RATSD: Retrieval Augmented Truthfulness Stance Detection from Social Media Posts Toward Factual Claims

Anonymous ACL submission

Abstract

The rapid spread of misinformation on social media presents challenges for gauging the truthfulness of public discourse. This paper introduces the concept of truthfulness stance, which assesses whether a textual utterance believes a factual claim to be true, false, or expresses a neutral stance or no stance toward the claim. Our systematic analysis fills a gap in the existing literature by offering the first in-depth exploration of truthfulness stance. We propose a new framework, RATSD (Retrieval Augmented Truthfulness Stance Detection), which utilizes large language models (LLMs) with retrieved-augmented generation to enhance the contextual understanding of tweets in relation to claims. RATSD is evaluated on the newly developed TSD-CT dataset, which contains 3,105 claim-tweet pairs, alongside benchmark datasets. Our experiments show that RATSD outperforms state-of-the-art methods, achieving a significant increase in Macro-F1 score. Our contributions lay the groundwork for advancing research in misinformation analysis and provide tools for understanding public perceptions.

1 Introduction

Online information provides a valuable lens through which we can gauge 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; Willaert et al., 2020). This paper introduces the novel concept of *truthfulness stance*—given a factual claim, whether a textual utterance believes the claim is true or false, or expresses a neutral or indeterminate stance toward its truthfulness. The paper focuses on social media posts (specifically, tweets on Twitter, now called X) as the particular type of textual utterances. Figure 1 shows a few examples in which the tweets express positive, neutral, negative, and no stance, respectively, regarding the truthfulness of the same factual claim.

Figure 1: Four tweets expressing different truthfulness stances toward the same factual claim.

Truthfulness stance can be a useful tool in discerning how misinformation spreads (Ecker et al., 2022) and influences people's decision-making in political discourse (Ognyanova et al., 2020) and health (Suarez-Lledo and Alvarez-Galvez, 2021). Such information can help social scientists measure the impact of misinformation and design effective interventions; health organizations can use the information to identify communities that may be more susceptible to health-related falsehoods. Truthfulness stance can also be valuable to marketers and media strategists for assessing the effectiveness of their campaigns (Dwivedi et al., 2021), as well as for tracking changing beliefs about a product or a political figure's popularity among social media users (Dimitrova and Matthes, 2018).

Definitions of stance in various studies share a common conceptual framework through which a declared stance comprises four components: a textual *utterance* that expresses the stance (e.g., a news article or a social media post), a *target*

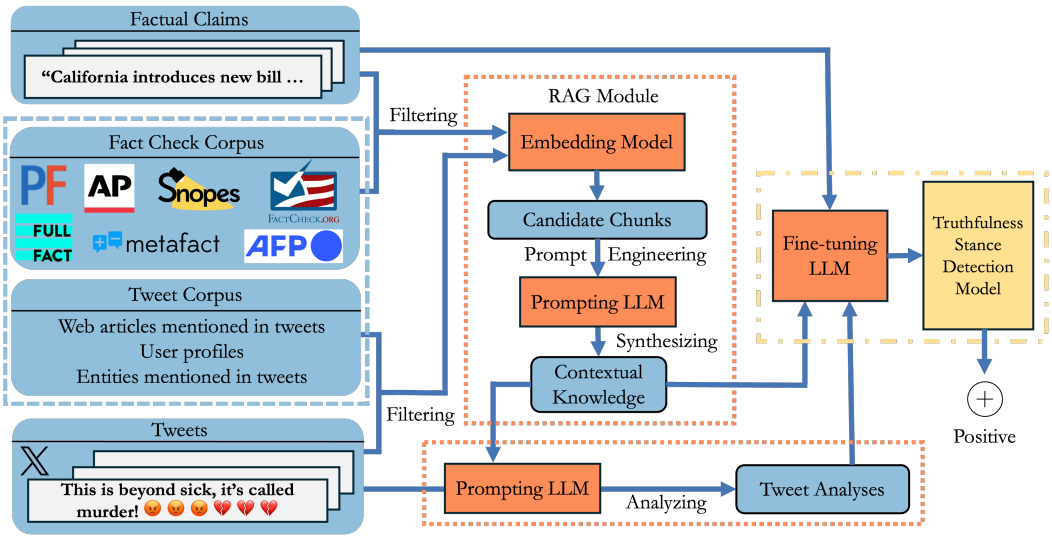

Figure 2: The RATSD framework.

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, 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 to happen, or the target claim's truthfulness).

Section 2 presents this conceptual framework of stance definitions in greater detail. Note that, although prior studies (Küçük and Can, 2020; Hardalov et al., 2022; Alturayeif et al., 2023) touch on various aspects of it, our systematic articulation of the conceptual framework is a significant contribution to the field, as such a nuanced and fine-grained analysis is currently absent from the literature. With regard to the particular definition of stance studied in this paper, ours is the first that focuses on social media posts' stance regarding *general* claims' truthfulness, as explained in Section 2 as well.

Section 3 proposes novel methods for truthfulness stance detection. Particularly, we have developed an LLM-empowered framework RATSD (**r**etrieval **a**ugmented **t**ruthfulness **s**tance **d**etection), as shown in Figure 2. 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' capability in analyzing tweets' stance, incorporating contextual information into the learning process, and also helps reduce the informal tone typically found in social media. *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.

Appendix A introduces our new benchmark dataset TSD-CT, which is specifically designed for training and assessing truthfulness stance detection models. The dataset comprises 3,105 claim-tweet pairs (hence the CT in its name) in which the stances of the tweets toward the truthfulness of factual claims from PolitiFact were labeled using our in-house annotation tool. This tool includes various quality control mechanisms to ensure the accuracy of the annotated dataset.

Section 4 discusses our experiments with RATSD on TSD-CT and three existing 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.

In summary, this paper's contributions are as follows:

- It introduces a novel conceptual framework for defining stance and focuses on a unique task formulation for truthfulness stance detection.
- We created a new benchmark dataset, TSD-CT, which has the potential to become a staple for this area of research and computational social science at large.
- We introduce RATSD, a novel approach that leverages RAG in stance analysis and highlights the effectiveness of LLM in detecting truthfulness stance.
- The TSD-CT dataset and RATSD's codebase are available at anonymous.4open.science/r/RATSD-3E7B. The repository is anonymized to comply with the review process and will be made publicly accessible upon paper acceptance, to promote research reproducibility and support further studies.

2 Conceptual Framework and Task Definition

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

As summarized in Section 1, the conceptual framework of stance consists of four components: the utterance, target, orientation, and type of the stance. This section explores each component in greater detail and discusses how our work both aligns with and diverges from existing definitions of stance in the literature.

Orientation of Stance. Figure 3 illustrates the relationship among all stance labels. Note that our truthfulness stance detection model does not consider unrelated pairs, because detecting the relevance between c and t falls under the realm of areas such as textual semantic similarity (Wang and Dong, 2020; Gomaa et al., 2013), which is beyond the scope of this work. We did annotate unrelated pairs in creating TSD-CT, though, in order to exclude such pairs in training and evaluating detection models.

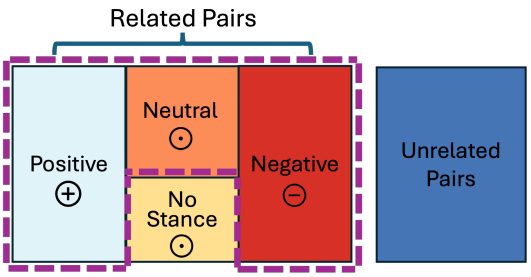

Figure 3: A claim c and a tweet t may be related or unrelated. Related claim-tweet pairs are partitioned into four cases of stance. The cases within the dashed, purple frame (positive, neutral, negative) are applicable when t expresses a stance regarding c's truthfulness. Though conceptually different, for practical purposes neutral and no stance are combined into one class $\odot$ in our consideration.

Conceptually, we recognize the difference between neutral stance and no stance. A tweet has the neutral stance if it expresses mixed verdict or uncertainty about a claim's truthfulness. On the other hand, it has no stance if the tweet, while being related to the claim topic-wise, does not express an intentional stance that reflects beliefs and desires (Dretske, 1988) towards the claim's truthfulness. This distinction is also similarly recognized in some existing studies, e.g., SemEval-19 and (Grimminger and Klinger, 2021), although they use different vocabulary of stance labels.

In practice, though, discerning no stance is highly challenging. The examples (3) in Figure 7 in the Appendix demonstrates one such challenging case. Although the tweet is highly pertinent to the claim as it mentions Paul Ryan, gun laws, and "action", it does not say whether Paul Ryan has blocked such actions or not. Its stance is not neutral. Rather, it does not express a stance toward the claim's truthfulness. Neutral stance and no stance in many cases have strong similarities. The difficulty is reflected in that, among all pairs of stance labels, the (neutral stance, no stance) pair received the lowest inter-annotator agreement among the expert annotators during our dataset preparation.

Given this intrinsic challenge, we decided to combine neutral stance and no stance into a single class $\odot$, both in dataset creation and detec-

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)	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)	Emergent (Ferreira and Vlachos, 2016); FNC-1 (Pomerleau and Rao, 2017); COVIDLies (Hossain et al., 2020); This work (TSD–CT)

Table 1: Various definitions of stance differ in the type, utterance, and target of stance.

tion model construction. The similar approach was taken in sentiment analysis, in which Koppel and Schler (2006) categorized documents' neutrality sentiment into two types. The first type of neutrality sentiment (akin to no stance in our scheme) is for documents containing objective information, without expressing clear sentiment, while the second type (akin to neutral stance) is for documents expressing mixed positive and negative sentiment.

Utterance of Stance. Table 1 compares the definitions of stance in existing datasets. It includes both dataset names (if available) and corresponding references. Our ensuing discussion refers to the names, whenever available. Researchers have proposed various methods and models for these datasets and corresponding stance detection tasks. Such models are referenced throughout this paper but not necessarily in Table 1.

To differentiate the utterances in existing and our definitions, Table 1 uses two colors—brown for news articles and blue for social media posts (mostly tweets, although SemEval-2019 includes Reddit posts and VAST considers comments in news websites). For 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 structures.

Target of Stance. Table 1 shows four main kinds of stance *targets* in prior studies, including 1) entities (e.g., Hillary Clinton) and topics (e.g., "legalization of abortion") in SemEval-2016, VAST, P-Stance, and (Grimminger and Klinger, 2021), 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 shootings at Parliament Hill in Ottawa), 3) factual claims (e.g., news claims in Emergent, news headlines in FNC-1, and COVD-19 related misconceptions in COVIDLies), and 4) fact triples in NewsClaims, where the stance is about whether a news article affirms or refutes a particular "claim object" (e.g., Vitamin C) being an origin of, transmission method of, cure for, or protection against the COVID-19 virus. Each stance target there can thus be viewed as a fact triple, e.g., (Vitamin C, cure, COVID-19 virus).

Note that the datasets also vary in the number of targets, ranging from fewer than ten (one in MGTAB, three in P-Stance and (Grimminger and Klinger, 2021), five in WT-WT, and six in SemEval-2016), to dozens and hundreds (86 misconceptions in COVIDLies, 300 news claims in Emergent, 889 fact triplets in NewsClaims, and several hundred latent rumors in PHEME, SemEval-2017 and SemEval-2019 as their source tweets, each mentioning a rumor, are on that scale), and thousands (1,520 factual claims in our TSD-CT, 2542 news headlines in FNC-1, 5634 topics in VAST).

Type of Stance. The *type* of stance in various existing definitions is about the likelihood of target events going to happen (e.g., WT-WT), about favorability (e.g., SemEval-2016, VAST, P-Stance, (Grimminger and Klinger, 2021), and MGTAB)—whether the stance expressed in an utterance is in favor of or against a given target, or about the truthfulness of a rumor (PHEME, SemEval-2017 and SemEval-2019), a news headline (FNC-1), or a claim (Emergent and COVIDLies). These stance types are not equivalent and thus must be tackled by different detection models. This is well demonstrated by the upper-left example in Figure 1—the tweet expresses a negative favorability stance but a positive truthfulness stance regarding the claim.

Among the aforementioned datasets and definitions of stance, TSD-CT resembles COVIDLies the most, since they both are about tweets'

stance regarding the truthfulness of factual claims. One key distinction is that COVIDLies (https://ucinlp.github.io/covid19/misinformation/) 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, coming from PolitiFact, are more complex both syntactically and semantically, and they cover a range of topics that interest fact-checkers. It is also worth noting that COVIDLies only includes false claims as targets, while TSD-CT contains a mix of true and false claims. The way truthfulness stance is expressed might be different regarding true vs. false claims.

3 Methodology

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

RATSD counters these challenges with two innovative data augmentation ideas, both leveraging LLMs' abilities. One is to employ RAG (retrieval augmented generation) to retrieve from external knowledge corpora relevant contextual information that compensates for the inherent lack of context. The other is to synthesize an analysis of the tweet t based on the context. The tweet analysis directly incorporates LLM's perspective regarding t's truthfulness stance toward c. It also has the benefit of reducing the aforementioned informal language from the process. 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 2 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.

3.1 Knowledge Corpora Construction

Two knowledge corpora were constructed for supplying contextual knowledge to other components in RATSD, for claims and tweets, respectively.

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 includes claims and corresponding fact-checks published by seven fact-checking websites from 1995 to 2023, and 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 of a fact-check 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, the summary of the fact-checking ruling provided in the article; and the main body of the fact-check article. The resulting d_c typically contains between 10,000 and 30,000 words.

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.

3.2 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 for training and evaluating the models, 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. 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 not identical to the set of tweets for which d_t was generated.

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) such as unfaithful information for fact-checking (Kim et al., 2024). The generation process of e_c and e_t follows three steps: 1) relevant document selection, 2) relevant chunk retrieval, and 3) prompting LLM.

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 parts of the selected top documents are relevant to c and t. We segment the documents into smaller chunks (i.e., continuous sequence of tokens), each with a token size of 512, and retrieve the most relevant chunks. This was accomplished by using the BAAI general embedding (BGE) model (Xiao et al., 2023) to generate text embeddings and applying cosine similarity to measure the semantic alignment between each chunk and a prompt. Given each c and and similarly each t, the top 10 most relevant chunks were retrieved. The BGE model, being a lightweight, pre-trained model, has demonstrated strong performance in the text embedding leaderboard (Muennighoff et al., 2023). The prompt used in this step is the same prompt instruction that is used in step (3) when prompting the LLM.

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 5 in the Appendix shows an example prompt. As described above, this prompt was used to find relevant chucks 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. An example of e_c and e_t can be found in Figure 5.

3.3 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 6.

When training the stance detection model, a will replace t in the input claim-tweet pair, as detailed in Section 3.4. 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. (Section 4 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.

3.4 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 vector is fed into a single fully connected layer and a softmax layer to produce the probability distribution of stance orientation labels $\{\hat{s}_i^{\oplus}, \hat{s}_i^{\odot}, \hat{s}_i^{\ominus}\} = softmax(Wh_i + b)$ 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 ground-truth 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 ground-truth stance distribution.

4 Evaluation

4.1 Experiment Datasets

As noted in Section 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 3.

4.2 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. 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.

4.3 Experiment Results

We evaluated the performance of two types of stance detection models: LM-based and LLM-based. And the evaluation was conducted in two different settings: fine-tuning the models for stance detection and applying them directly in a zero-shot setting. Consistent with previous studies, we used F1 scores for each class—denoted as $F_{\oplus}$, $F_{\odot}$, and $F_{\ominus}$—and the Macro F1 score (F_M) as our evaluation metrics.

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., 2019)), 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 2, RATSD demonstrates strong performance across all datasets compared to other stance detection models. On the TSD-CT dataset, $\text{RATSD}_{\text{GPT-3.5}}$ achieves the highest scores across all metrics. For the SemEval-2019 dataset, $\text{RATSD}_{\text{GPT-3.5}}$ surpasses other models in $F_{\odot}$ score and achieves the highest macro F1 score. $\text{RATSD}_{\text{Zephyr}}$ demonstrates its strength on the WT-WT dataset, where it secures the highest performance across all metrics. While on the COVIDLies dataset, BERTSCORE+NLI and BART+NLI slightly outperform RATSD, RATSD still delivers competitive results. These results suggest that while RATSD models, especially $\text{RATSD}_{\text{GPT-3.5}}$, 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.

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 4. Among the models, $\text{RATSD}_{\text{Zephyr}_{\text{zero}}}$ achieves the highest overall performance, with the F_M of 36.55. This suggests that $\text{RATSD}_{\text{Zephyr}_{\text{zero}}}$ is a strong framework for truthfulness stance detection in the zero-shot setting. $\text{RATSD}_{\text{Zephyr}_{\text{zero}}}$ outperforms $\text{RATSD}_{\text{GPT-3.5}_{\text{zero}}}$ 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, $\text{GPT-3.5}_{\text{zero}}$ uses direct prompting, as described in Section 3.3, which demonstrates its difficulty in

Model	TSD–CT				SemEval-2019				WT-WT				COVIDLies			
	$F_{\oplus}$	$F_{\odot}$	$F_{\ominus}$	F_M	$F_{\oplus}$	$F_{\odot}$	$F_{\ominus}$	F_M	$F_{\oplus}$	$F_{\odot}$	$F_{\ominus}$	F_M	$F_{\oplus}$	$F_{\odot}$	$F_{\ominus}$	F_M
BUT-FIT	83.38	72.00	65.11	80.11	49.09	50.98	92.01	64.03	81.29	94.73	79.29	85.10	47.62	97.82	23.53	56.32
BLCU_NLP	85.37	71.43	63.29	73.36	**70.15**	40.00	88.12	66.09	81.02	94.74	77.09	84.28	52.38	97.71	45.46	65.18
BERTSCORE+NLI	88.68	72.53	81.04	80.75	46.96	60.67	91.32	66.32	82.02	95.06	79.11	85.39	**57.14**	**98.20**	58.33	**71.22**
BART+NLI	88.00	73.42	74.25	78.56	47.96	51.71	91.90	63.86	82.82	95.52	81.75	86.70	50.00	98.00	**60.87**	69.62
TESTED	84.09	72.37	67.90	74.75	46.43	58.04	**92.08**	65.52	81.75	94.98	78.00	85.91	40.00	97.12	51.85	62.99
RATSD$_{\text{Zephyr}}$	88.67	77.38	80.28	82.10	41.71	55.42	91.80	62.97	**83.85**	**95.72**	**82.66**	**87.44**	51.42	97.63	54.55	67.87
RATSD$_{\text{GPT-3.5}}$	**93.27**	**80.24**	**87.90**	**87.13**	56.12	**63.79**	83.67	**67.86**	75.78	92.98	75.07	81.27	51.16	98.06	52.63	67.30

Table 2: Performance comparison on datasets TSD-CT, SemEval-2019, WT-WT, and COVIDLies.

Dataset	($\oplus$)	($\odot$)	($\ominus$)	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 3: Label distribution of SemEval-2019, WT-WT, COVIDLies and TSD-CT datasets.

achieving strong performance without fine-tuning.

Model	$F_{\oplus}$	$F_{\odot}$	$F_{\ominus}$	F_M
BUT-FIT$_{\text{zero}}$	12.82	0.00	33.88	15.56
BLCU_NLP$_{\text{zero}}$	27.05	0.00	32.81	19.95
BERTSCORE+NLI$_{\text{zero}}$	6.82	**41.71**	17.65	22.06
BART+NLI$_{\text{zero}}$	33.55	40.58	3.96	26.03
TESTED$_{\text{zero}}$	**55.84**	38.91	4.04	32.93
GPT-3.5$_{\text{zero}}$	34.04	16.81	**39.74**	30.20
RATSD$_{\text{Zephyr}_{\text{zero}}}$	49.74	32.14	27.78	**36.55**
RATSD$_{\text{GPT-3.5}_{\text{zero}}}$	28.76	29.71	33.46	30.64
RATSD$_{\text{Zephyr}}$	**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 4: The zero-shot model performance comparison and ablation study on the TSD-CT dataset.

4.4 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$_{\text{Zephyr}}$ without stance analysis (w/o analysis) and RATSD$_{\text{Zephyr}}$ without contextual knowledge generation and stance analysis (w/o context & analysis). The results in the bottom three rows of Table 4 reveal the impact of key components on the performance of RATSD$_{\text{Zephyr}}$. 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.

5 Related Work

Recent studies show that LLM such as GPT-3.5 can achieve impressive results in stance detection (Zhang et al., 2022, 2023). Researchers have explored the incorporation of contextual knowledge to enhance stance detection model performance. For example, Li et al. (2023) developed a topic-based heuristic algorithm to retrieve relevant Wikipedia documents for input instances; however, their approach does not utilize RAG and stance analysis as ours does. Zhang et al. (2024) and Singhal et al. (2024) proposed methods to prompt LLMs to generate contextual knowledge, but their approaches do not rely on external knowledge corpora. Additionally, the stance considered in these works (Li et al., 2023; Zhang et al., 2024; Singhal et al., 2024) focuses on favorability rather than truthfulness.

6 Conclusion

In this paper, we revisit the concept of stance detection, develop a new dataset based on this revised definition, and propose an approach to truthfulness stance detection. Our method, which merges machine learning techniques with an understanding of the linguistic and contextual factors that influence stance, has yielded encouraging results. We have tested its performance on various datasets, demonstrating its adaptability and resilience. In conclusion, our research contributes to the ongoing efforts in the automatic detection of truthfulness stance. We hope that our findings in truthfulness stance problem definition, TSD-CT dataset and RATSD will be useful for future research in this dynamic and rapidly progressing field.

Limitations

One major limitation of RATSD is its inability to distinguish between neutral stance and no stance, which were combined into a single category due to their inherent similarities. While conceptually different, it is challenging to reliably differentiate between the two in practice. This difficulty was evident in the low inter-annotator agreement for these labels during dataset preparation. By merging them into one category for simplicity, RATSD loses the ability to capture this finer distinction, which could be valuable in certain applications where discerning a lack of stance versus neutral stance is important. This trade-off exposes the model's limitation in handling nuanced stance differentiation.

Another limitation of RATSD is its inability to detect truthfulness stance across multiple sub-claims within a single factual claim. For example, in the factual claim "We won and we won a lot," which contains two sub-claims ("We won" and "We won a lot"), RATSD treats the entire statement as a single unit, potentially oversimplifying the semantic complexity of such claims. This approach may overlook the differences in stance that can exist across multiple sub-claims.

Furthermore, while RATSD is capable of processing multi-sentence claims and even paragraphs, the dataset used for its training and evaluation consists exclusively of one-sentence claims. As the dataset was not designed for handling multi-sentence factual claims, the model's performance on such inputs may be less effective. This points to the need for a more comprehensive dataset that includes multi-sentence claims to fully assess RATSD's capabilities in these scenarios.

Ethics and Risks

Bias and Misinformation. One significant risk involves the potential for bias in the model's predictions, arising from both the training data and annotations. The dataset may reflect biases from social media environments or inherent biases in the annotators themselves, which could influence the model's outputs. Additionally, while RATSD is designed to detect stances, there is a risk that users might misinterpret the model's outputs as a measure of truthfulness. This could inadvertently amplify misinformation if the model's stance detection is taken as an endorsement or validation of false claims. Mitigating these risks requires careful curation of training data, clear disclaimers, and transparency in how the model's outputs should be interpreted.

Impact on Public Discourse. By automating stance detection on factual claims, RATSD has the potential to influence public discourse, particularly in highly polarized contexts, such as politics, public health, or social justice. Misuse of the system could lead to the selective presentation of results, thereby reinforcing biased narratives or suppressing dissenting opinions. To prevent this, it is critical to ensure that the model's use is transparent and its limitations are well understood, emphasizing that it is a supportive tool for human decision-making rather than a replacement.

Privacy and Data Protection. Privacy concerns are particularly relevant when analyzing social media posts. Users often share personal opinions in semi-public spaces. Applying automated tools, such as RATSD, must adhere to data protection and privacy regulations. It is imperative to respect user consent, anonymize data, and handle information responsibly in compliance with legal and ethical standards. Failure to do so could lead to violations of privacy and erode trust in the system.

Ethical Considerations in Annotation. Ethical considerations extend to the annotation process itself. The content of the tweets may negatively impact annotators, exposing them to harmful or distressing material. To address this, we obtained Institutional Review Board (IRB) approval before recruiting annotators and implemented guidelines to protect their well-being.

References

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

Jamal Al Qundus, Adrian Paschke, Shivam Gupta, Ahmad M Alzouby, and Malik Yousef. 2020. Exploring the impact of short-text complexity and structure on its quality in social media. *Journal of Enterprise Information Management*, 33(6):1443–1466.

Emily Allaway and Kathleen R. McKeown. 2020. Zero-shot stance detection: A dataset and model using generalized topic representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 8913–8931.

Nora Saleh Alturayeif, Hamzah Luqman, and Moataz A. Ahmed. 2023. A systematic review

of machine learning techniques for stance detection and its applications. *Neural Comput. Appl.*, 35(7):5113–5144.

Erik Arakelyan, Arnav Arora, and Isabelle Augenstein. 2023. Topic-guided sampling for data-efficient multi-domain stance detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 13448–13464.

Lasse Bergroth, Harri Hakonen, and Timo Raita. 2000. A survey of longest common subsequence algorithms. In *Proceedings Seventh International Symposium on String Processing and Information Retrieval*, pages 39–48.

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

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. 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.

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

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

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

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

Martin Fajcik, Pavel Smrz, and Lukás Burget. 2019. BUT-FIT at semeval-2019 task 7: Determining the rumour stance with pre-trained deep bidirectional transformers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1097–1104.

William Ferreira and Andreas Vlachos. 2016. 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.

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

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

Lara Grimminger and Roman Klinger. 2021. Hate towards the political opponent: A twitter corpus study of the 2020 US elections on the basis of offensive speech and stance detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 171–180.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022. A survey on stance detection for mis- and disinformation identification. In *Findings of the Association for Computational Linguistics*, pages 1259–1277.

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

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating LLM hallucination via self reflection. In *Findings of the Association for Computational Linguistics*, pages 1827–1843.

Kyungha Kim, Sangyun Lee, Kung-Hsiang Huang, Hou Pong Chan, Manling Li, and Heng Ji. 2024. Can LLMs produce faithful explanations for fact-checking? towards faithful explainable fact-checking via multi-agent debate. *arXiv preprint arXiv:2402.07401*.

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

Moshe Koppel and Jonathan Schler. 2006. The importance of neutral examples for learning sentiment. *Computational Intelligence*, 22(2):100–109.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37.

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

Ang Li, Bin Liang, Jingqian Zhao, Bowen Zhang, Min Yang, and Ruifeng Xu. 2023. Stance detection on social media with background knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15703–15717.

Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 2021. P-Stance: A large dataset for stance detection in political domain. In *Findings of the Association for Computational Linguistics*, pages 2355–2365.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, pages 31–41.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. MTEB: massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2006–2029.

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

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

Revanth Gangi Reddy, Sai Chetan Chinthakindi, Zhenhailong Wang, Yi R. Fung, Kathryn Conger, Ahmed Elsayed, Martha Palmer, Preslav Nakov, Eduard H. Hovy, Kevin Small, and Heng Ji. 2022. Newsclaims: A new benchmark for claim detection from news with attribute knowledge. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6002–6018.

Shuhao Shi, Kai Qiao, Jian Chen, Shuai Yang, Jie Yang, Baojie Song, Linyuan Wang, and Bin Yan. 2023. MGTAB: A multi-relational graph-based Twitter account detection benchmark. *arXiv preprint arXiv:2301.01123*.

Ronit Singhal, Pransh Patwa, Parth Patwa, Aman Chadha, and Amitava Das. 2024. Evidence-backed fact checking using RAG and few-shot in-context learning with LLMs. *arXiv preprint arXiv:2408.12060*.

Ivan Smirnov. 2017. The digital flynn effect: Complexity of posts on social media increases over time. In *Social Informatics: 9th International Conference*, pages 24–30. Springer.

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

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

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

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*.

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

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

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

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

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

Ruoyao Yang, Wanying Xie, Chunhua Liu, and Dong Yu. 2019. Blcu_nlp at semeval-2019 task 7: An inference chain-based GPT model for rumour evaluation. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1090–1096.

Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, and Li Yuan. 2023. LLM lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469*.

Bowen Zhang, Daijun Ding, Zhichao Huang, Ang Li, Yangyang Li, Baoquan Zhang, and Hu Huang. 2024. Knowledge-augmented interpretable network for zero-shot stance detection on social media. *IEEE Transactions on Computational Social Systems*.

Bowen Zhang, Daijun Ding, and Liwen Jing. 2022. How would stance detection techniques evolve after the launch of chatgpt? *arXiv preprint arXiv:2212.14548*.

Bowen Zhang, Xianghua Fu, Daijun Ding, Hu Huang, Yangyang Li, and Liwen Jing. 2023. Investigating chain-of-thought with chatgpt for stance detection on social media. *arXiv preprint arXiv:2304.03087*.

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

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 11(3):e0150989.

A Creation of the TSD-CT Dataset

A.1 Fact-check Collection

We developed a tool to collect the fact-checks from seven well-known fact-checking websites, including AFP Fact Check, AP Fact Check, FactCheck.org, FullFact, Metafact, PolitiFact, and Snopes. The collected data is coded using ClaimReview's data schema, a widely adopted standard for structuring fact-checks. The data schema includes fields such as Publisher, ClaimReviewed, Summary, Review, Verdict, Author, ClaimPublishedDate, FactcheckPublishedDate, ThumbnailURL, URL and Tags. Particularly, the Summary field provides a summary of the fact-check, while the Review field contains the main body of the fact-checking article, including background information and the evidence supporting the verdict.

A.2 Claim-tweet Pair Collection

In constructing the claim-tweet pair collection, we selected factual claims from PolitiFact in the fact-check collection, excluding those phrased as questions. We then using *spaCy* to extract keywords (nouns, verbs, adjectives, pronouns, and numbers) from the claims. For each factual claim, we retrieved related tweets via Twitter API v2 using a conjunctive (ANDed) query formed by the extracted keywords from the claim. We filtered out tweets with fewer than 30 characters, as well as retweets, replies, and quotes, to avoid duplicates. To ensure temporal relevance between tweets and factual claims, we limited the API search to tweets posted within one month before the factual claim and one year after it. This process led to 36,154 claim-tweet pairs.

A.3 Claim-tweet Pair Sanitization

A claim-tweet was removed from the collection if it triggers one of the following conditions: (1) the tweet closely resembles the factual claim, with a similarity score higher than 0.9; (2) the tweet was nearly identical to the tweet in another pair (the pair containing the tweet collected earlier was kept and the one later was removed), with a similarity score above 0.8; and (3) the tweet was published by a fact-checker and it fact-checks the claim. For (1) and (2), the similarity scores were calculated by removing links and hashtags from tweets and applying the longest contiguous matching subsequence (LCS) algorithm (Bergroth et al., 2000). Removing these pairs can avoid wasting efforts in annotating similar pairs and can help diversify the dataset. For (3), such pairs were identified using a heuristic rule: the tweet contains the claim and any of the following: a hyperlink to a fact-check, the name of a fact-checking website, or the claimant's name at the beginning of the tweet. Removing these pairs can avoid wasting efforts in annotating easy cases where the tweet's stance could be highly accurately labeled according to the fact-checker's verdict regarding the claim. After removing 30,032 pairs in cases (1) and (2) and 329 pairs in case (3), we were left with 2,283 unique factual claims paired with 5,793 tweets from 5,227 distinct twitter accounts.

A.4 Claim-tweet Pair Annotation

Annotation Website. The human annotation was conducted through our in-house website, as shown in Figure 4. For each claim-tweet pair, the annotation task is to choose one of the five distinct options, as shown in the figure. These options correspond to the positive stance, neutral/no stance, negative stance, and unrelated pairs ("different topics") in Figure 3, and an additional "problematic" option. This "problematic" option allows annotators to flag tweets that are created for sarcasm only or contain invalid links.

To help annotators better understand the task, the website also includes a detailed instruction page that provides a clear definition of the task along with thorough explanation of each classification class. For each class, we included three examples of claim-tweet pairs, each accompanied by the correct class and analysis of the choices. Additionally, we developed an administrative progress monitoring page to streamline the annotation process. This page enabled us to monitor the overall progress,

DataSource	AFP Fact Check	AP Fact Check	FactCheck.org	FullFact	Metafact	PolitiFact	Snopes
Claims	0*	297	0*	2,783	3,428	21,023	18,097
Review Summary	4,204	297	3,452	2,783	0*	21,023	2,638
Review	4,304	297	3,452	2,783	3,428	21,022	18,474
Verdict	0*	225	0*	0*	3,428	21,023	13,947

Table 5: Numbers of claims, review summaries, reviews, and verdicts in 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. In such cases, a document d_c is generated in the knowledge corpus to represent the latent claim c in each fact-check article.

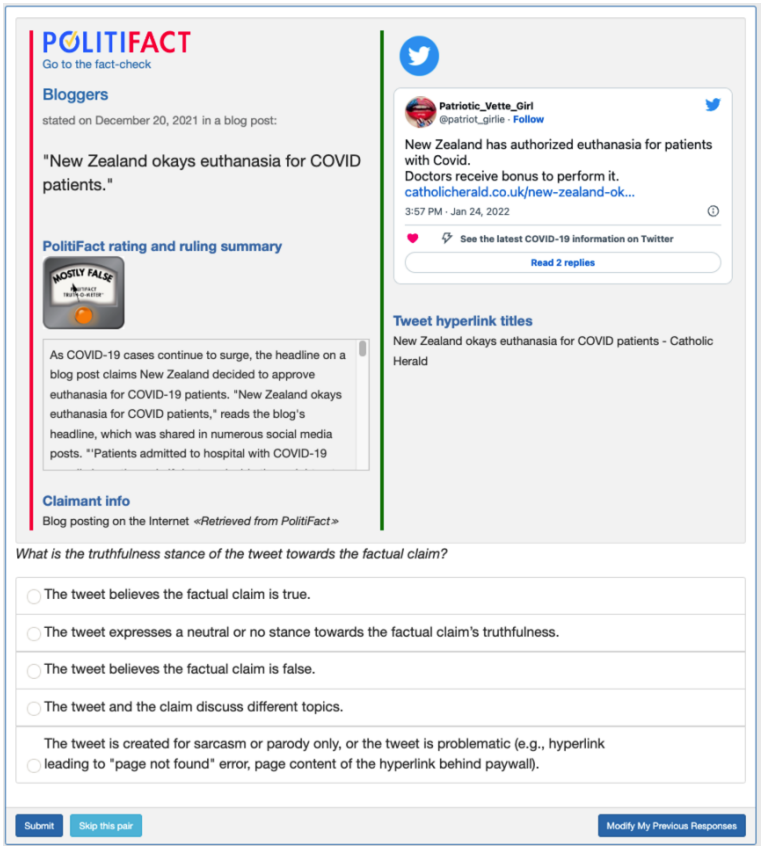

Figure 4: The annotation interface.

track annotator performance, and examine detailed annotation history of each individual claim-tweet pair.

Annotators. During annotator recruitment, we distributed flyers and emailed announcements across the campus of a university. All annotators were at least 18 years old and fluent in English. Compensation for their work is provided either in the form of gift cards or through the payroll system of the university. Their earnings were determined by the quality of their annotations, with the potential to earn up to 20 US cents for each claim-tweet pair they annotate.

Quality Control. To identify high-quality annotators, we used 287 carefully selected screening pairs. Each pair received consistent labeling from five researchers. These pairs were mixed with the pairs that need real annotation. They were randomly chosen and presented to an annotator at an average frequency of one in every ten pairs, without the annotator's knowledge. Annotators were scored based on how well their labels match the experts' labels on the screening pairs. Annotations from low-quality annotators were excluded from the dataset. A pair's annotation is considered complete when at least three high-quality annotators have worked on it, and more than half of their labels agree. The pair is then given that majority

label.

Among all 206 annotators, 30 were deemed high-quality based on the approach mentioned above. A total of 18,584 annotations were collected, with 13,594 from these high-quality annotators. This resulted in 3,105 completed pairs, containing 1,520 unique claims. Of the completed pairs, 216 were labeled as "different topics" and 669 as "problematic." As explained in Section 2, detecting unrelated pairs (i.e., "different topics") is a separate task and therefore falls outside the scope of our study, as does the detection of "problematic" pairs. Therefore, while these pairs are included in the released TSD-CT dataset, they were excluded from model training and evaluation. The class label distribution of TSD-CT is shown in Table 3.

B Challenging Examples in Truthfulness Stance Detection

To highlight the nuances among our stance labels within our conceptual framework, as well as the challenges encountered during the truthfulness stance annotation process, we present six challenging examples in Figure 7.

C Use of AI Assistants

Some of our code was developed using GitHub Copilot, and the writing was polished using ChatGPT and Grammarly.

Prompt: Consider this factual claim: {California introduces new bill that would allow mothers to kill their babies up to 7 days after birth.}. 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.}. Please enumerate entities along with their information and provide related details regarding those entities. Please generate the context knowledge for the factual claim and tweet separately in a structured way.

Contextual knowledge for factual claim:

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.
Assembly Member Buffy Wicks

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.
Facebook's Misinformation Policy

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.
Chelsea Becker and Adora Perez

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.
Verdict: False
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.

Contextual knowledge for Tweet:

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.
America First Report

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.
Verdict: False
Nature of the Tweet: The tweet is based on the same misinformation found in the Facebook post. It falsely portrays AB 2223 as a bill allowing infanticide, which is not the case.

Figure 5: A contextual knowledge generation example.

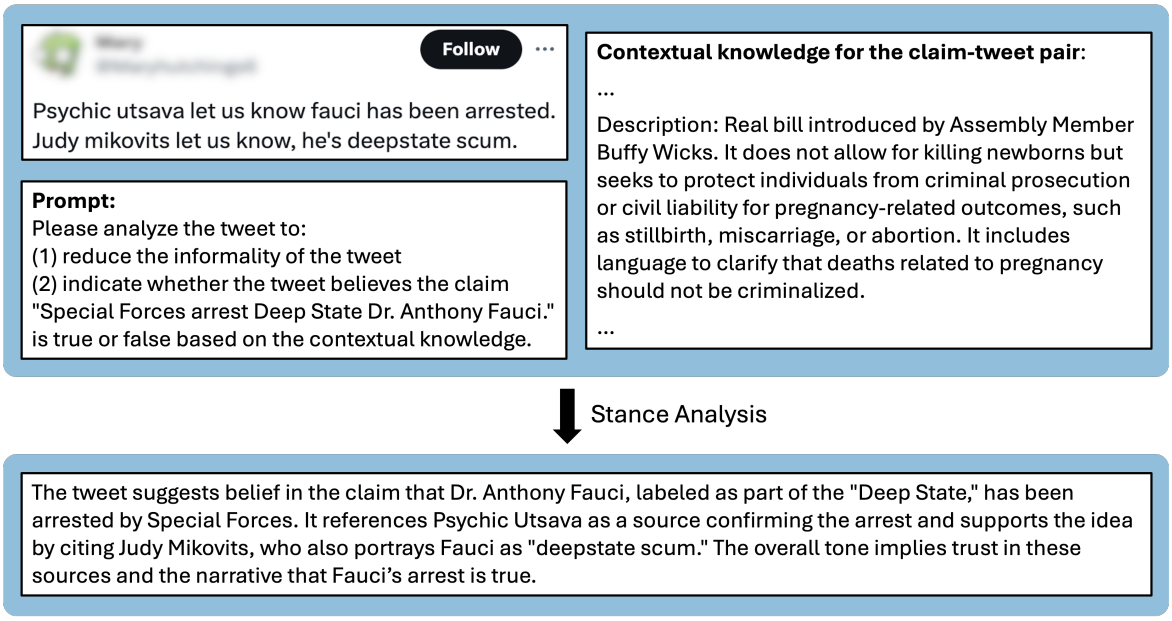

Figure 6: An example of stance analysis.

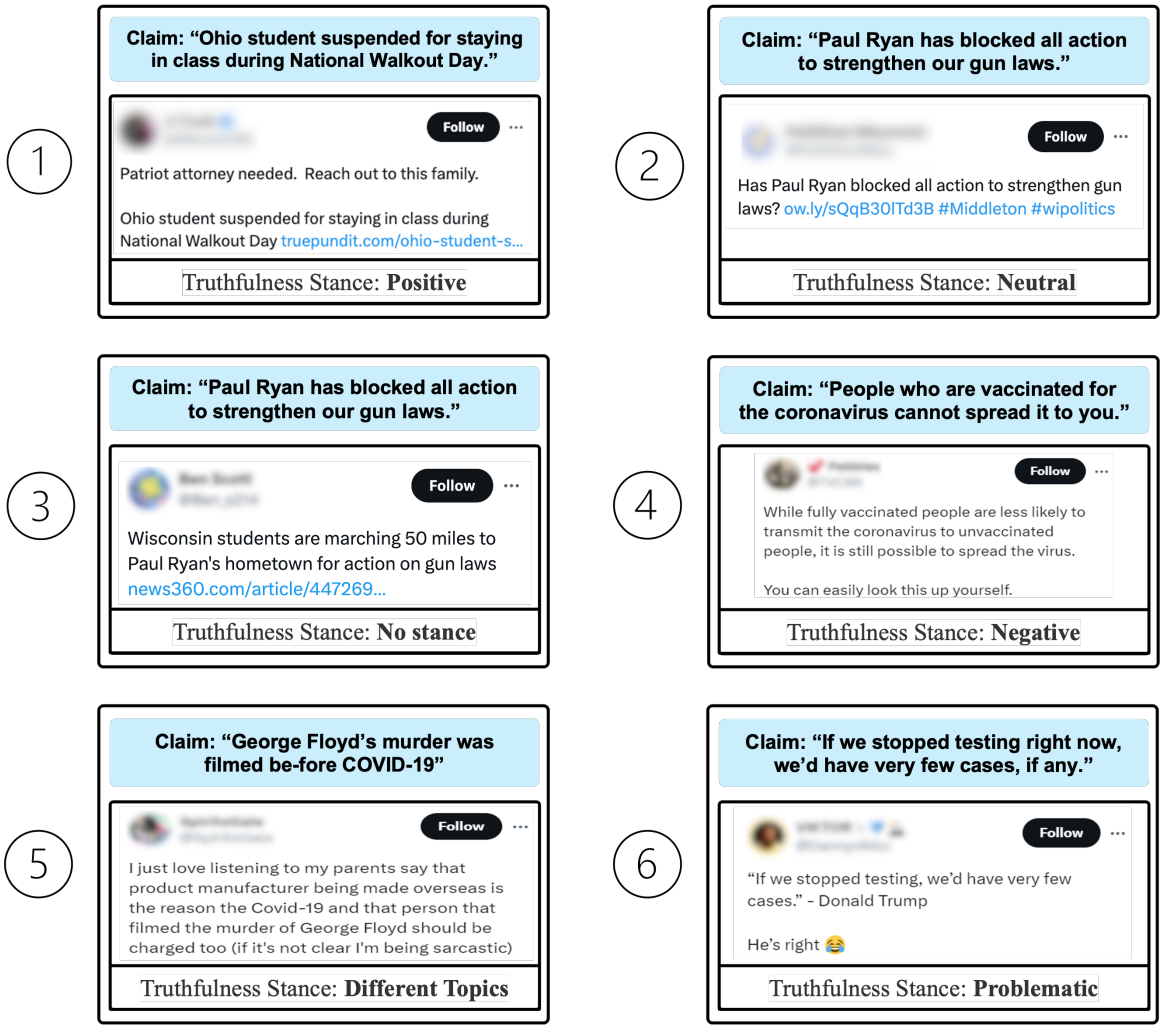

Figure 7: A few samples of challenging cases in truthfulness stance annotation.