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

University of Texas Arlington

Zhengyuan Zhu, Zeyu Zhang, Haiqi Zhang, Chengkai Li

iDIR Innovative Data Intelligence Research Laboratory

Truthfulness Stance Detection

- 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.
- Truthfulness stance: given a factual claim, assesses whether a textual utterance affirms its truth, disputes it as false, or expresses a neutral or indeterminate position.
- Truthfulness stance has the potential to be a useful tool in discerning how misinformation spreads and shapes decision-making in political discourse and health-related contexts.

A Conceptual Framework for Stance

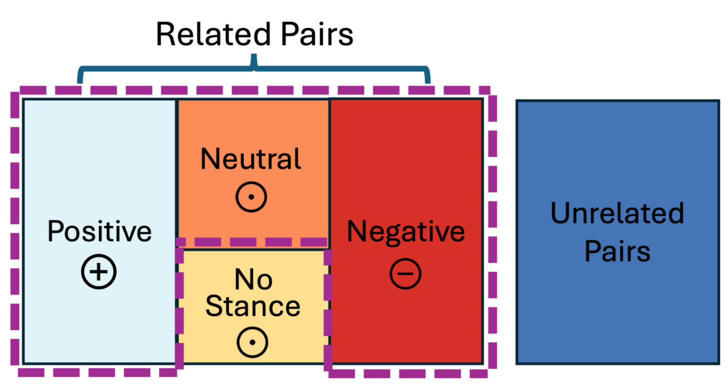
Type of Stance	Target of Stance: Entities or Topics	Events or Rumors	Fact Triples	Factual Claims
Favorability	SemEval-2016 (Mohammad et al., 2016); VAST (Allaway and Mckeown, 2020); P-Stance (Li et al., 2021); (Geimminger and Klinger, 2021); (Aleksandric el al., 2024)	MGTAB (Shi et al., 2023)		
Likelihood		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); FactBank (Sauri 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)

Utterance of Stance: ① news articles (in brown); ② social media posts (in blue).

Target of Stance: ① entities (e.g., Hillary Clinton) and topics (e.g., "legalization of abortion"); ② events (e.g., mergers and acquisitions of companies); ③ factual claims (e.g., news claims and news headlines) ④ fact triples extracted from the utterance itself.

Type of Stance: ① likelihood of target events occurring; ② favorability — determining whether the stance expressed in an utterance is in favor of or against a given target; ③ the truthfulness of a rumor, a news headline, a fact triple, or a claim.

Orientation of Stance: ① positive: a tweet conveys the belief that a claim is true; ② negative: a tweet believes a claim is false; ③ neutral/no stance: a tweet either expresses uncertainty about the truthfulness of a claim (neutral) or does not explicitly take a position on the claim's truthfulness (no stance).



TSD-CT Dataset

Fact-check Collection: Seven websites, 52,596 fact-checks (including associated factual claims) from 1995 to 2023.

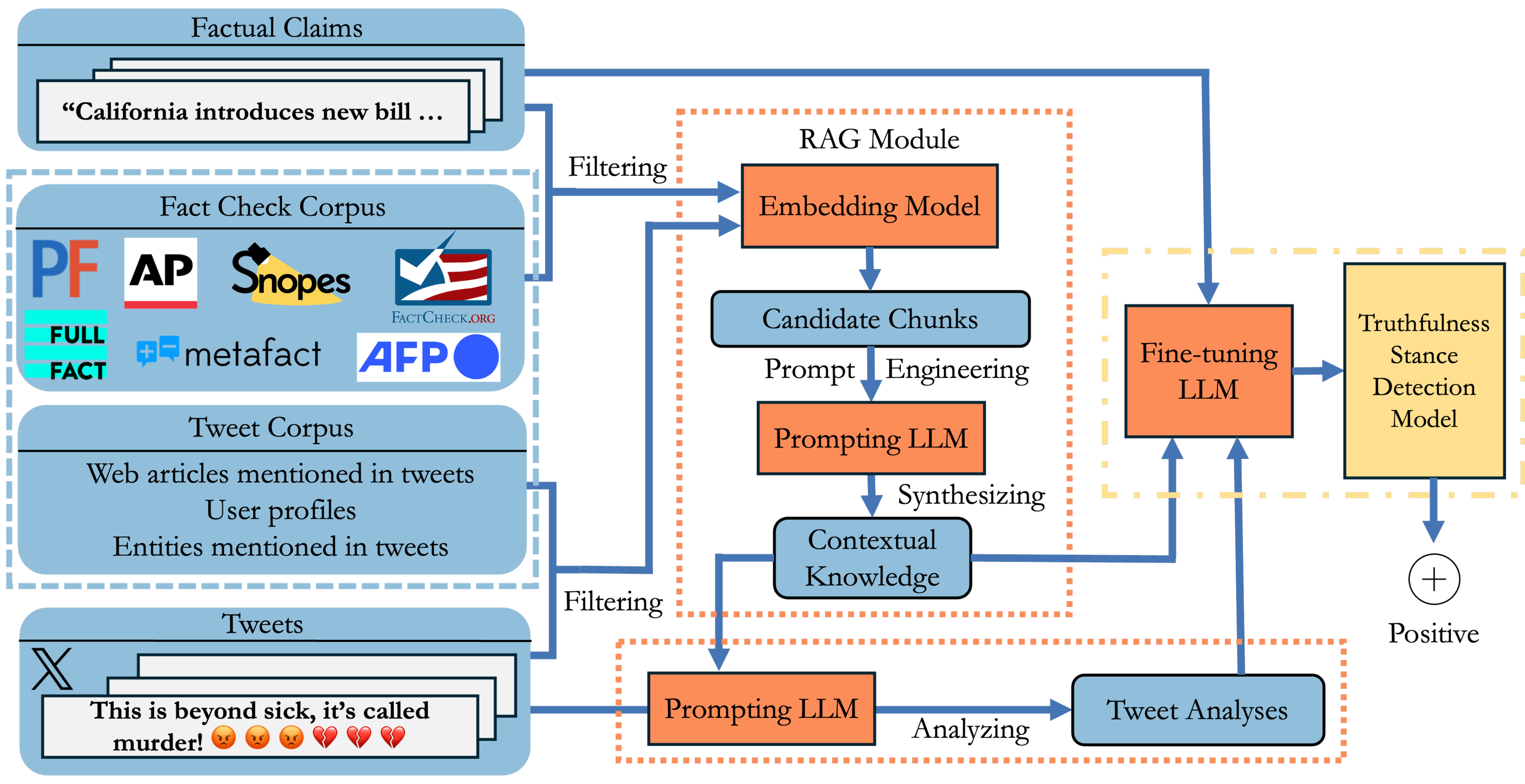
Claim-tweet Pair Collection:

- Extracted keywords from factual claims and retrieved related tweets using Twitter API v2, resulting in 36,154 pairs.
- After sanitization, retained 5,793 pairs for human annotation.

Claim-tweet Pair Annotation:

- In-house annotation website with detailed instructions, a progress monitoring page, and a leaderboard.
- Out of 206 annotators, 30 were deemed high-quality.
- Collected 3,105 annotated pairs containing 1,520 unique claims.

The RATSD Framework



Knowledge Corpora Construction: ① Factual claim knowledge corpus encompasses 52, 596 synthesized documents; ② Tweet knowledge corpus consists of 8, 236 synthesized documents from 2010 to 2023.

Contextual Knowledge Generation: (1) Document preprocessing (2) Relevant document selection (3) Relevant Chunk Retrieval (4) Prompting LLM

Stance Analysis: LLM is prompted using claim, tweet, contextual knowledge as the input to generate a narrative of tweet's truthfulness stance regarding claim.

Classification Model: LLM is fine-tuned using the claim-tweet pairs, stance analysis results, and contextual knowledge.

Evaluation

Our experiments used TSD-CT along with three benchmark datasets—SemEval-2019, WT-WT, and COVIDLies.

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

Model	TSD-CT $F_{\oplus}$	$F_{\odot}$	$F_{\ominus}$	F_M	SemEval-2019 $F_{\oplus}$	$F_{\odot}$	$F_{\ominus}$	F_M	WT-WT $F_{\oplus}$	$F_{\odot}$	$F_{\ominus}$	F_M	COVIDLies $F_{\oplus}$	$F_{\odot}$	$F_{\ominus}$	F_M
BUTFIT	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**	65.32	81.75	94.98	78.00	85.91	40.00	97.12	51.85	62.99
$RATSD_{Zephyr}$	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	98.63	54.55	67.87
$RATSD_{GPT-3.5}$	**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

Fine-tuned Model Performance

- Both RATSD variants demonstrate strong performance across all datasets.
- $RATSD_{GPT-3.5}$ achieved the highest scores across all metrics on the TSD-CT dataset
- Different fine-tuned LLM in RATSD may excel in specific datasets or stance categories, which highlights the importance of model selection based on dataset characteristics.

Zero-shot Performance on TSD-CT

- $RATSD_{Zephyr\ zero}$ achieves the highest overall performance (F_M=36.55).
- $RATSD_{Zephyr\ zero}$ and $RATSD_{GPT-3.5\ zero}$ are better suited for zero-shot scenarios on the TSD-CT dataset.

Model	$F_{\oplus}$	$F_{\odot}$	$F_{\ominus}$	F_M
$BUT\text{-}FIT_{zero}$	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	50.58	3.96	26.03
$TESTED_{zero}$	**55.84**	38.91	4.04	32.93
$GPT\text{-}3.5_{zero}$	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

Ablation Study

- Stance analysis provides useful additional context for both positive and neutral pairs.
- Contextual knowledge generation is crucial in handling negative pairs.

Model	$F_{\oplus}$	$F_{\odot}$	$F_{\ominus}$	F_M
$RATSD_{Zephyr}$	**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

Acknowledgements

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

https://github.com/idirlab/RATSD

GitHub

NAACL 2025