

Using Persuasive Writing Strategies to Explain and Detect Health Misinformation

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

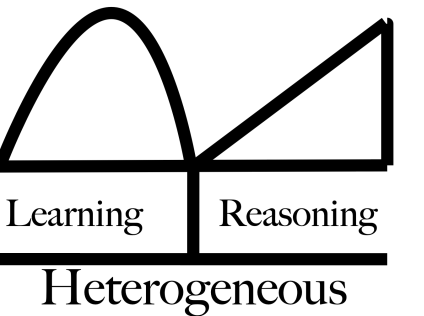
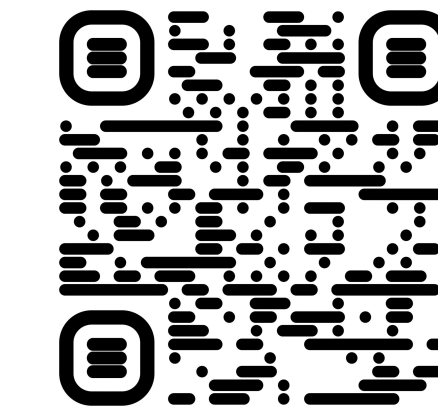
* Department of Computer Science and Engineering, Michigan State University

† 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



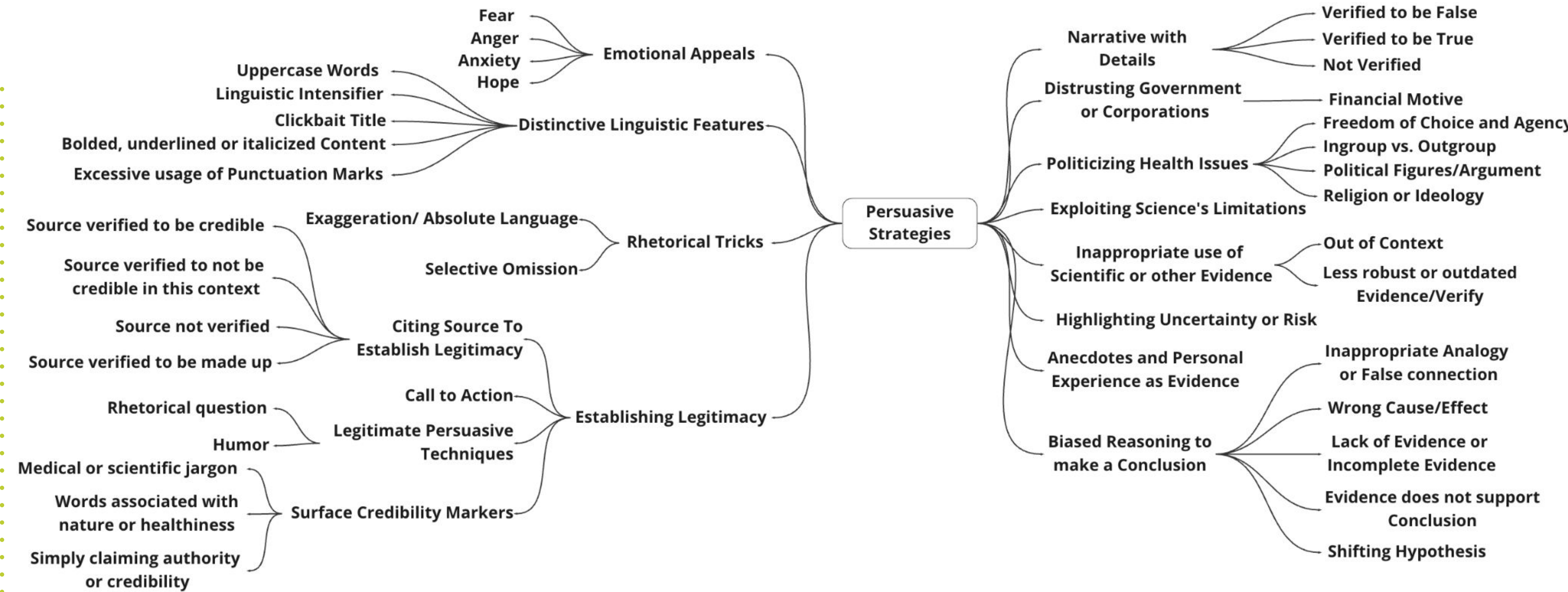
Highlights

Research questions

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

Outcomes:

- Created a comprehensive annotation scheme for persuasive writing in health misinformation.
- Showed that including persuasive strategies boosts accuracy and explainability of detection models.
- Evaluated improvements in detection performance using BERT and GPT models.



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

Misinformation Detection results(Task 3). Pred (context) refers the context level, and GT refers to ground truth strategy labels.

RAWFC benchmark

- Shows real-world capability
- Achieved SOTA 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 the RAWFC benchmark. † denotes the access to search engines.

Explainability

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

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.

An example a reasoning generated by GPT-4 with in context learning

Take away messages

- Persuasive strategies significantly improve misinformation detection.
- Our approach provides clear explanations for misinformation classifications.
- Our novel scheme offers a systematic way to analyze persuasive writing.
- Our method is applicable across various misinformation domains.

REFERENCES

- Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4685–4697, Hong Kong, China. Association for Computational Linguistics.
- Xuan Zhang and Wei Gao. 2023. Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method.
- Peng, W., Lim, S., & Meng, J. (2023). Persuasive strategies in online health misinformation: a systematic review. Information, Communication & Society, 26(11), 2131–2148. <https://doi.org/10.1080/1369118X.2022.2085615>

Motivation

Health misinformation is important

- Societal impacts of health misinformation.
 - Undermining trust in medical advisories.
 - Dire consequences.
- Using experts is costly.

Automatic Misinformation Detection

- Must be **Accurate** and **Transparent**
- Validating with knowledge base.
- Using linguistic features.
 - Persuasive Writing Strategies

Dataset Construction

Source

- Health-related subset of MultiFC.

Article Retrieval

- Manually retrieved a supporting article of each claim.

Annotation Process

- Developed a **novel annotation scheme** for persuasive writing strategies.
- Annotated by three experts in communication and misinformation studies.

Persuasive Strategies Identified

- 12 main different strategies e.g. *Emotional Appeals*
- Three hierarchical levels to enhance specificity.

Quality Control

- Cohen's Kappa Inter-annotator reliability: 88% average

Final Dataset

- 242 articles with veracity (binary) labels.
- 5,666 sentences annotated with one or more persuasive writing strategy.

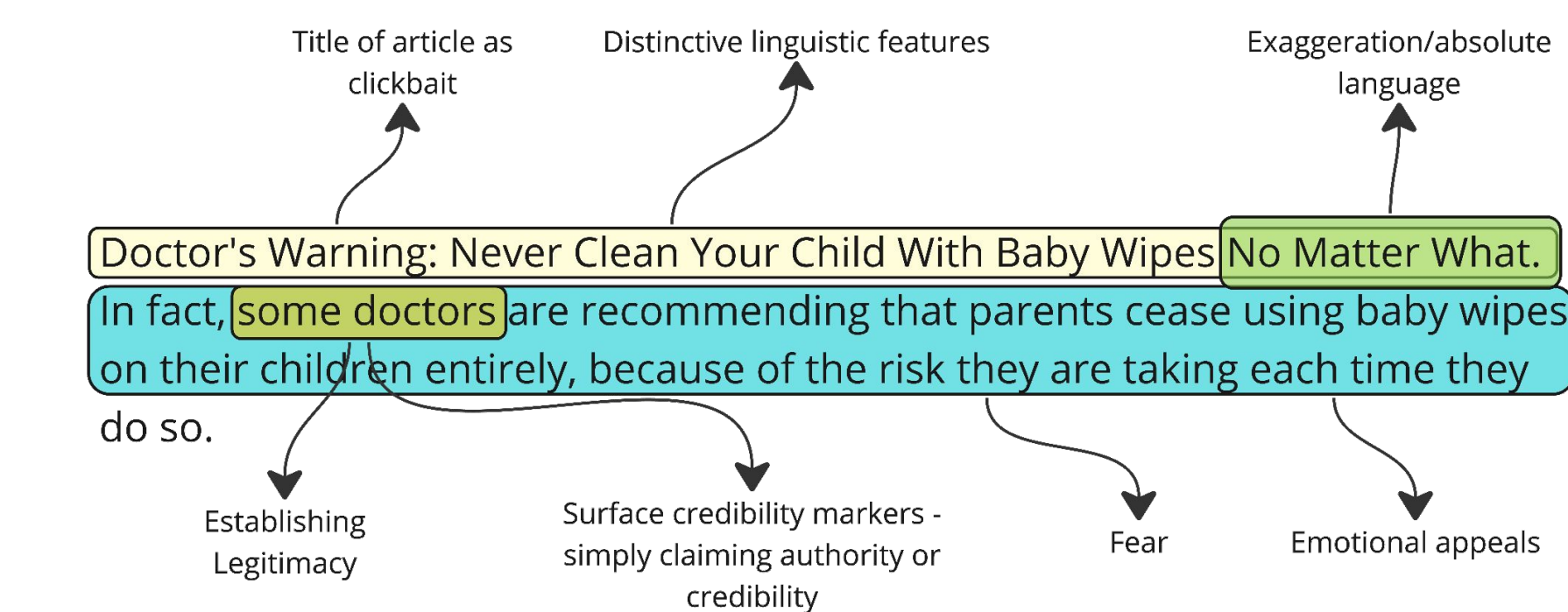
Problem Setting

Task 1: Misinformation Detection

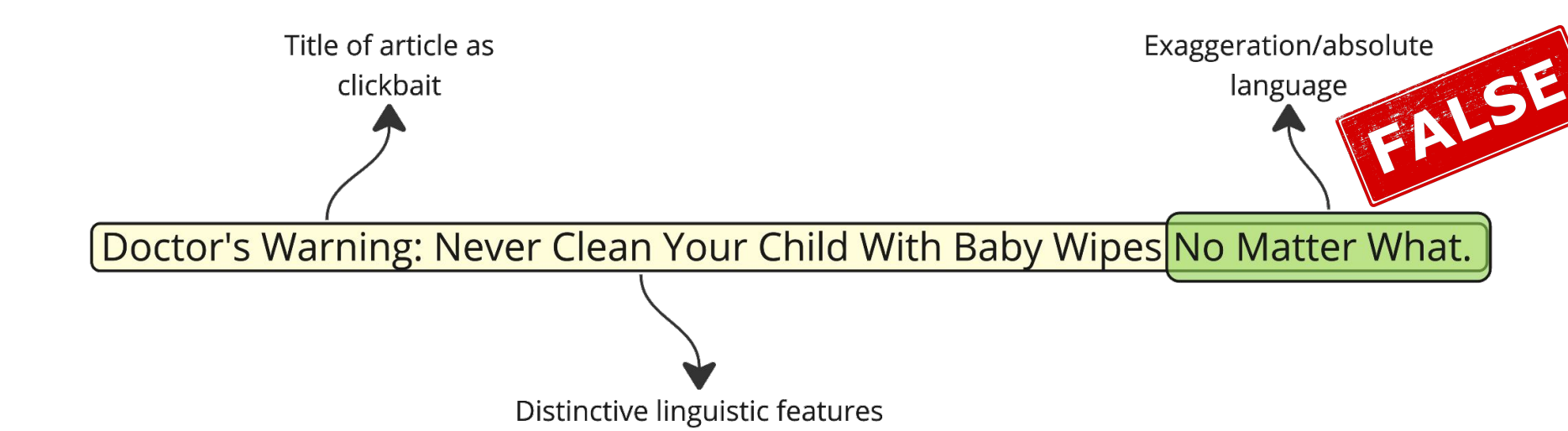
Doctor's Warning: Never Clean Your Child With Baby Wipes No Matter What.

FALSE

Task 2: Persuasive Strategy Labeling



Task 3: Misinformation Detection with Persuasive Strategy Labeling



Experiments

Persuasive Strategy Labeling (Task 2)

- Fine-tuned RoBERTa
- Experimented with three context levels.
- Evaluated in four different level of hierarchy.

Misinformation Detection (Tasks 1 & 3)

- Prompted GPT with different methods.
- Fine-tuned RoBERTa for misinformation.
- GPT-3/4 used in zero-shot and in-context learning.

Misinformation Detection

- Superior performance of GPT over RoBERTa
- GPT-4 significant 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

Misinformation Detection results (Task 1)

Persuasive Strategy Detection

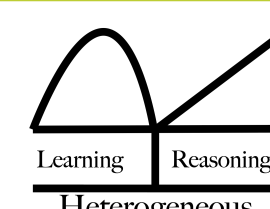
- Low performance shows the complexity of this task.
- Increasing context generally improves detection accuracy.
- only Level-4 doesn't increase as context grows

Level 1			Level 2		
Context	F1 _{Micro}	F1 _{Macro}	Context	F1 _{Micro}	F1 _{Macro}
None	0.668	0.662	None	0.468	0.342
Low	0.677	0.676	Low	0.501	0.372
High	0.699	0.692	High	0.497	0.383
Level 3			Level 4		
Context	F1 _{Micro}	F1 _{Macro}	Context	F1 _{Micro}	F1 _{Macro}
None	0.430	0.270	None	0.428	0.291
Low	0.464	0.283	Low	0.434	0.265
High	0.463	0.288	High	0.441	0.256

Sentence-level results on four levels of persuasive strategies, and three context levels

Misinformation Detection with Persuasive Strategies

- Significant improvement with Ground-Truth (GT) persuasive labels.
- GT alone performs better than claim itself.
- For *Claim + Article + GT*
 - GPT-4 Performs best.
 - RoBERTa Fails due to token limitation.
- In context learning
 - Outperforms RoBERTa predictions
 - Close to ground-truth labels



Using Persuasive Writing Strategies to Explain and Detect Health Misinformation

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

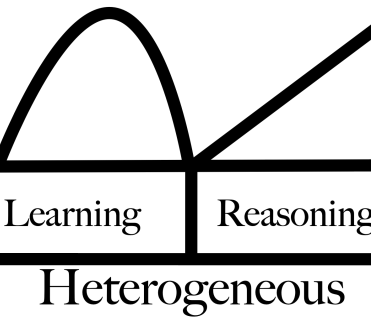
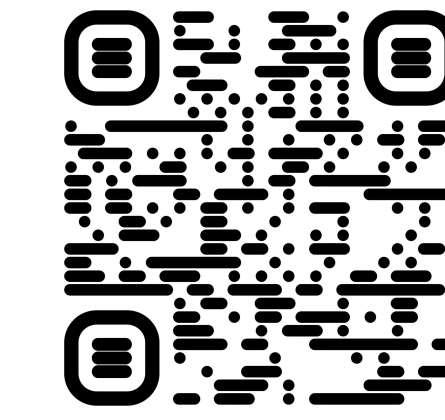
* Department of Computer Science and Engineering, Michigan State University

† 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



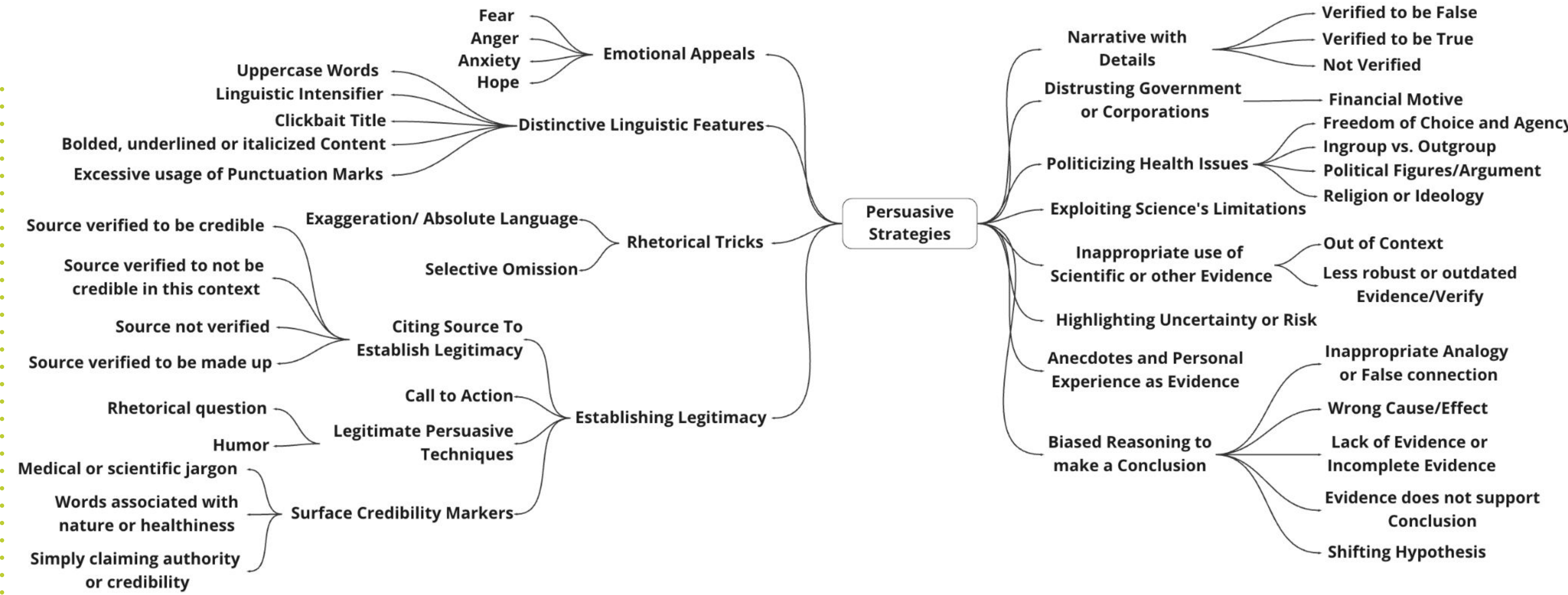
Highlights

Research questions

- How to identify persuasive writing strategies in health misinformation systematically?
- Can persuasive strategies improve misinformation detection models?
- How effectively can language models incorporate persuasive strategies for enhanced detection?

Outcomes:

- Created a comprehensive annotation scheme for persuasive writing in health misinformation.
- Showed that including persuasive strategies boosts accuracy and explainability of detection models.
- Evaluated improvements in detection performance using BERT and GPT models.



Misinformation Detection + Persuasive Strategies

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

Misinformation Detection results with different sources (Task 3). Pred context refers to persuasive strategy prediction labels concerning the context level, and GT refers to the strategy ground truth labels. Articles in the RoBERTa experiments are summarized to satisfy its token limit. GPT-4 (in-context learning) result has persuasive strategies information in the prompt.

- Integration of persuasive strategies led to an accuracy improvement of over in GPT-4 and RoBERTa.
- Integration of noisy persuasive strategy labeling deteriorates the performance while GT helps.
- Comparing GT and Pred models underscores potential improvements and existing gaps.
- Using In-context learning we can achieve better accuracy with limit sample size.

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

Experimental results of claim verification on the RAWFC. Bold denotes the best performance. † denotes the access to search engines. We used the davinci-003 to ensure a fair comparison.

- Using In-context Learning of persuasive strategies we achieved SOTA on RAWFC dataset.

Take away messages

- Persuasive strategies significantly improve misinformation detection.
- Our approach provides clear explanations for misinformation classifications.
- A novel scheme offers a systematic way to analyze persuasive writing.
- This method is applicable across various misinformation domains.

REFERENCES

- Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4685-4697, Hong Kong, China. Association for Computational Linguistics.
- Xuan Zhang and Wei Gao. 2023. Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method.
- Peng, W., Lim, S., & Meng, J. (2023). Persuasive strategies in online health misinformation: a systematic review. Information, Communication & Society, 26(11), 2131-2148. <https://doi.org/10.1080/1369118X.2022.2085615>

Problem Setting

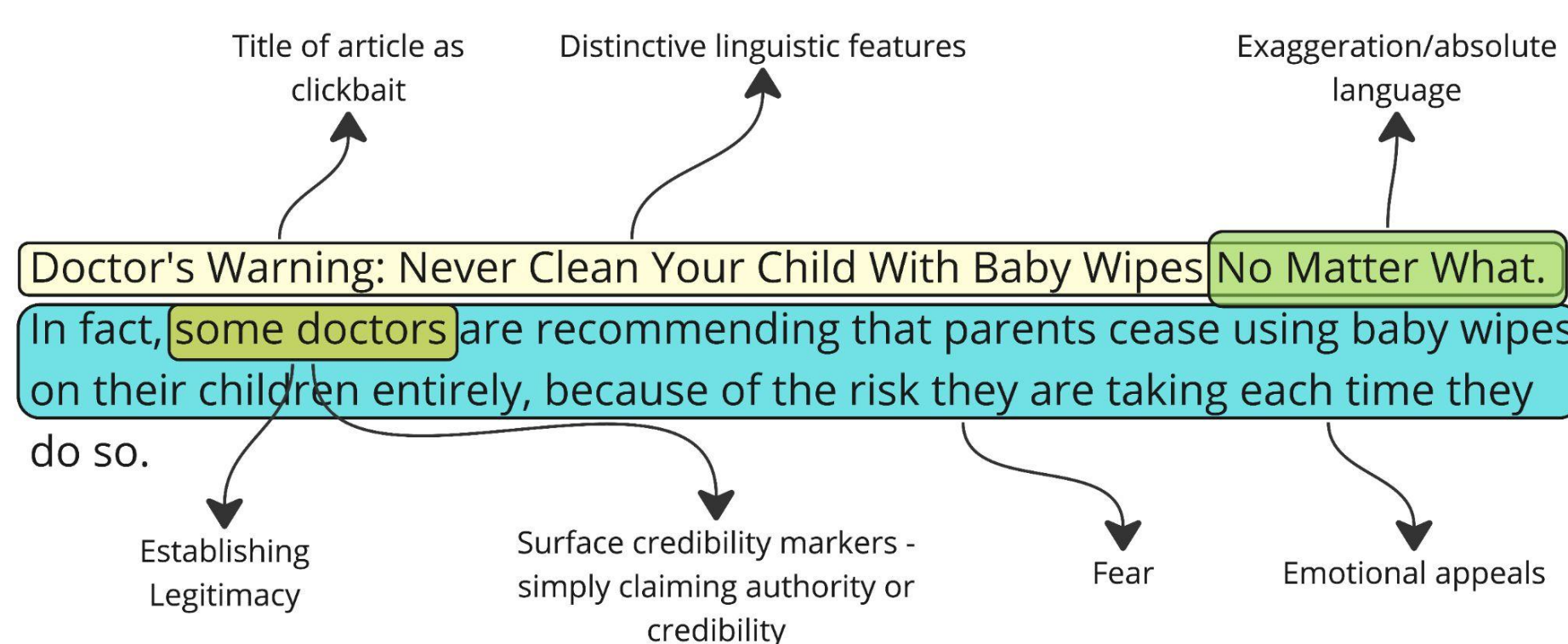
Tasks

1. Misinformation Detection

- Objective:** Classify misinformation news using their claim and articles.
- Challenge:** Determining truthfulness involves analyzing content's accuracy, context, and intent.

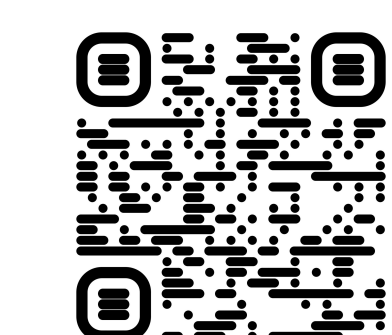
2. Persuasive Strategy Labeling

- Objective:** Identify and categorize text spans with persuasive strategy writing.
- Use of hierarchical labels to classify strategies.
- Challenge:** Techniques vary widely and require deep linguistic analysis and specialized annotations.



3. Misinformation Detection + Persuasive Strategy Labeling

- Combines the tasks of misinformation detection and explanation of persuasive strategies.
- Multi-task Learning or Pipeline setting.



Motivation

Labeling Techniques:

- Identify and expose hidden persuasive techniques in health misinformation.

Improving Detection:

- Combine persuasive strategy analysis with traditional text analysis to enhance detection accuracy.

Building Trustful AI:

- Provide detailed explanations for misinformation classifications to foster trust in AI systems.

Dataset Construction

Source

- Health-related subset of MultiFC.
- Filtered initial 599 claims to retain 242 relevant 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 experts in communication and misinformation studies using WebAnno.

Persuasive Strategies Identified

- 12 main different strategies such as emotional appeals
- Three hierarchical levels to enhance specificity.
- Each sub-tag is also marked with its corresponding higher-level tags.

Quality Control

- Inter-annotator reliability was assessed to ensure consistency, with Cohen's Kappa used as the measure of agreement.

Final Dataset

- 242 articles with veracity (binary) labels.
- 5,666 sentences, each sentence annotated for the presence of one or more persuasive writing strategy.

Experiments

Persuasive Strategy Labeling

- Fine-tuned RoBERTa
- Experimented with three context levels.
- Evaluated in four different level of hierarchy.

Misinformation Detection

- Employed RoBERTa and GPT family models.
- Prompted GPT with different methods.
- Fine-tuned RoBERTa for misinformation.
- GPT-3/4 used in zero-shot and in-context learning.

Persuasive Strategy Detection

- Low performance shows the complexity and challenge of this task.

Context	Level 1		Level 2	
	F1 _{Micro}	F1 _{Macro}	F1 _{Micro}	F1 _{Macro}
None	0.668	0.662	0.468	0.342
Low	0.677	0.676	0.501	0.372
High	0.699	0.692	0.497	0.383

Context	Level 3		Level 4	
	F1 _{Micro}	F1 _{Macro}	F1 _{Micro}	F1 _{Macro}
None	0.430	0.270	0.428	0.291
Low	0.464	0.283	0.434	0.265
High	0.463	0.288	0.441	0.256

RoBERTa results for four levels of persuasive strategies, with three levels of context.

Misinformation Detection

- Superior performance of GPT over RoBERTa
- GPT-4 significant 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

Misinformation Detection results (Task 1). Model_{source} represents the foundation model classifier with source as the input source

