UNIVERSITY OF TEXAS ARLINGTON

iDiR

A Truthfulness Stance Map for 2024 Election-Related Factual Claims.

Zhengyuan Zhu

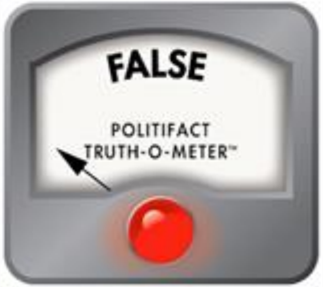
Fifth year PhD supervised by Dr. Chengkai Li

Research focus: Natural Language Processing

The Innovative Data Intelligence Research Laboratory (IDIR Lab)

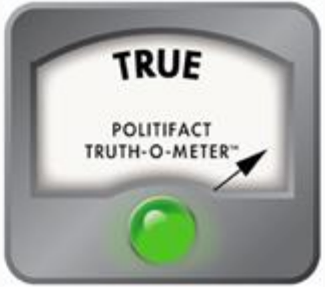
Problem Formulation

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



POLITIFACT

"Florida has the highest homeowners insurance in the nation."



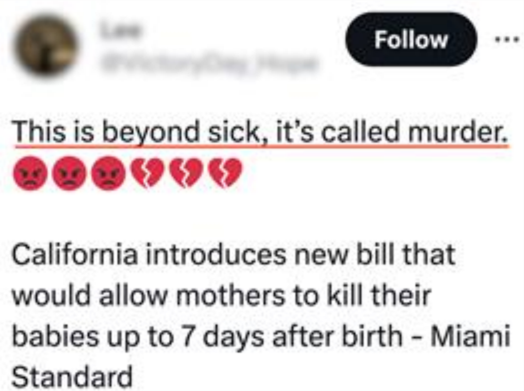
Truthfulness Stance Detection

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

believes the claim is true

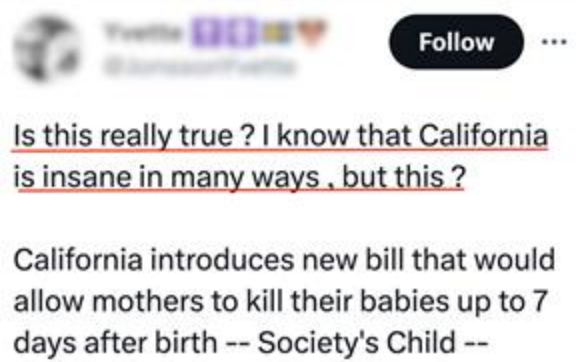
expresses a neutral stance

believes the claim is false



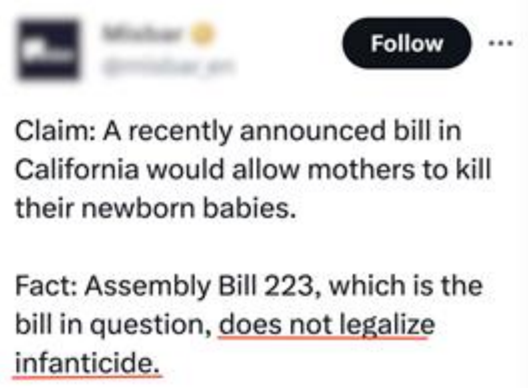
This is beyond sick, it's called murder. 😡😡😡💔💔💔

California introduces new bill that would allow mothers to kill their babies up to 7 days after birth - Miami Standard



Is this really true ? I know that California is insane in many ways , but this ?

California introduces new bill that would allow mothers to kill their babies up to 7 days after birth -- Society's Child --



Claim: A recently announced bill in California would allow mothers to kill their newborn babies.

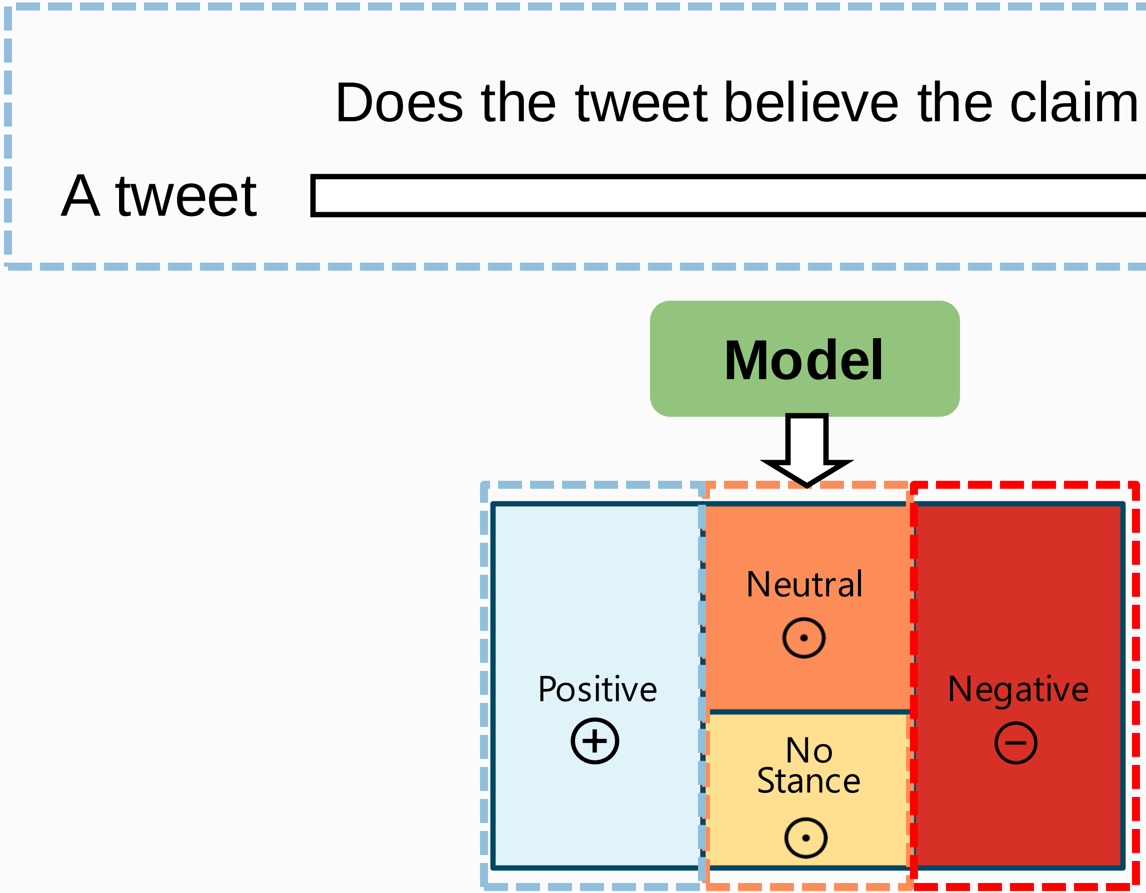
Fact: Assembly Bill 223, which is the bill in question, does not legalize infanticide.

Truthfulness stance detection: We detect the stance taken by tweets toward the truthfulness of factual claims.

Stance Detection Conceptual Framework

Type of stance	Target of stance			
	Entities or Topics	Events or Rumors	Fact Triples	Factual claims
Favorability	SemEval-2016; VAST; P-Stance; H-Stance	MGTAB	-	-
Likelihood	-	WT-WT	-	-
Truthfulness	-	PHEME; SemEval-2017; Semeval-2019	NewsClaims	Emergent; FNC-1; COVIDLies; TSD-CT

Truthfulness Stance Detection



Does the tweet believe the claim is true?

A tweet

A claim

Model

Positive ⊕

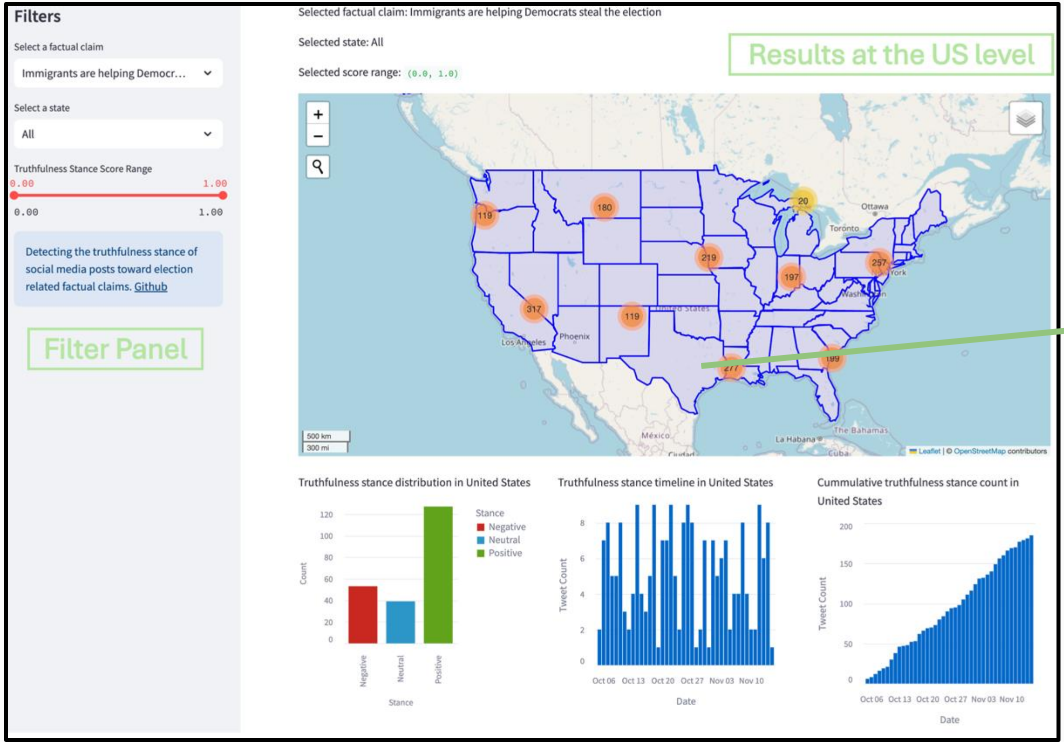
Neutral ⊙

No Stance ⊙

Negative ⊖

Truthfulness Stance Helps Tackle Misinformation

- Gauge public perception toward factual claims
- Comprehend how misinformation spreads, e.g., identify susceptible communities.



Filters

Select a factual claim: Immigrants are helping Democr...

Select a state: All

Truthfulness Stance Score Range: 0.00 – 1.00

Detecting the truthfulness stance of social media posts toward election related factual claims. Github

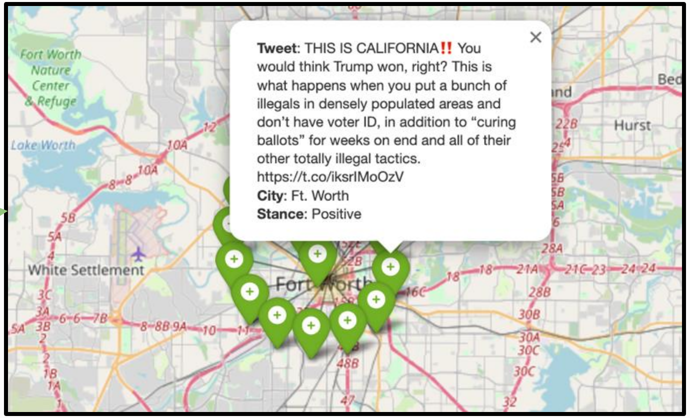
Filter Panel

Selected factual claim: Immigrants are helping Democrats steal the election

Selected state: All

Selected score range: `(0.0, 1.0)`

Results at the US level



Tweet: THIS IS CALIFORNIA‼️ You would think Trump won, right? This is what happens when you put a bunch of illegals in densely populated areas and don't have voter ID, in addition to "curing ballots" for weeks on end and all of their other totally illegal tactics. https://t.co/iksrIMoOzV

City: Ft. Worth

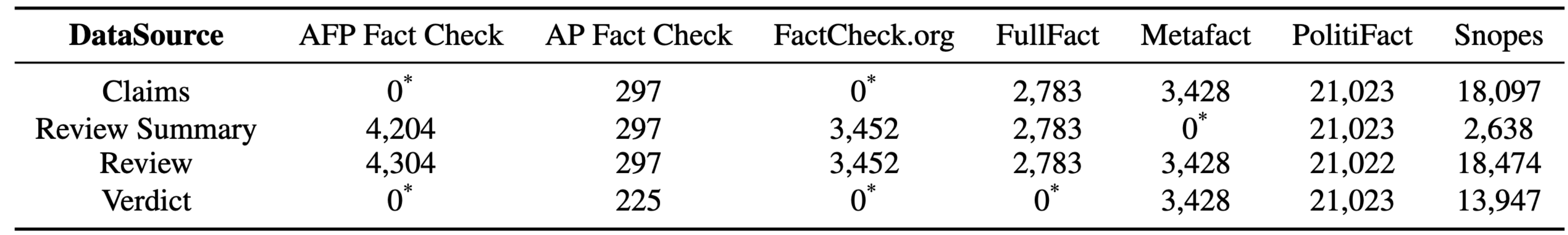
Stance: Positive

Thank you!

Backup Slides: Data Collection

Fact-check collection

- We developed a tool to collect fact-checks from seven well-known fact-checking websites, including AFP Fact Check, AP Fact Check, FactCheck.org, FullFact, Metafact, PolitiFact, and Snopes.



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

Data Collection

Claim-tweet Pair Collection

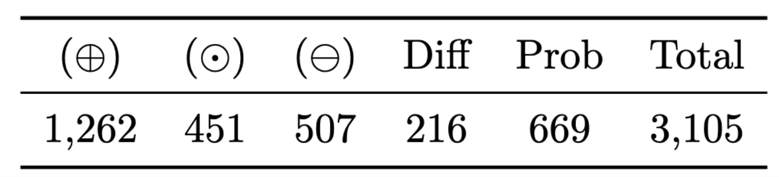
1. We selected factual claims from PolitiFact in the fact-check collection, excluding those phrased as questions.
2. We then use Spacy to extract keywords (nouns, verbs, adjectives, pronouns, and numbers) from the claims.
3. For each factual claim, we retrieved related tweets via Twitter API v2 using a conjunctive (ANDed) query formed by the extracted keywords from the claim.
4. We filtered out tweets with fewer than 30 characters, as well as retweets, replies, and quotes, to avoid duplicates.
5. This process led to 36,154 claim-tweet pairs.

Claim-tweet Pair Annotation

- In-house annotation website
- Annotators:
  - **96** annotators contributed to the annotation.
  - 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:
  - We used 287 carefully selected screening pairs. Each pair received consistent labeling from five researchers.
  - These pairs were mixed with the pairs that needed real annotation.
  - 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.
  - Among all 206 annotators, 30 were deemed high-quality.

TSD-CT Dataset

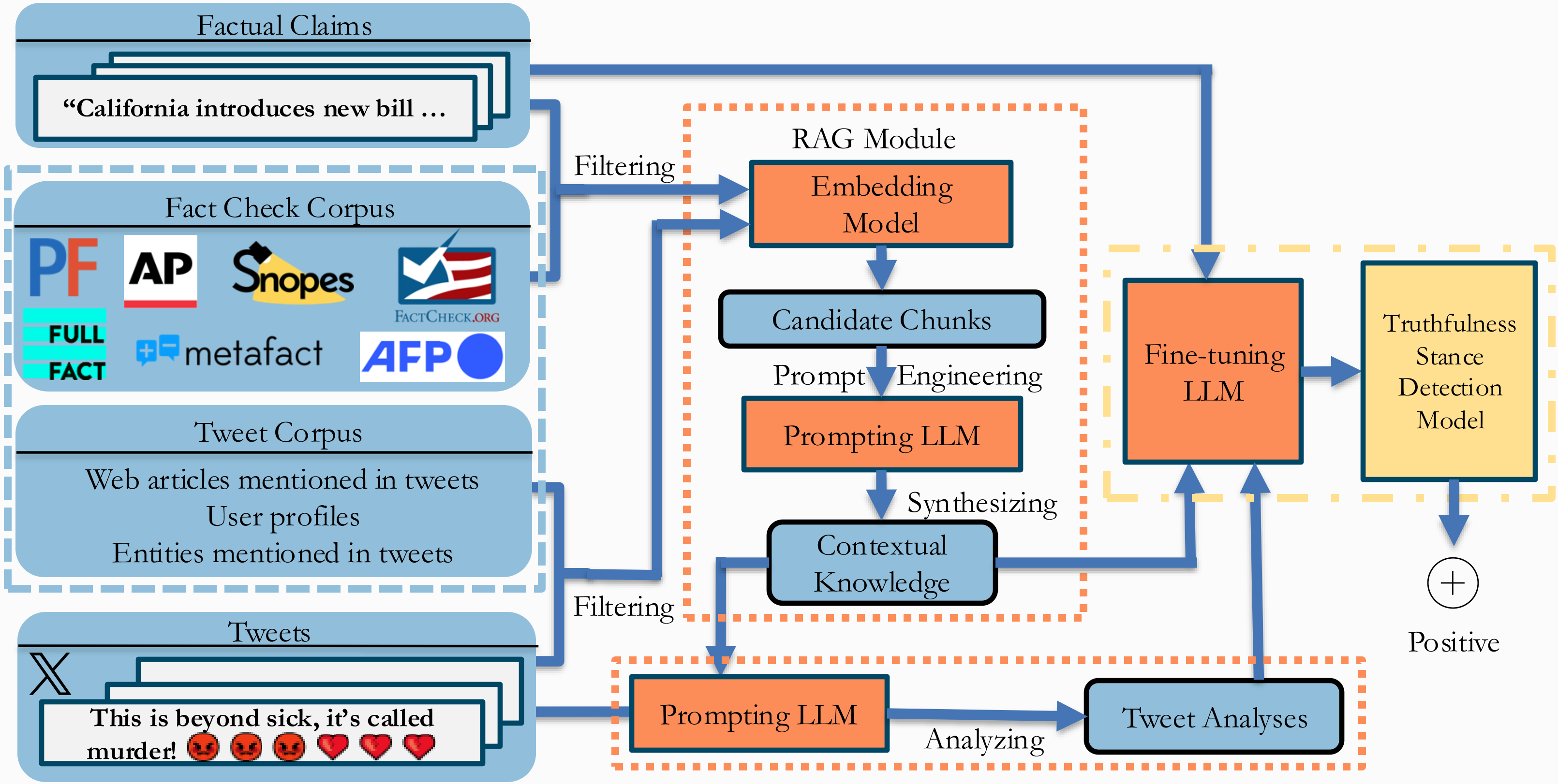
- Truthfulness stance detection for claim-tweet pair(TST-CT) dataset
- A total of 18,584 annotations were collected, with 13,594 from these high-quality annotators.
- This resulted in 3,105 completed pairs, containing 1,520 unique claims.
- Of the completed pairs, 216 were labeled as *different topics* and 669 as *problematic*.



(⊕)	(⊙)	(⊖)	Diff	Prob	Total
1,262	451	507	216	669	3,105

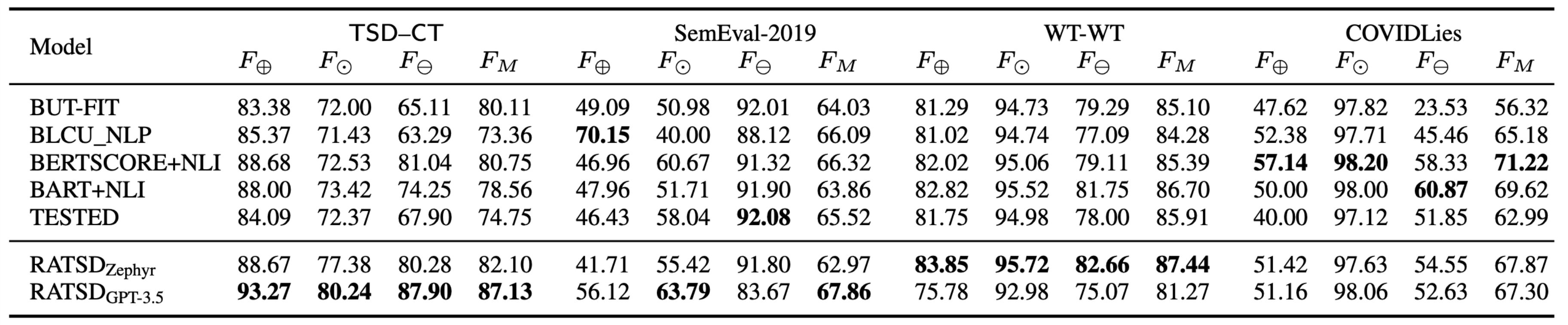
RATSD Framework

Retrieval **A**ugmented **T**ruthfulness **S**tance **D**etection



Evaluation

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), generative pre-trained model (BLCU_NLP), domain-adaptive pre-trained model (BERTSCORE+NLI, BART+NLI, and TESTED).



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