S. Saeef, P. Pathak, and C. Li. A dashboard for mitigating the COVID-19 misinfodemic. In *Proceedings of the 16th Conference of*

the European Chapter of the Association for Computational Linguistics (EACL): System Demonstrations, pages 99–105, Apr. 2021.

Z. Zhu, Z. Zhang, F. Patel, and C. Li. Detecting stance of tweets toward truthfulness of factual claims. In *Proceedings of the 2022 Computation+Journalism Symposium*, 2022.

Z. Zhu, H. Zhang, Z. Zhang, and C. Li. 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*, Coex, Seoul, Korea, Nov. 2025a. Association for Computing Machinery.

Z. Zhu, Z. Zhang, H. Zhang, and C. Li. RATSD: Retrieval augmented truthfulness stance detection from social media posts toward factual claims. In *Findings of the Association for Computational Linguistics: NAACL 2025*, page 3366–3381, 2025b.

A. Zubiaga, M. Liakata, R. Procter, G. Wong Sak Hoi, and P. Tolmie. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 11(3):e0150989, 2016.

BIOGRAPHICAL STATEMENT

Zhengyuan Zhu received his Ph.D. in Computer Science and Engineering from the University of Texas at Arlington in 2025. He was supervised by Dr. Chengkai Li. His research spans natural language processing, knowledge graphs, and large language models for social good, with a focus on building large-scale systems for truthfulness stance detection, multimodal question-answering, and text generation from knowledge graphs. Zhengyuan's work has been published in premier conferences, including NAACL, EMNLP, ACL, EACL, COLING, WSDM, and ASONAM. His contributions have been recognized through invited talks at the Texas Advanced Computing Center and teaching roles in Google-funded workshops. His publication, based on work conducted at Adobe, received the Best Demo Paper Award at COLING 2025.