DETECTING MEDICAL MISINFORMATION IN SOCIAL MEDIA

CLAIM VERIFICATION AND TOPIC MODELING

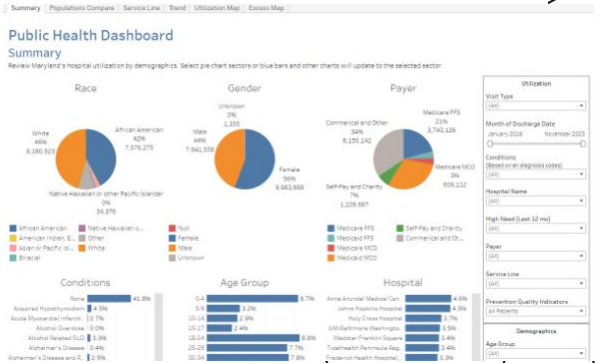
BY: MAHAMADOU DIALLO, SOLOMON GRUSE, ROBERT SIDNEY COX, AND LIA TESTA

A HEALTH CRISIS: VIRAL MEDICAL MISINFORMATION

- Social platforms amplify fake health claims.
- Real world harm: vaccine hesitancy, fatal remedies.

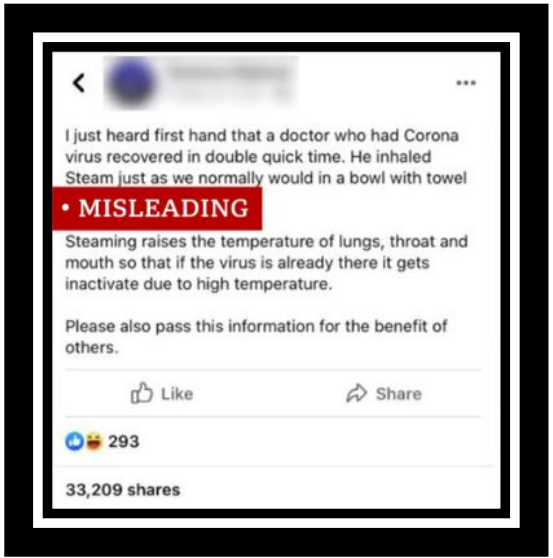
WHY IT MATTERS?

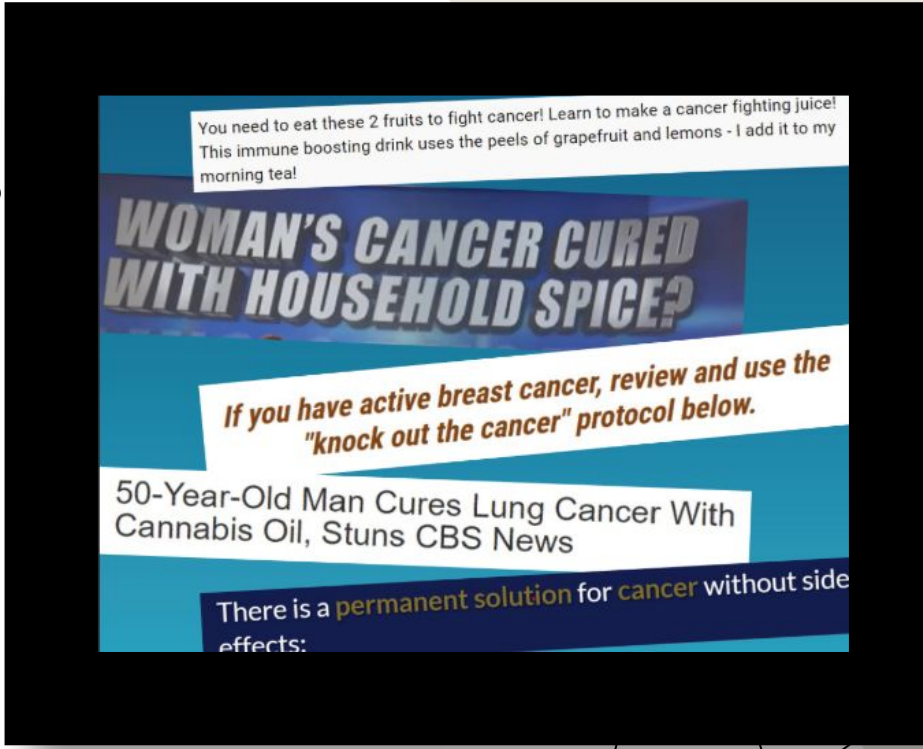
- Tools for public health.
- Assist with content moderation.
- Build educational campaigns.



WHAT DOES MEDICAL MISINFORMATION LOOK LIKE?

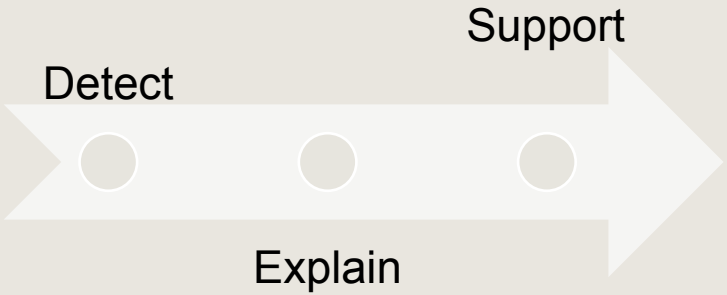
- "COVID Vaccines Cause Infertility"
- "Essential Oils Cure Cancer"
- "Masks Lower Oxygen Levels Dangerously"





OUR GOAL

- Detect and understand medical misinformation.
- Make detection explainable and actionable.



DATA - GENERAL PREPROCESSING:

- Lowercase, remove URLs, punctuation, and HTML
- Cleaned over 4,000 documents
- Example: 'COVID CURE!!!' → 'covid cure'

Modeling Tools

- Baseline: SBERT + cosine similarity to prototype class sentence
- Fine-tuned BERT with
  - (a) LoRA adapter
  - (b) Full fine-tuning
  - (c) PeFT prefix fine-tuning
- HuggingFace Transformers for training

Training Details

- Train/Val Split: 80/20
- CrossEntropy Loss, batch size optimized
- Accuracy, Precision, Recall, F1 evaluated (confusion matrix)

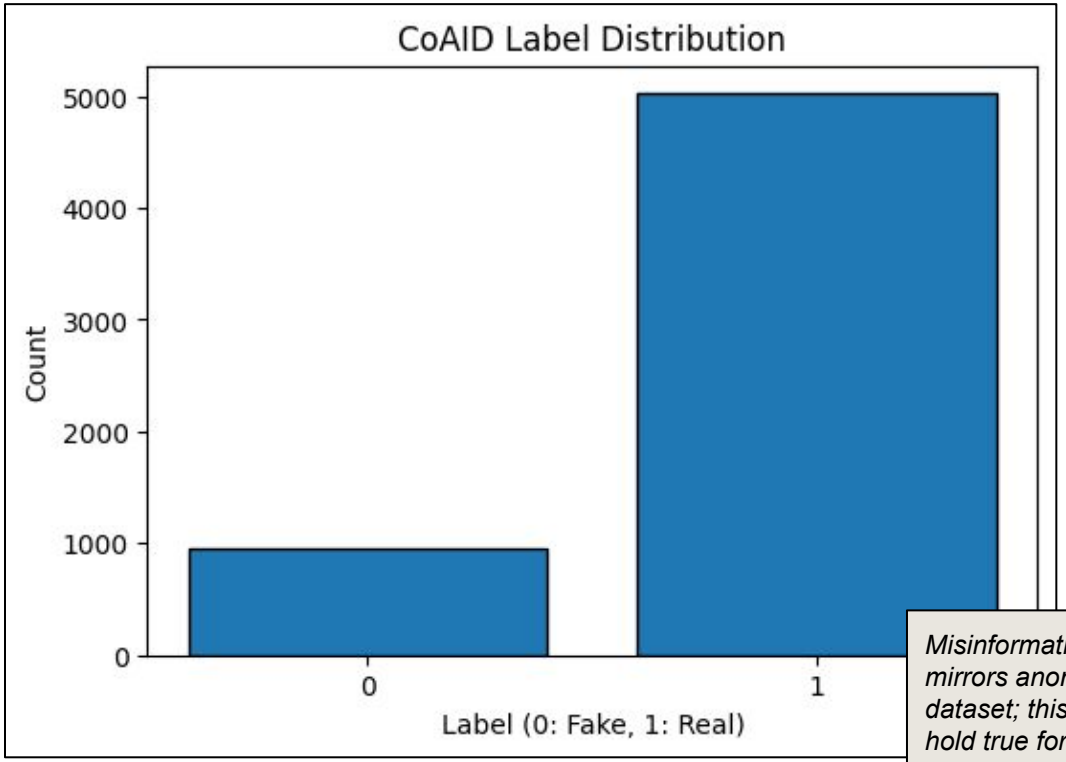
Data Sources

- **CoAID:** Diverse range of COVID-19 healthcare misinformation including fake news articles, social media posts, and user engagement data
  - 4,251 news articles
  - ~296,000 related user engagements
  - 926 social media posts
  - *Ground truth labels included*
- **HealthStory/FakeHealth:** Comprehensive set of labeled social media posts with user engagement data
  - ~1,600 news articles with associated rating by health experts
- **TREC:** General misinformation data labelled by several methods of reliability and accuracy (we only use "helpful" or "unhelpful").
- **Reddit Validation:** Additional testing data For health-related subreddit posts, demonstration of user interaction with our model

CoAID Dataset: COVID-19 Misinformation

- CSV files of fake and authenticated claims
- 3565 entries, but high quality verified labels
- Primary dataset for training/fine tuning misinformation and factual information
- https://github.com/cuilimeng/CoAID

Label Distribution: CoAID

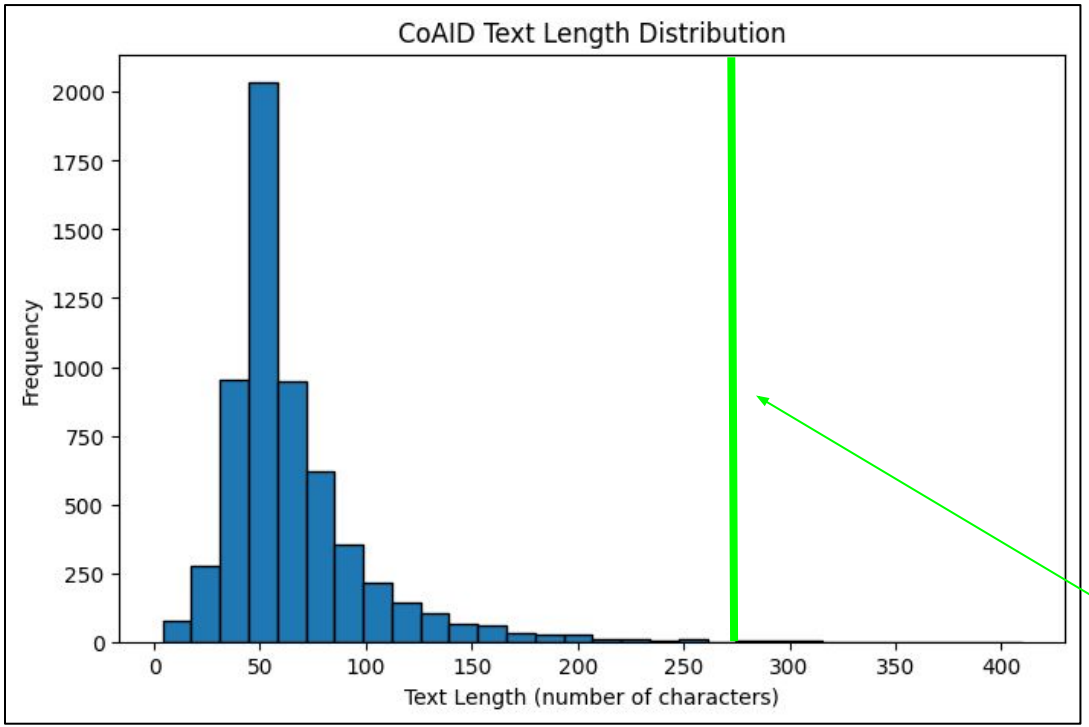


*Class imbalance between **fake** and **real** labels in dataset*

Stratification used when constructing training/testing split

Misinformation classification mirrors anomaly detection for this dataset; this assertion may not hold true for all real-world data

Data Size Distribution: CoAID



Distribution reflects majority of examples being short-form user engagement responses

Common character limit for social media responses (Twitter)

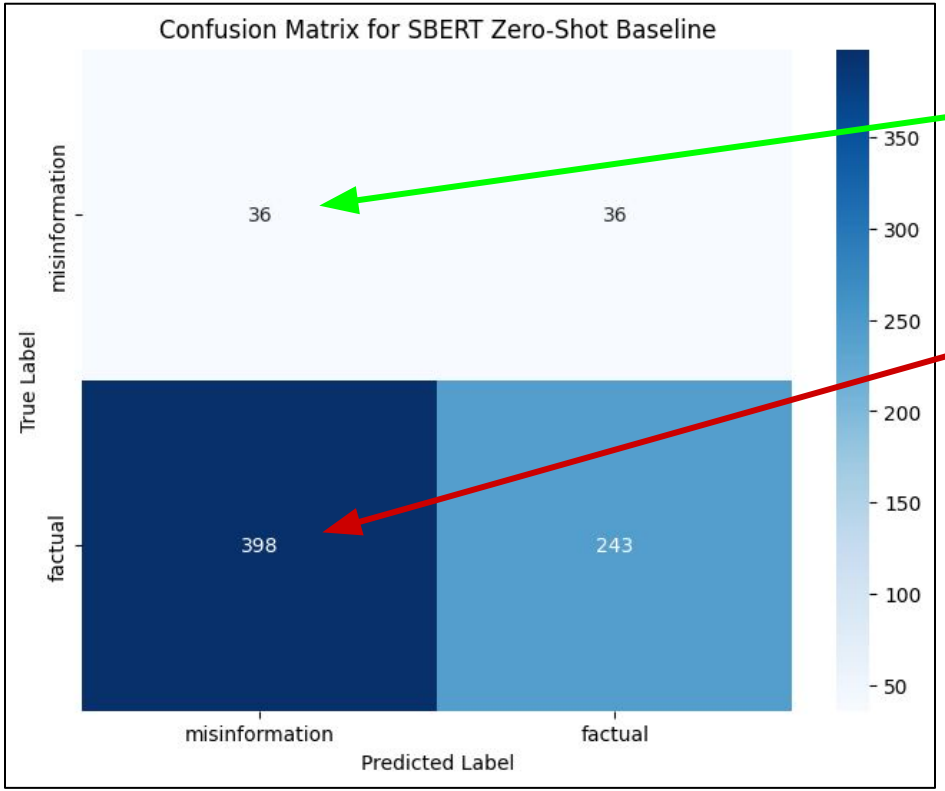
Model Architecture

all-MiniLM-L6-v2/bert-base-uncased*		
Zero-Shot (SBERT)	**Peft (LoRA/Prefix)**	**Full Fine-Tuning (BERT)**
2-class sequence classification using similarity scores Out-of-the-box BERT with no additional training for zero shot evaluation	2-class sequence classification LoRA fine-tuning and prefix finetuning **~300K trainable parameters**	2-class sequence classification Unfroze and Update all weights in BERT model **~100M trainable parameters**

*`all-MiniLM-L6-v2` *and* `bert-base-uncased` *models (available via HuggingFace* `transformers` *and* `sentence-transformers`*) used as starting point for each experiment*

Results

Results: SBERT Zero-Shot Classification (CoAID)

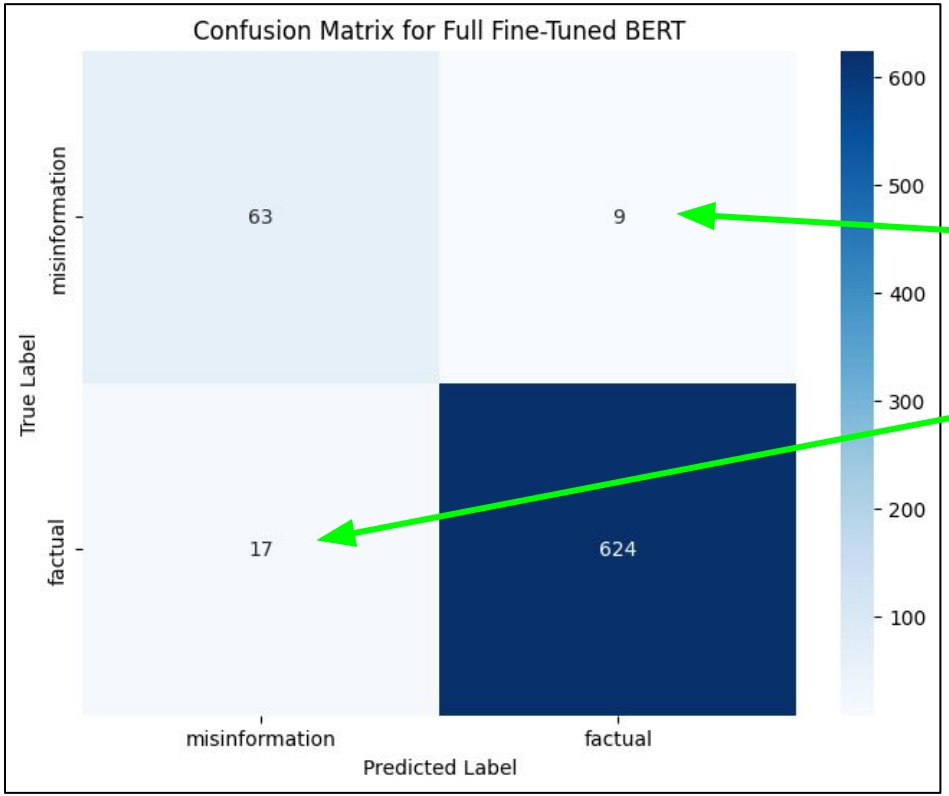


Correctly identifies half of misinformation examples

Tendency to overcall factual news as misinformation

Simply comparing sentences for similarity is not enough; a more focused training approach is needed to create a reliable misinformation detector

Results: Full Fine-Tuning BERT (CoAID)



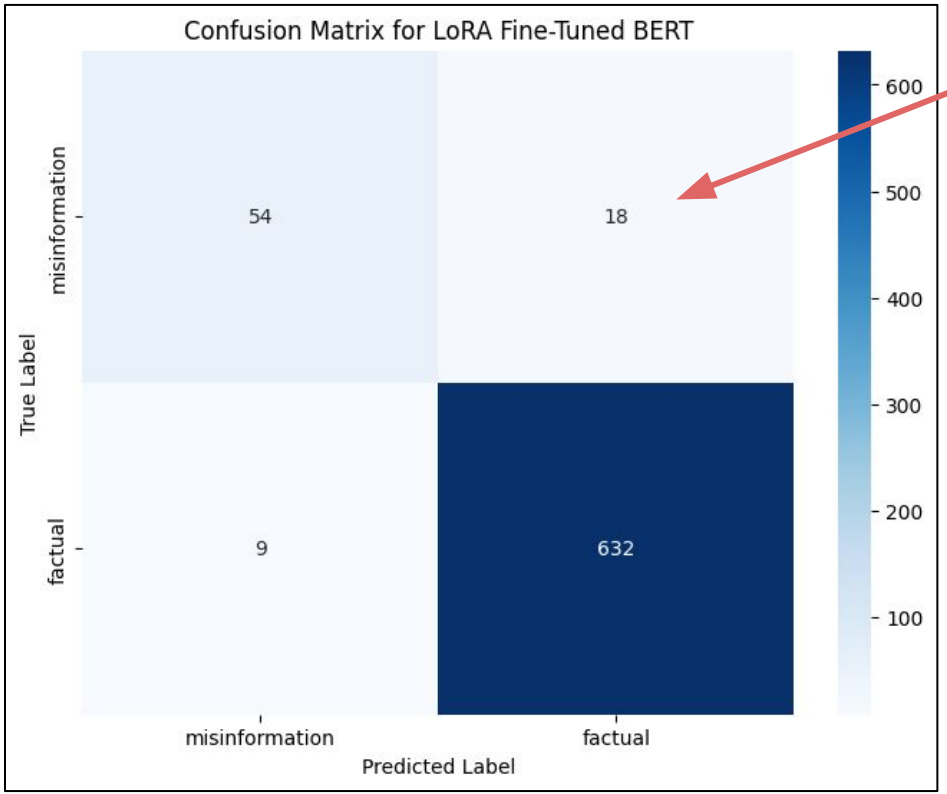
96.4% accuracy on test set

Low FN count

Low FP count

Fine-tuning a pre-trained language model on a domain specific dataset is great for optimal performance on tasks like detecting medical misinformation

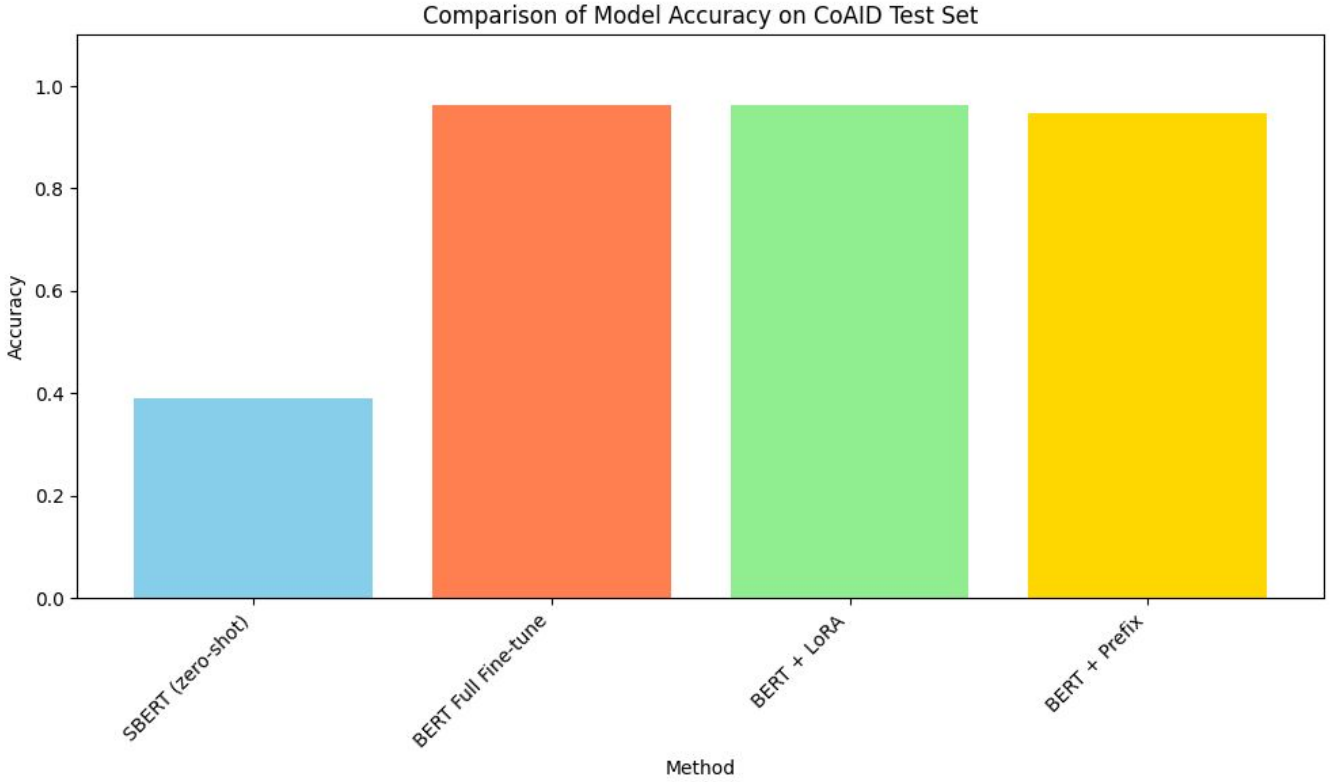
Results: LoRA Fine-Tuned BERT (CoAID)



Larger FN rate than full fine-tuned model

Effectively discriminates between factual and misinformation news, ***while being more computationally efficient during training***

Results: Fine-tuning accuracy (CoAID)



Results: Metrics Overview (CoAID)

Method	Precision (Misinfo.)	Recall (Misinfo.)	F1 (Misinfo.)	Precision (Factual)	Recall (Factual)	F1 (Factual)	F1 (Macro)	Accuracy
SBERT (zero-shot)	0.083	0.500	0.142	0.871	0.379	0.528	0.335	0.391
BERT Full Fine-tune	0.787	0.875	0.829	0.986	0.973	0.980	0.904	0.964
BERT + LoRA	0.857	0.750	0.800	0.972	0.986	0.979	0.890	0.962
BERT + Prefix	0.947	0.500	0.655	0.947	0.997	0.971	0.813	0.947

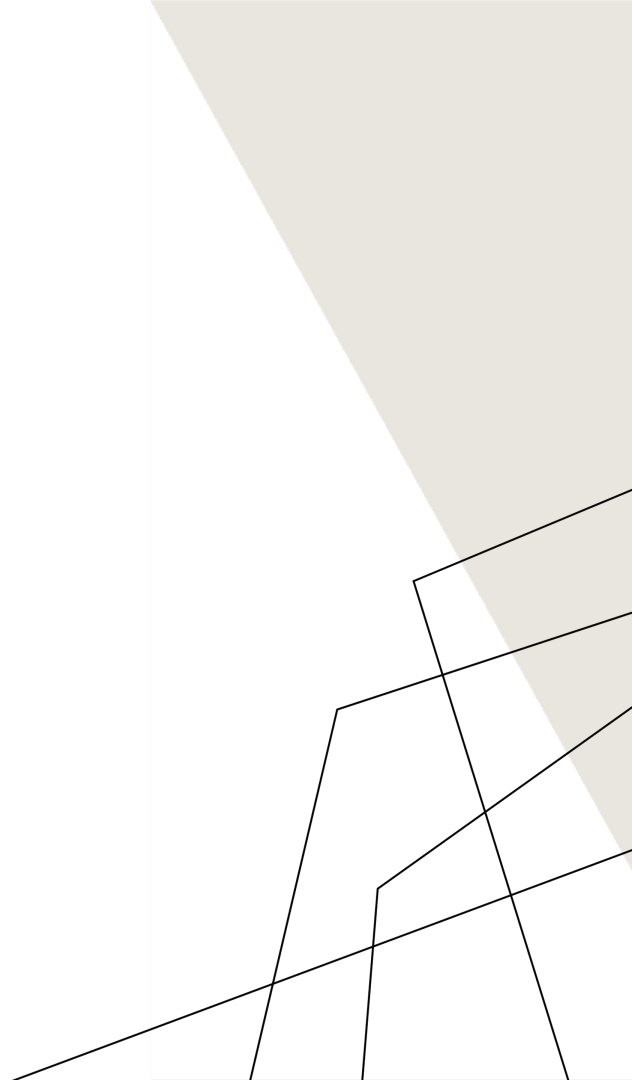
Results: Takeaways

- Fully fine-tuned BERT was the best performer in the raw metrics with the highest F1
- LoRA and prefix tuning not far behind, achieving between 95-98% of the full model's performance
  - Using far fewer trainable weights
- SBERT zero-shot was a useful baseline, and performed poorly at the CoAID classification task
  - Claim-level verification and fine-tuning improve accuracy

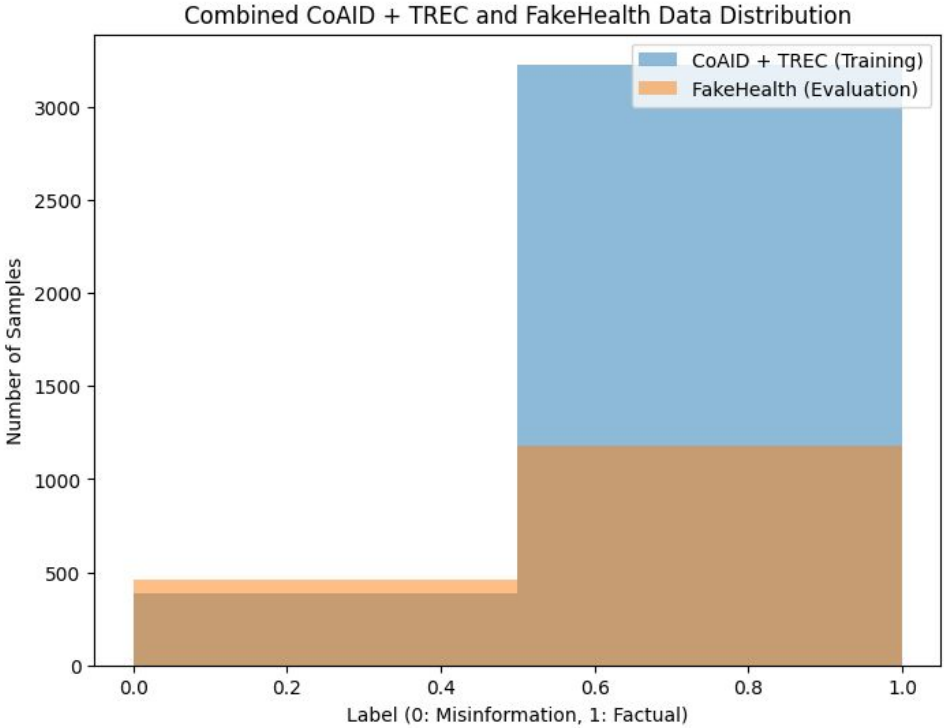
TREC 2021 Health Misinformation

TREC 2021 Data

- Compiled misinformation Data from NIST
- General health misinformation--not limited to Covid-19 related stories
- Highly labelled: "helpful", "harmful", "useful", "correct", etc.
    - We only use whether or not a topic is "helpful" or not
    - "helpful" we take to be equivalent to "factual"
    - not "helpful" we take to be equivalent to "misinformation"
- https://trec-health-misinfo.github.io/
- We use this data as a supplement to the CoAID data to make it more general than just COVID-19 related misinformation



Label distribution



FakeHealth/HealthStory Verification

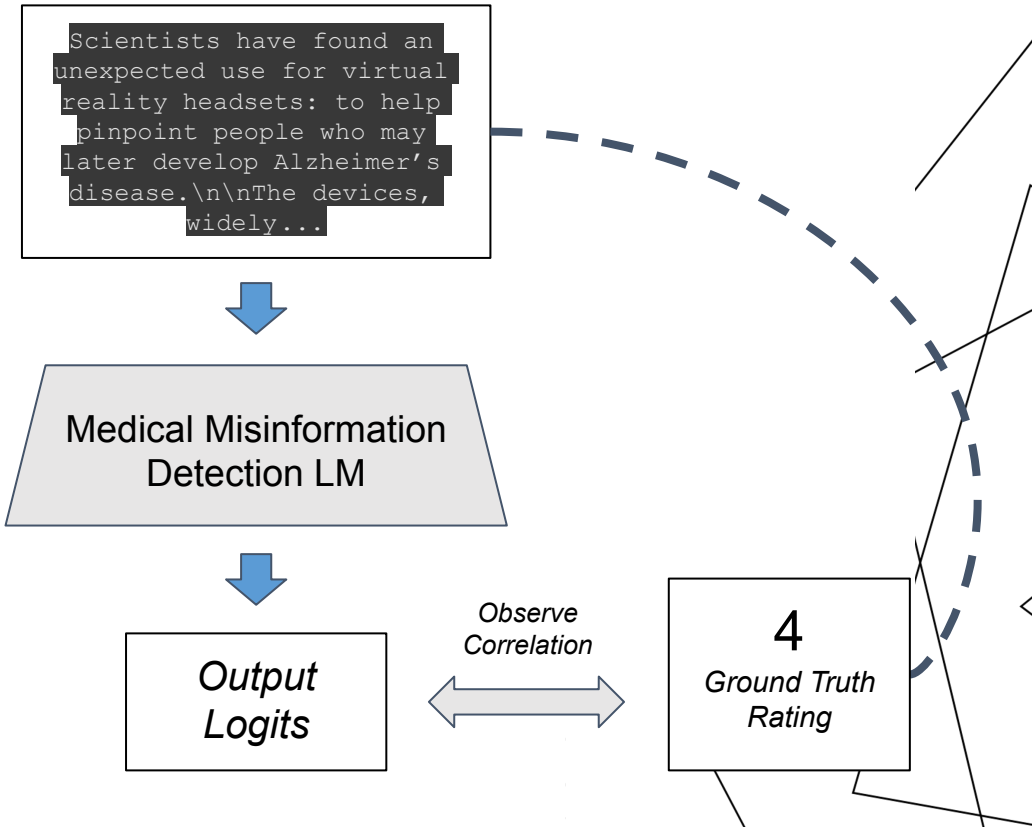
FakeHealth/HealthStory Dataset

- 1638 stories with titles and user reviews
- 5-star scale for user reviews (0-5) with reported average user review for each story
- Not validated as "true" or "false", more of a user generated score of "truthiness"

Adaptation to Article Rating Prediction: Use the CoAID models and compare the model logits to the user score for each story

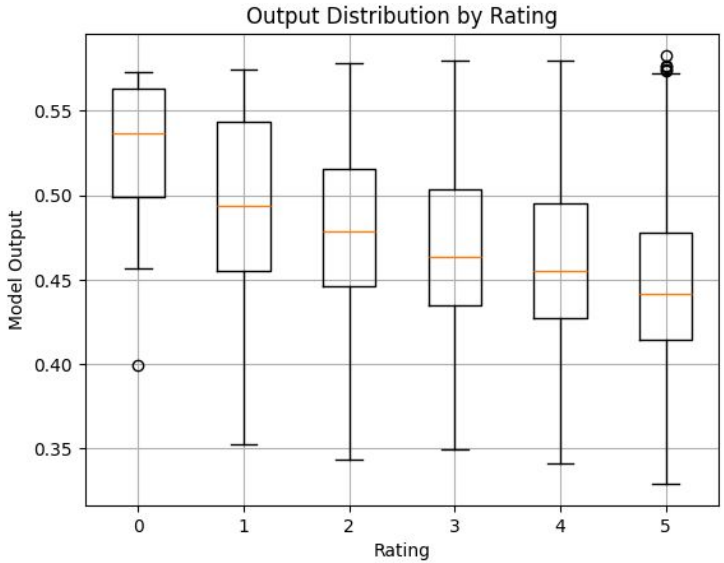
HealthStory dataset containing text articles and associated [0,5] reviews from health experts

Inference on article chunks to determine if correlation exists between ground truth ratings and ***model assessment of article "truthiness"***

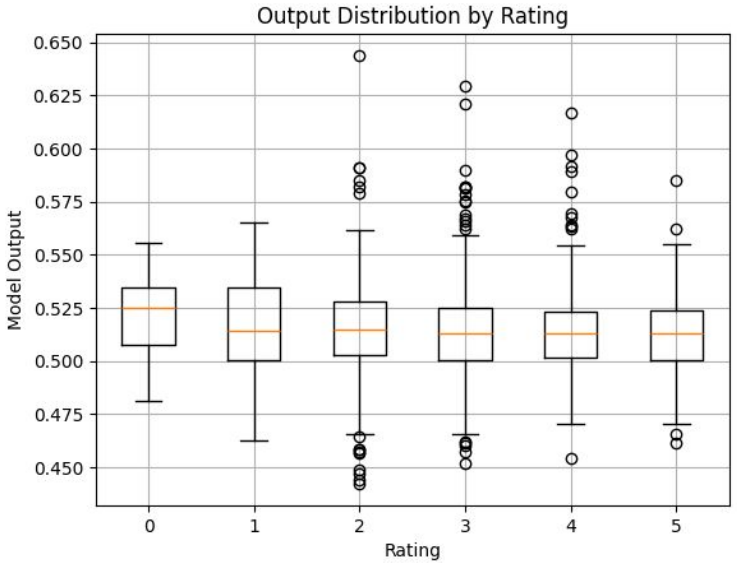


Adaptation to Article Rating Prediction: LoRA Model

LoRA Checkpoint



Final LoRA Model



Average predicted probability of misinformation across all chunks in an article

Adaptation to Article Rating Prediction

- **No obvious correlation between article rating and misinformation prediction**
  - Possibly due to linguistic differences between social media posts and formatted news articles
  - More difficult to accurately measure misinformation across longer text sequences

Reddit Story Testing

Reddit Data test

- Pull data from reddit r/Health
- Use API to get the 50 top discussion titles
- Tokenize and create embeddings for each title using our model

Reddit Data from r/Health

Claim: World may be ‘post-herd immunity’ to measles, top US scientist says
Predicted label: misinformation

Claim: Milk samples collected from over 100 Karachi tea shops found ‘contaminated’
Predicted label: misinformation

Claim: Trump cuts demolish agency focused on toxic chemicals and workplace hazards
Predicted label: misinformation

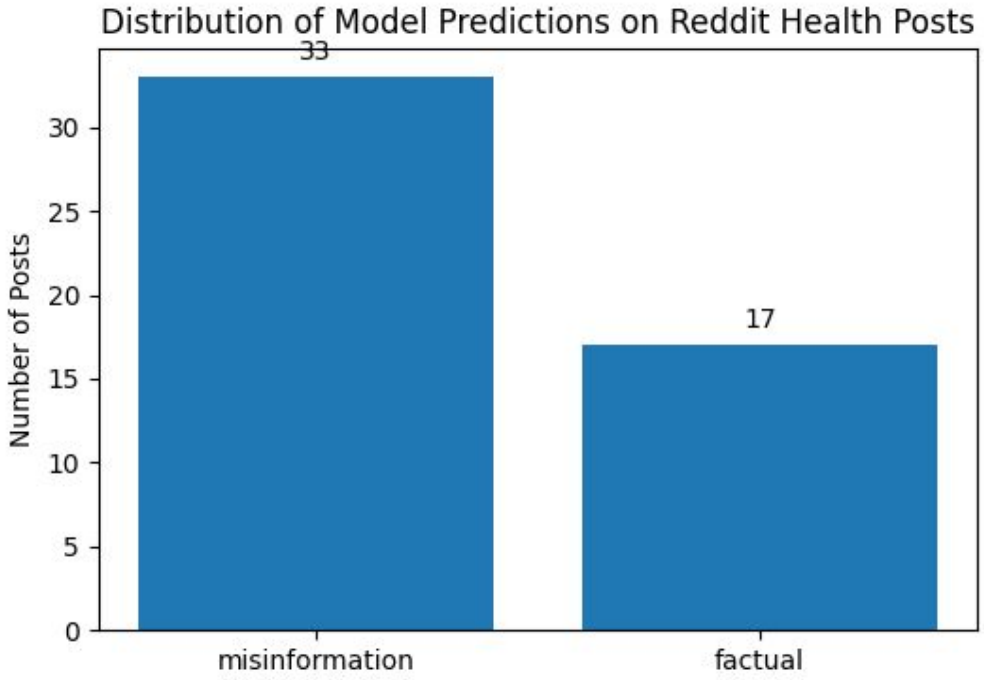
Claim: Stomach cancer cases are rising among younger people, but there is hope for early detection
Predicted label: factual

Claim: First case of measles reported in North Dakota since 2011; "The case involves an unvaccinated child from Williams County who is believed to have contracted the illness from an out-of-state visitor."
Predicted label: misinformation

Claim: I’m an oncologist. Trump’s cuts will devastate cancer research.
Predicted label: misinformation

Demonstration of fully fine-tuned model inference capability on real data

Model prediction distributions



Misinformation Claims

Claim: World may be ‘post-herd immunity’ to measles, top US scientist says

Claim: Milk samples collected from over 100 Karachi tea shops found ‘contaminated’

Claim: Trump cuts demolish agency focused on toxic chemicals and workplace hazards

Claim: First case of measles reported in North Dakota since 2011 ; "The case involves an unvaccinated child from Williams County who is believed to have contracted the illness from an out-of-state visitor."

Claim: I’m an oncologist. Trump’s cuts will devastate cancer research.

Claim: Texas goes after toothpaste in escalating fight over fluoride | Colgate and Crest toothpastes are in the crosshairs.

Claim: Health care company says Trump tariffs will cost it $60M-$70M this year | The health care sector is bracing for higher prices and potential shortages.

Claim: Texas AG opens investigation into toothpaste companies over fluoride exposure. But, dentists say it’s safe

Claim: Wisconsin Man Who’s Spent Years Letting Deadly Snakes Bite Him May Have Unlocked The Ultimate Antivenom

Claim: Education Department stops $1 billion in funding for school mental health

Claim: Thailand reports first anthrax death in decades, hundreds potentially exposed

Claim: Journée Mondiale: The protein order mistake I fixed at meals (9 pounds lost in 30 days)

Claim: Snakes have bitten this man hundreds of times. His blood could help save lives

Claim: Texas attorney general targets toothpaste companies amid increased scrutiny of fluoride

Claim: HHS redirects $500 million to Trump appointee 's vaccine project, bypassing reviews

Claim: Hot Cheetos & Takis: South Texas jailers blame woman' s fatal illness on junk food | Lawsuit alleges jailers ignored woman 's pleas for days before taking her to a hospital.

Claim: Universal vaccines have eluded scientists for years. RFK, Jr., is betting big on the approach.

Claim: Measles jumps borders in North America with outbreaks in Canada, Mexico and U.S.

Claim: NIH cancels participation in Safe to Sleep campaign that decreased infant deaths

Claim: RFK Jr. rejects cornerstone of health science: Germ theory | In his 2021 book vilifying Anthony Fauci, RFK Jr. lays out support for an alternate theory

Claim: We Need to Talk About AI' s Impact on Public Health

Claim: New research contradicts RFK Jr.’s claim that severe autism cases are rising

Claim: RFK Jr.’s HHS Orders Lab Studying Deadly Infectious Diseases to Stop Research

Claim: Why doctors say the viral ‘fart walk’ trend is actually good for you

Claim: Why Ozempic and Wegovy might change your favorite food

Claim Verification

Given a corpus of verification documents, use this as a means of providing evidence for misinformation detection.

Use cosine similarity between SBERT embeddings of evidence corpus and **factual** claim - most similar document is evidence

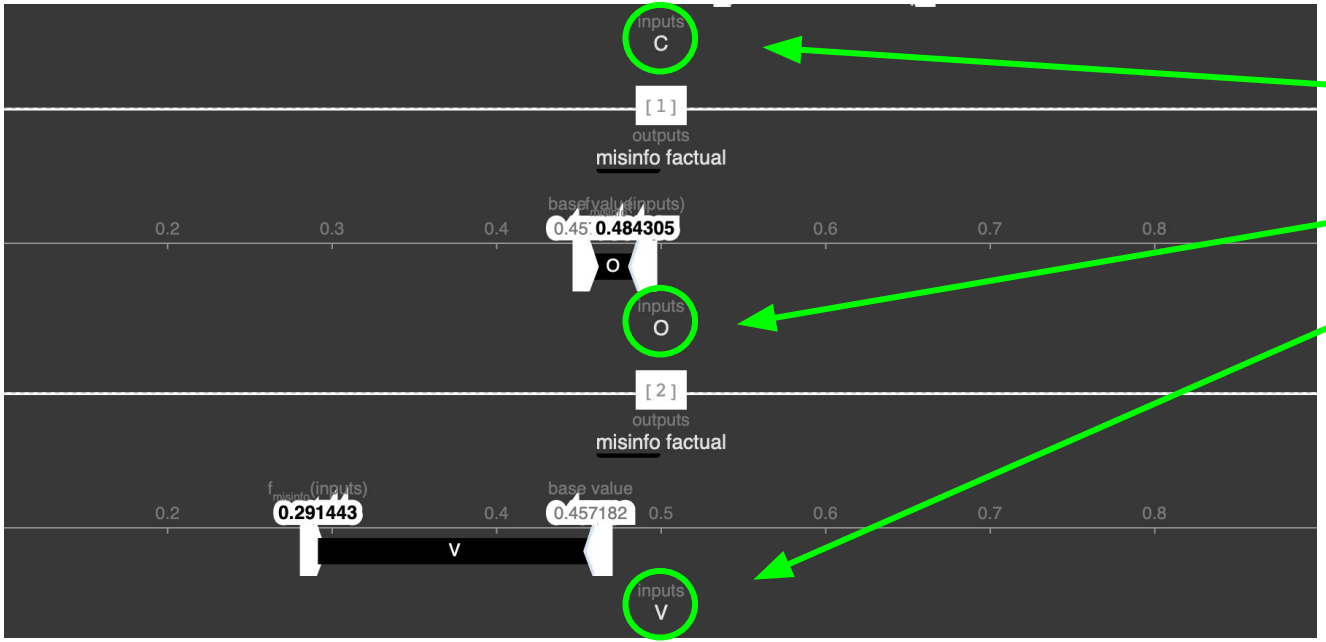
```
Claim: Will wearing an ankle brace help heal achilles tendonitis?
→ Factual
Evidence [FDA]: Ivermectin is not approved to treat COVID-19 in humans.
Sim: 0.09
tensor([[0.1742, 0.1021, 0.1936, 0.3287]], device='cuda:0')

Claim: Is a tepid sponge bath a good way to reduce fever in children?
→ Factual
Evidence [CDC]: Masks reduce spread of respiratory viruses by blocking dr
Sim: 0.33
tensor([[ 0.2304, -0.0122,  0.0611, -0.0064]], device='cuda:0')

Claim: Can folic acid help improve cognition and treat dementia?
→ Factual
Evidence [CDC]: Vaccines do not cause autism; studies find no link.
Sim: 0.23
```

Evidence Corpus	
CDC	"Vaccines do not cause autism; studies find no link."
WHO	"5G networks do not spread COVID-19; viruses can't travel on radio waves."
FDA	"Ivermectin is not approved to treat COVID-19 in humans."
CDC	"Masks reduce spread of respiratory viruses by blocking droplets."

Explainability Analysis

Explainability Analysis with SHAP

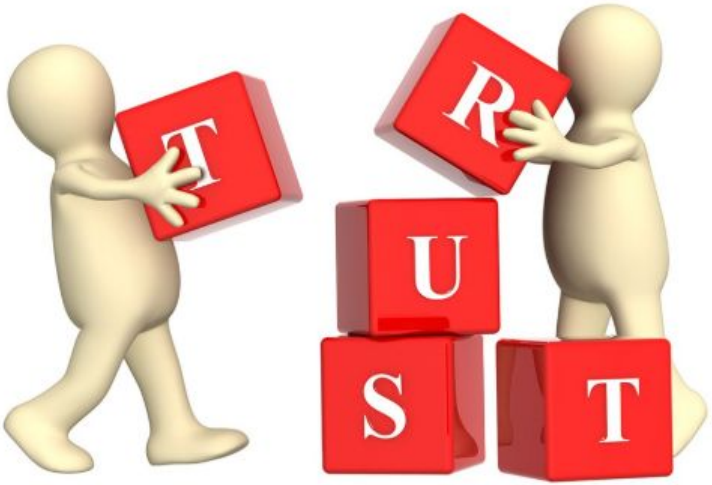
Claim: COVID-19 vaccines completely eliminate the risk of infection.



Token-level explanation of misinformation detection for a given claim

SUMMARY

- Modular, explainable system.
- Tackling a growing public health threat.



REFERENCES

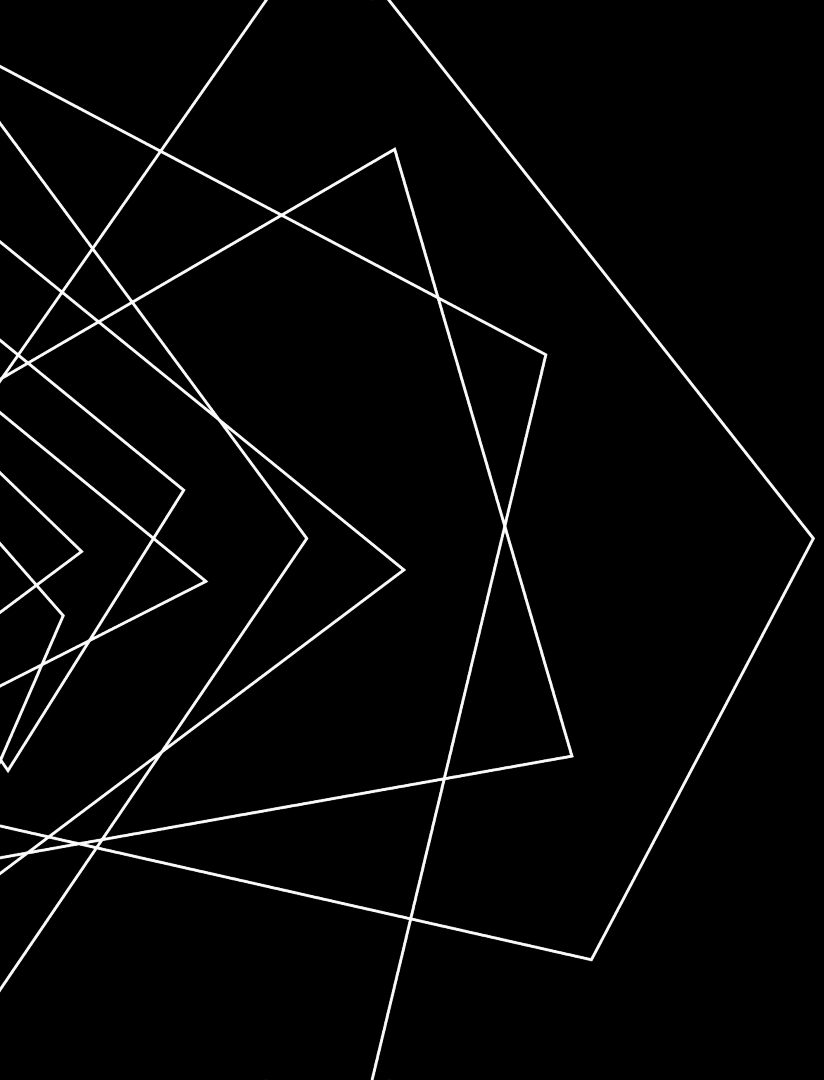
Di Sotto, Stefano, and Marco Viviani. 2022. “Health Misinformation Detection in the Social Web: An Overview and a Data Science Approach.” *International Journal of Environmental Research and Public Health* 19 (4): 2173.

Islam, Md Rafiqul, Shaowu Liu, Xianzhi Wang, and Guandong Xu. 2020. “Deep Learning for Misinformation Detection on Online Social Networks: A Survey and New Perspectives.” *Social Network Analysis and Mining* 10 (1): 82.

Su, Qi, Mingyu Wan, Xiaoqian Liu, and Chu-Ren Huang. 2020. “Motivations, Methods and Metrics of Misinformation Detection: An NLP Perspective.” *Natural Language Processing Research* 1 (1-2): 1.

Valdez, Danny, Arthur D. Soto-Vásquez, and María S. Montenegro. 2023. “Geospatial Vaccine Misinformation Risk on Social Media: Online Insights from an English/Spanish Natural Language Processing (NLP) Analysis of Vaccine-Related Tweets.” *Social Science & Medicine (1982)* 339 (116365): 116365.

Wang, Jun, Xiulai Wang, and Airong Yu. 2025. “Tackling Misinformation in Mobile Social Networks a BERT-LSTM Approach for Enhancing Digital Literacy.” *Scientific Reports* 15 (1): 1118.



THANK YOU