Using Persuasive Writing Strategies to Explain and Detect Health Misinformation

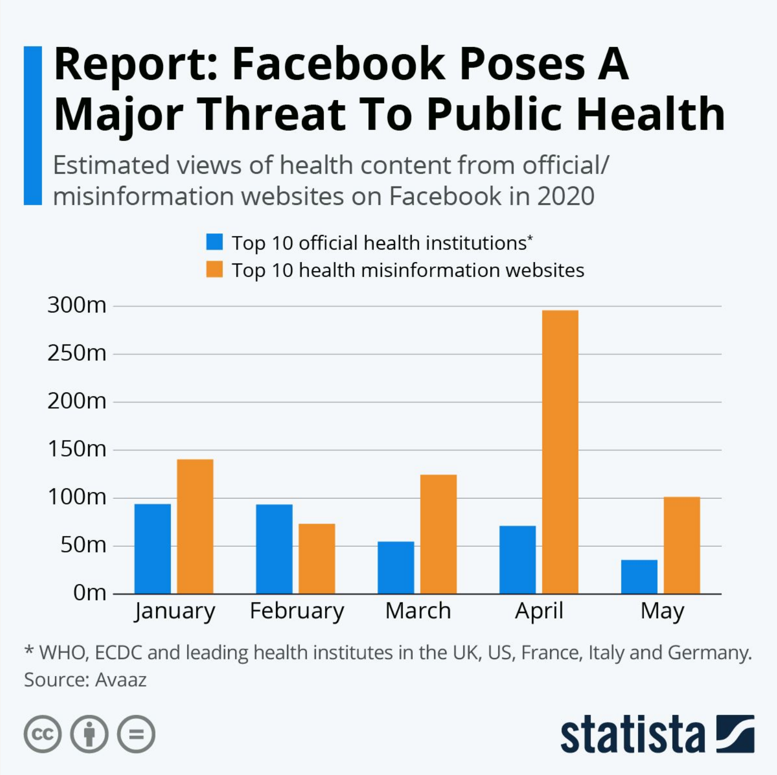
Danial Kamali[*], Joseph Romain[*], Huiyi Liu[†], Wei Peng[‡], Jingbo Meng[§], Parisa Kordjamshidi[*]

*Department of Computer Science and Engineering, Michigan State University,
†Department of Communication, Michigan State University,
‡Department of Media and Information, Michigan State University,
§School of Communication, Ohio State University
{kamalida, romainj2, liuhui5, pengwei}@msu.edu,
meng.28@osu.edu, kordjams@msu.edu

Why?

Health Misinformation

- Societal impacts of health misinformation.
    - Undermining trust in medical advisories.
    - Dire consequences.
- Experts are costly.



Report: Facebook Poses A Major Threat To Public Health

Estimated views of health content from official/ misinformation websites on Facebook in 2020

* WHO, ECDC and leading health institutes in the UK, US, France, Italy and Germany.

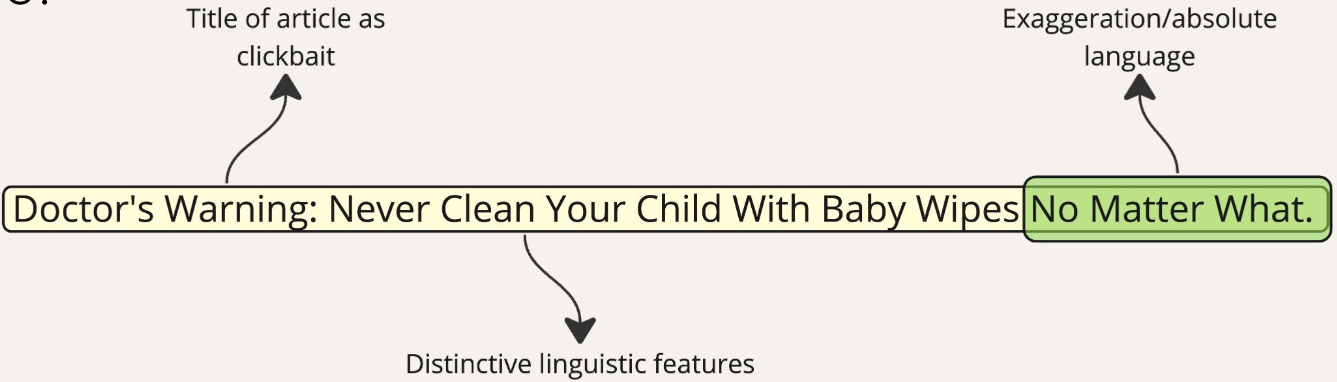
Source: Avaaz

https://www.statista.com/chart/22660/health-misinformation-on-facebook/

What

Automatic Misinformation Detection

Is this true?



Title of article as clickbait

Exaggeration/absolute language

Doctor's Warning: Never Clean Your Child With Baby Wipes No Matter What.

Distinctive linguistic features

Accuracy

Transparency

Research questions

- Can language models detect misinformation?
- Can persuasive writing strategies help LLMs in misinformation detection?
- Can language models identify persuasive strategies?

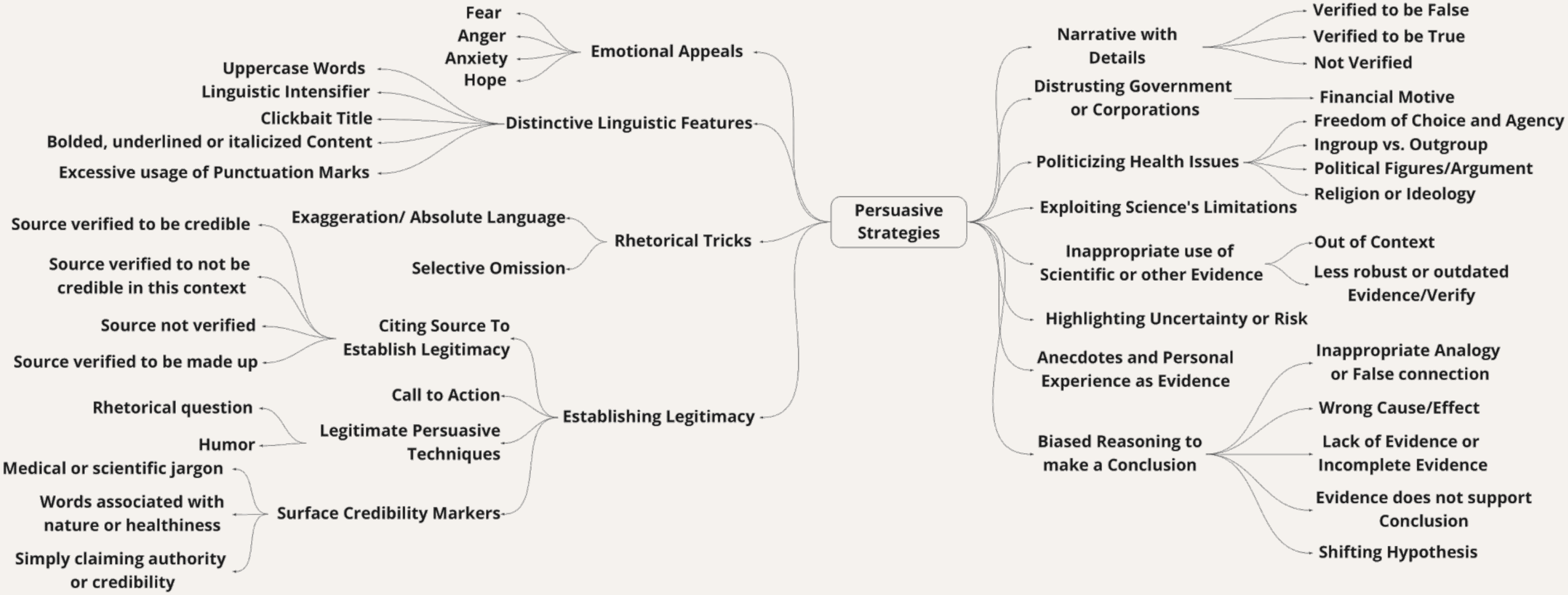
No resource

New Dataset

Persuasive Writing Strategy Dataset

- **Source**
  - A health-related subset of MultiFC.
  - Filtered 599 claims to to 242 claims.
- **Article Retrieval**
  - Manually retrieved full articles associated with each claim.
  - Articles were gathered from publicly accessible domains.
- **Annotation Process**
  - Developed a novel annotation scheme for persuasive writing strategies.
  - Annotated by three communication experts in studies.

Annotation scheme



- Persuasive Strategies
  - Emotional Appeals
    - Fear
    - Anger
    - Anxiety
    - Hope
  - Distinctive Linguistic Features
    - Uppercase Words
    - Linguistic Intensifier
    - Clickbait Title
    - Bolded, underlined or italicized Content
    - Excessive usage of Punctuation Marks
  - Rhetorical Tricks
    - Exaggeration/ Absolute Language
    - Selective Omission
  - Establishing Legitimacy
    - Citing Source To Establish Legitimacy
      - Source verified to be credible
      - Source verified to not be credible in this context
      - Source not verified
      - Source verified to be made up
    - Call to Action
    - Legitimate Persuasive Techniques
      - Rhetorical question
      - Humor
    - Surface Credibility Markers
      - Medical or scientific jargon
      - Words associated with nature or healthiness
      - Simply claiming authority or credibility
  - Narrative with Details
    - Verified to be False
    - Verified to be True
    - Not Verified
  - Distrusting Government or Corporations
    - Financial Motive
  - Politicizing Health Issues
    - Freedom of Choice and Agency
    - Ingroup vs. Outgroup
    - Political Figures/Argument
    - Religion or Ideology
  - Exploiting Science's Limitations
  - Inappropriate use of Scientific or other Evidence
    - Out of Context
    - Less robust or outdated Evidence/Verify
  - Highlighting Uncertainty or Risk
  - Anecdotes and Personal Experience as Evidence
  - Biased Reasoning to make a Conclusion
    - Inappropriate Analogy or False connection
    - Wrong Cause/Effect
    - Lack of Evidence or Incomplete Evidence
    - Evidence does not support Conclusion
    - Shifting Hypothesis

Dataset Details

Persuasive Writing Strategy Dataset

- **Quality Control**
    - Inter-annotator reliability:
        - Cohen's Kappa.
        - 88% average pairwise agreement
- **Final Dataset**
    - **242** claim and articles with veracity (binary) labels.
    - **5,666** sentences, each can be annotated with one or more strategy.

Task 1: Misinformation Detection

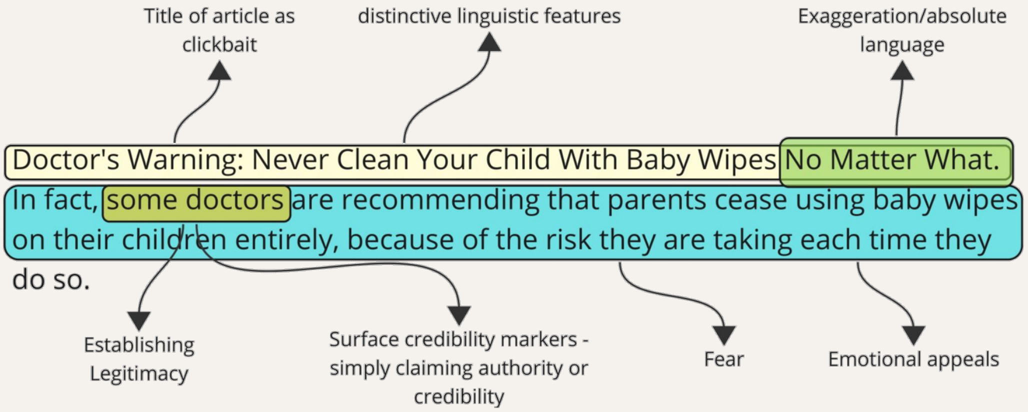
- **Objective:** Classify misinformation news using their claim and articles.
- **Challenge:**
    - Involves analyzing content veracity, context, and intent.

Is this true?

Doctor's Warning: Never Clean Your Child With Baby Wipes No Matter What. ✗

Task 2: Persuasive Strategy Labeling

- **Objective:** Identify and categorize text spans with persuasive strategy writing.
- **Challenge:** Requires in-depth linguistic analysis.
- **Settings:**
  - Unit:
    - Sentence
    - Article
    - Span
  - Information context:
    - No context.
    - Low: 1 unit from each side.
    - High: 2 unit from each side.


Title of article as clickbait
distinctive linguistic features
Exaggeration/absolute language
Establishing Legitimacy
Surface credibility markers - simply claiming authority or credibility
Fear
Emotional appeals

Doctor's Warning: Never Clean Your Child With Baby Wipes No Matter What.
In fact, some doctors are recommending that parents cease using baby wipes on their children entirely, because of the risk they are taking each time they do so.

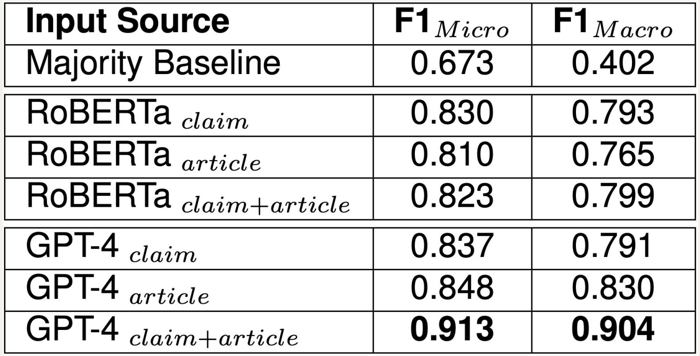
Task 3: Combined

- **Objective:** Combine the detection of misinformation with the analysis of persuasive writing strategies.
- **Challenge:** Inherits all!
- **Settings:** Multi-task Learning or Pipeline setting.

Experiments

1. Misinformation Detection

- Setting
  - Different input variations
    - claim, article, and their combination.
- Models Tested
  - RoBERTa: Fine-tuned on the train data.
  - GPT-4: Evaluated in zero-shot setting.
- Observation
  - Superior performance of GPT-4 over RoBERTa
  - GPT-4 improvement with combined inputs.



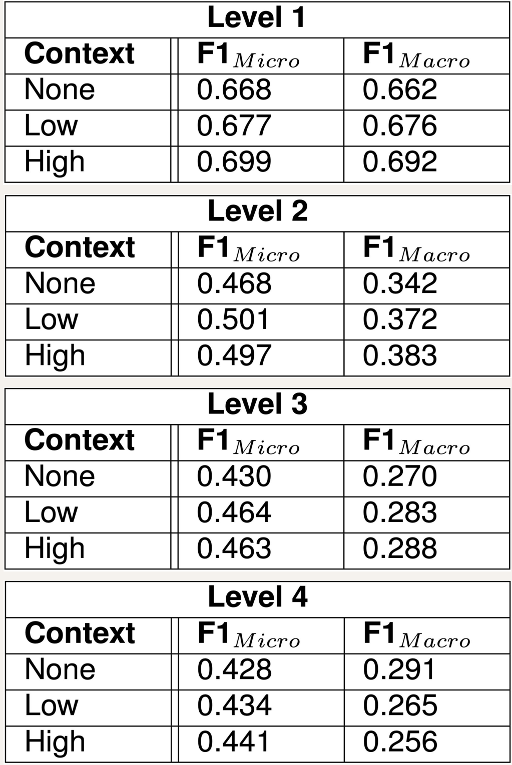
Input Source	**F1** $_{Micro}$	**F1** $_{Macro}$
Majority Baseline	0.673	0.402
RoBERTa $_{claim}$	0.830	0.793
RoBERTa $_{article}$	0.810	0.765
RoBERTa $_{claim+article}$	0.823	0.799
GPT-4 $_{claim}$	0.837	0.791
GPT-4 $_{article}$	0.848	0.830
GPT-4 $_{claim+article}$	**0.913**	**0.904**

Results with different inputs

Experiments

2. Persuasive Writing Strategy Detection

- Sentence Level
- Tested fine-tuned RoBERTa
- F1-Macro is a more important metric.
- Results
    - Increasing context generally improves detection accuracy.
    - Level-4 only doesn't increase as context grows
        - Not related to context
            - *Scientific jargon*
            - *Words associated with nature*
    - The low results demonstrate the complexity of this task.



Level 1

Context	F1$_{Micro}$	F1$_{Macro}$
None	0.668	0.662
Low	0.677	0.676
High	0.699	0.692

Level 2

Context	F1$_{Micro}$	F1$_{Macro}$
None	0.468	0.342
Low	0.501	0.372
High	0.497	0.383

Level 3

Context	F1$_{Micro}$	F1$_{Macro}$
None	0.430	0.270
Low	0.464	0.283
High	0.463	0.288

Level 4

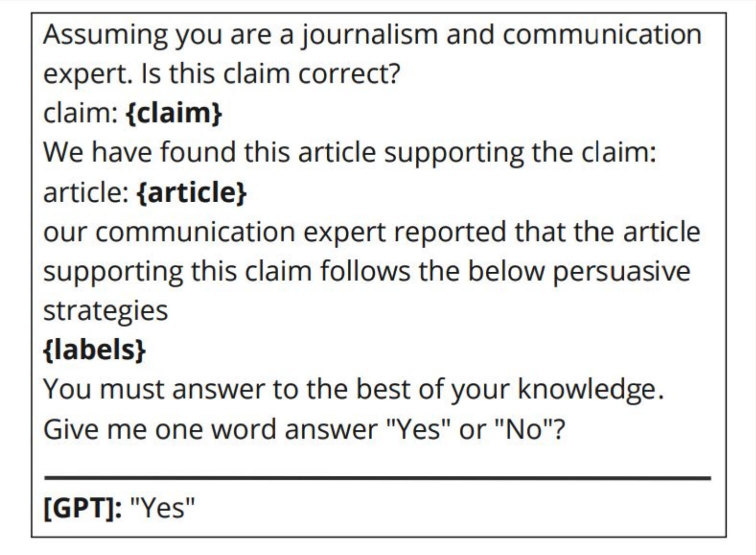
Context	F1$_{Micro}$	F1$_{Macro}$
None	0.428	0.291
Low	0.434	0.265
High	0.441	0.256

Results across different context size and layers

Experiments

3. Misinformation Detection with Persuasive Strategy Detection

- Used pipeline results of RoBERTa.
- Models:
    - RoBERTa
        - Fine-tuned
    - GPT-4
        - Zero-shot
        - In-context Learning



Assuming you are a journalism and communication expert. Is this claim correct?
claim: **{claim}**
We have found this article supporting the claim:
article: **{article}**
our communication expert reported that the article supporting this claim follows the below persuasive strategies
{labels}
You must answer to the best of your knowledge.
Give me one word answer "Yes" or "No"?

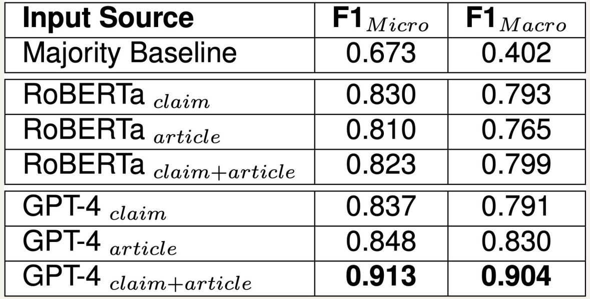
[GPT]: "Yes"

GPT-4 Prompt Template.

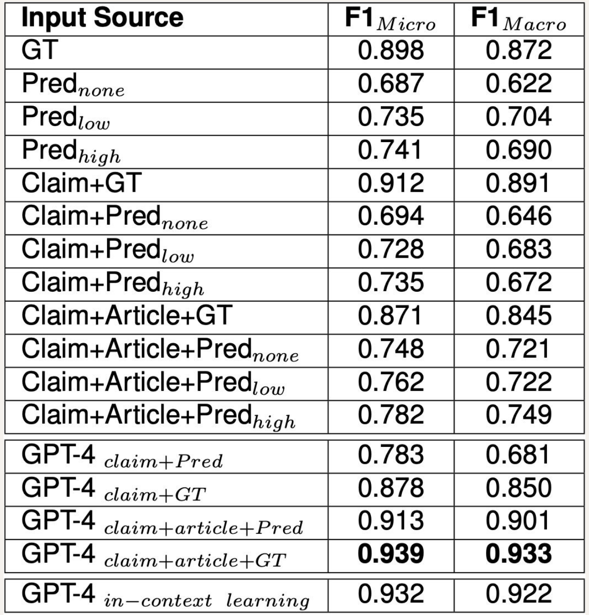
Experiments

3. Misinformation Detection with Persuasive Strategy Detection

- Results
    - Significant improvement with GT persuasive labels.
    - GT alone performs better than claim itself.
        - It has useful information.
    - Claim + Article + GT
        - GPT-4 Performs best.
        - RoBERTa Fails due to token limitation.
    - **In context learning**
        - Outperforms RoBERTa predictions
        - Close to ground-truth labels



Input Source	**F1**$_{Micro}$	**F1**$_{Macro}$
Majority Baseline	0.673	0.402
RoBERTa $_{claim}$	0.830	0.793
RoBERTa $_{article}$	0.810	0.765
RoBERTa $_{claim+article}$	0.823	0.799
GPT-4 $_{claim}$	0.837	0.791
GPT-4 $_{article}$	0.848	0.830
GPT-4 $_{claim+article}$	**0.913**	**0.904**

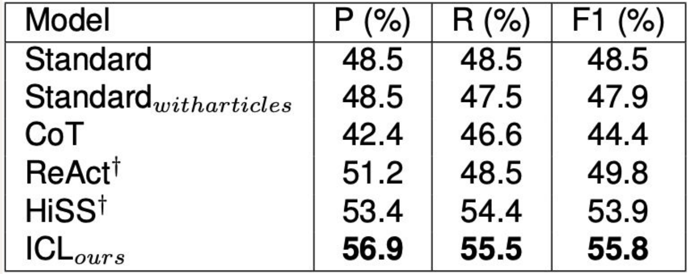


Input Source	**F1**$_{Micro}$	**F1**$_{Macro}$
GT	0.898	0.872
Pred$_{none}$	0.687	0.622
Pred$_{low}$	0.735	0.704
Pred$_{high}$	0.741	0.690
Claim+GT	0.912	0.891
Claim+Pred$_{none}$	0.694	0.646
Claim+Pred$_{low}$	0.728	0.683
Claim+Pred$_{high}$	0.735	0.672
Claim+Article+GT	0.871	0.845
Claim+Article+Pred$_{none}$	0.748	0.721
Claim+Article+Pred$_{low}$	0.762	0.722
Claim+Article+Pred$_{high}$	0.782	0.749
GPT-4 $_{claim+Pred}$	0.783	0.681
GPT-4 $_{claim+GT}$	0.878	0.850
GPT-4 $_{claim+article+Pred}$	0.913	0.901
GPT-4 $_{claim+article+GT}$	**0.939**	**0.933**
GPT-4 $_{in-context\ \ learning}$	0.932	0.922

Experiments

3. Misinformation Detection with Persuasive Strategy Detection

- Experiment on additional benchmarks
  - RAWFC dataset
    - Claim verification with supporting articles
  - Shows **real-world** capability
  - Not domain specific.
  - Achieved **SOTA** on the RAWFC dataset.
    - Without search engine access.

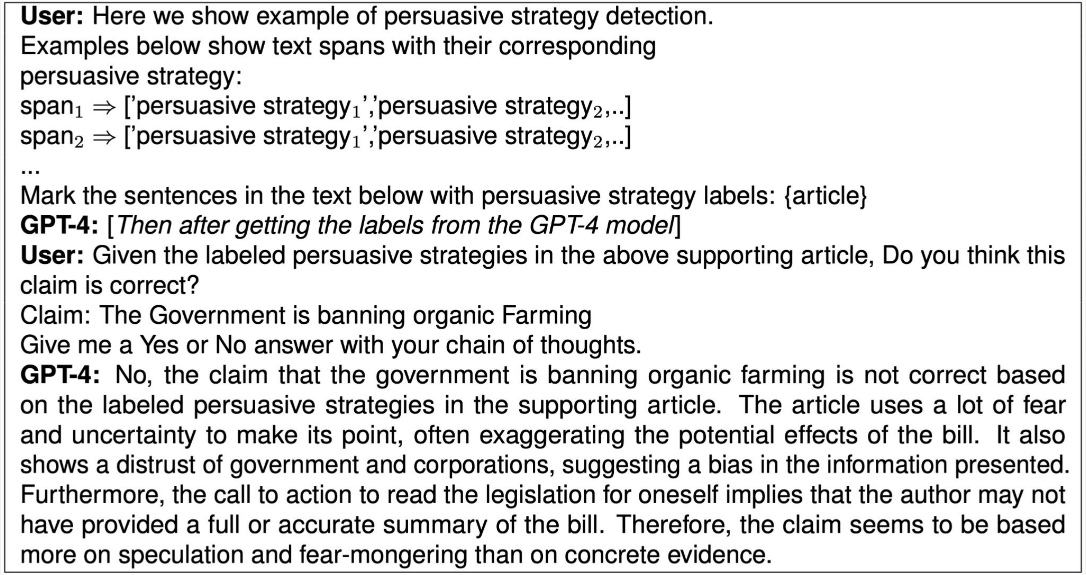


Model	P (%)	R (%)	F1 (%)
Standard	48.5	48.5	48.5
Standard$_{witharticles}$	48.5	47.5	47.9
CoT	42.4	46.6	44.4
ReAct†	51.2	48.5	49.8
HiSS†	53.4	54.4	53.9
ICL$_{ours}$	**56.9**	**55.5**	**55.8**

Results on RAWFC benchmark.

Explainability

Persuasive writing strategies as reasoning in **Chain of Thoughts** prompting can generate explanation for the veracity labels.



User: Here we show example of persuasive strategy detection.
Examples below show text spans with their corresponding persuasive strategy:
span$_1$ ⇒ ['persuasive strategy$_1$','persuasive strategy$_2$,..]
span$_2$ ⇒ ['persuasive strategy$_1$','persuasive strategy$_2$,..]
...
Mark the sentences in the text below with persuasive strategy labels: {article}
GPT-4: [*Then after getting the labels from the GPT-4 model*]
User: Given the labeled persuasive strategies in the above supporting article, Do you think this claim is correct?
Claim: The Government is banning organic Farming
Give me a Yes or No answer with your chain of thoughts.
GPT-4: No, the claim that the government is banning organic farming is not correct based on the labeled persuasive strategies in the supporting article. The article uses a lot of fear and uncertainty to make its point, often exaggerating the potential effects of the bill. It also shows a distrust of government and corporations, suggesting a bias in the information presented. Furthermore, the call to action to read the legislation for oneself implies that the author may not have provided a full or accurate summary of the bill. Therefore, the claim seems to be based more on speculation and fear-mongering than on concrete evidence.

Example of a explanation generated by GPT-4

Conclusion

- Key Contributions:
    - A **novel annotation scheme** for persuasive writing strategies.
    - Utilizing these strategies to enhance the detection of health misinformation.
- Significance:
    - Improved **accuracy** and **explainability** of misinformation detection models.
    - Potential to aid in combating health misinformation..
- Closing Remark:
    - "Our approach paves the way for more nuanced and effective tools in the fight against misinformation, contributing to a healthier, better-informed society."

Thanks!

Do you have any questions?

kamalida@msu.edu