UE20CS334 Natural Language Processing Project

Logical Fallacy Detection

PRESENTED BY

Ronit Srivastav - PES1UG21CS499

Basavaraj Haravi - PES1UG21CS132

Abhijeet M Baloji - PES1UG21CS019

DEPARTMENT OF CSE

PES UNIVERSITY

UNDER THE GUIDANCE OF

DR. MAMATHA.H.R

PROFESSOR

DEPARTMENT OF CSE

PES UNIVERSITY

Agenda

- Introduction
- Problem Statement
- Literature Survey
- References

What are logical fallacies?

A logical fallacy is a flaw in reasoning. Logical fallacies are like tricks or illusions of thought, and they're often very sneakily used by politicians and the media to fool people.

Problem statement

Develop a system capable of detecting fine-grained logical fallacies within textual content and accurately identifying the specific type of fallacy present.

Literature Survey

Taxonomies of logical fallacies need alignment with existing benchmarks.

Language model-based approaches are insufficient due to the abstraction required from syntax to semantics.

Background knowledge is crucial for understanding fallacies.

Implicit knowledge in fallacies must be made explicit.

Data sparsity poses a challenge for supervised learning.

Scalable mechanisms are needed to combat data sparsity

Robust and explainable identification of logical fallacies in natural language arguments

Zhivar Sourati [a b], Vishnu Priya Prasanna Venkatesh [a b], Darshan Deshpande [a b], Himanshu Rawlani [a b], Filip Ilievski [a b], Hông-Ân Sandlin [c], Alain Mermoud [c]

Show more

	Robust and Explainable Identification of Logical Fallacies in Natural Language Arguments[2] 2023	CBR with Language Model For classification of logical Fallacies 2023	Logical Fallacy Detection 2022
Aim	• Formalizing logical fallacy detection for AI. • Focus on robustness, explainability, and addressing data sparsity.	To classify new instances of logical fallacies in natural language arguments with a focus on improving the performance of language models in reasoning over complex logical structures	The aim of the research paper is to propose a new task of logical fallacy classification and to develop a model to detect logical fallacies in text. The authors also created a dataset of logical fallacies to train and test their model.
Dataset	BIG Bench Logical Fallacy Dataset, LOGIC and LOGIC Climate Datasets	LOGIC dataset LOGIC Climate Dataset	LOGIC dataset, which contains 2,449 logical fallacy instances across 13 logical fallacy types. LOGICCLIMATE
Feature	Formal Framework for Logical Fallacy Identification	Case-Based Reasoning(CBR) Model Evaluation	The feature used in the research paper is the text of the claim or argument.
Model Used	NLI Electra, NLI FCL Electra, IBR Electra, PBR Electra, KI BERT	Freq-Based Codex ELECTRA RoBERTa BERT	The model used in the research paper is a structure-aware classifier based on a pretrained language model (Electra). The structure-aware classifier is designed to focus on the logical form of the text rather than the content words.
Gap	Abstraction and Generalization, Knowledge Integration, Data Sparsity, Model Interpretability, Real-World Applicability	• Knowledge Transfer and Similarity Function • Complementary information From Explanations	• There is no existing task of logical fallacy classification for NLP models. • Existing datasets on argument quality are limited in size and scope. • Existing NLP models have limited performance in detecting logical fallacies.
Accuracy	Best of 99.7 on BIG Bench And best of 83 on LOGIC and LOGIC climate datasets	an accuracy of 0.613 on in-domain settings and 0.616 on out-of-domain settings	The accuracy of the structure-aware classifier on the LOGIC dataset is 58.77% F1 score. The accuracy on the LOGICCLIMATE challenge set is 23.81% F1

	Argument-based Detection and Classification of Fallacies in Political Debates 2023	COCOLOFA: News Comment Sections with Common Logical Fallacies 2024	Learning about informal fallacies and the detection of fake news: An experimental intervention 2023
Aim	The aim of the paper is to detect and classify fallacies in political debates using transformer-based architectures and a contextual framework.	The paper aims to address the limitations of existing datasets by providing a larger collection of text units labeled with logical fallacies, spanning a broad array of topics and featuring longer text units on average	Explore if online learning about informal fallacies improves fake news discernment. To investigate if learning about informal fallacies improves people's ability to detect fake news.
Dataset	ElecDeb60to20 dataset, US debates between 1960-2020.	online news comments is COCOLOFA (8 types)	Not explicitly mentioned in the paper, but likely consisted of: Textual materials for the learning interventions on informal fallacies and fake news Informal fallacy identification tasks Real and fake news articles
Feature	Formal Framework for Logical Fallacy Identification Dataset Extension, Fallacy Detection Model	Construction of COCOLOFA dataset Attention-check questions for quality control Data-driven approach for article selection	• Independent variable: Type of learning intervention (informal fallacies vs. fake news) • Dependent variables: ◦ Ability to spot informal fallacies ◦ Discernment between real and fake news articles
Model Used	. Specifically, it introduces the MultiFusion BERT model	ERT, NLI with RoBERTa as the backbone	• Not a formal statistical model, but a causal relationship is examined: ◦ Learning intervention -> Ability to spot informal fallacies -> Discernment of real vs. fake news (mediated effect)

	Multitask Instruction-based Prompting for Fallacy Recognition 2022	How susceptible are LLMs to Logical Fallacies? 2023	Teaching Informal Logical Fallacy Identification with a Cognitive Tutor
Aim	Identifying fallacies in text.	assess the robustness of Large Language Models (LLMs) against logical fallacies in multi-round argumentative debates using the LOGICOM.	The big-picture goal is to create a computer program, like a smart filter, that can automatically find informal fallacies in written text. Imagine a program that underlines these fallacies in your writing the same way a spellchecker underlines misspelled words.
Dataset	ARGOTARIO, CLIMATE, COVID , LOGIC	5k pairs of logical vs. fallacious arguments	They don't have a collection of text examples (data) yet.
Feature	The key feature is the introduction of a multitask instruction-based prompting , covering great number of fallacies	a new dataset of over 5,000 pairs of logical vs. fallacious arguments extracted from multi-round debates to assess LLMs' logical reasoning capabilities.	These are the specific characteristics the program will look for in text to identify fallacies. For instance, a feature might be identifying certain phrases commonly used in straw man arguments
Model Used	T5 (Text-to-Text Transfer Transformer) model.	GPT-3.5 and GPT-4	There's no computer program yet (model) to identify fallacies. This will be built later based on what they learn from this initial study
Gap	Unified Framework for Fallacy Recognition, Dataset Expansion, no general data	Lack of Comprehensive Evaluation Inconsistency in LLMs' Reasoning:	Currently, there isn't a good system to automatically detect these fallacies in text. This research aims to bridge that gap by creating a program that can do this automatically.
Accuracy	Max of 70% min of 19% across	both GPT-3.5 and GPT-4 are 41% and	They haven't built the system that would identify

	Detecting Argumentative Fallacies in the Wild: Problems and Limitations of Large Language Models 2023	The Search for Agreement on Logical Fallacy Annotation of an Infodemic 2022	Performance of Critical Thinking and Existence of Logical Fallacies in Indonesian Varsity English Debate 2020 in Jakarta 2021
Aim	analyze the limitations of data-driven approaches in detecting argumentative fallacies, particularly focusing on the challenges of applying these approaches in real-world scenarios	establish a consensus on the best practices for annotating logical fallacies in the context of an infodemic, aiming to improve the understanding and analysis of misinformation	The aim of the research can be inferred to be analyzing critical thinking skills and logical fallacies in English debates.
Dataset	Fallacy Detection Corpus(5 types), Argumentation Scheme Validation Dataset(7 types)	COVID-19 related texts, which includes 26 documents. The documents cover six COVID-19 topics	The "data" used was the spoken utterances from the Grand Final National University Debate Competition 2020 in Indonesia.
Feature	critical analysis of the limitations of current data-driven approaches , highlight the challenges of relying solely on deep learning algorithms	the identification and annotation of logical fallacies within COVID-19 related texts, using the Argotario fallacy annotation schema, to improve the understanding and analysis of misinformation.	It analyzed aspects of the debaters' arguments like identifying reasons, assumptions, and types of fallacies used.
Model Used	DL, GPT-3.5-TURBO, GPT-4	Pattern-Exploiting Training	There wasn't a machine learning model used in this research
Gap	They highlight the limitation of current approaches that focus on labeling short spans of text with fallacy labels without considering the underlying logic that makes an argument fallacious or not	Need for a Clear Annotation Schema Challenge of Identifying Logical Fallacies	The gap this research tries to address is the lack of understanding of how critical thinking and logical fallacies play out in English debates.

Consolidation of Research gap

Variable accuracy in different datasets

Data for a large number of logical fallacies.

Not detecting the logic of the statement in detail.

Fallacies to include(can’t do strawman, tu quoque, ambiguity, texas sharpshooter)

1) Bandwagon fallacy
2) Appeal to Authority
3) Appeal to Majority
4) Appeal to Nature
5) Appeal to Tradition
6) Appeal to Worse Problems
7) False Dilemma
8) Hasty Generalization
9) Slippery Slope
10) Red Herring
11) Ad Hominem
12) Loaded Question
13) Burden of Proof
14) The Gambler’s Fallacy
15) Anecdotal
16) Genetic Fallacy
17) Middle Ground Fallacy
18) Appeal to Emotion
19) False Cause
20) Circular Reasoning
21) Non Sequitur
22) Irrelevant Authority
23) Personal Incredulity
24) Special Pleading
25) The Fallacy Fallacy

General datasets to include

1) LOGIC dataset
2) Big Bench dataset
3) COCOLOFA

Summary of Literature Survey

1. The document suggests exploring two parallel streams of AI methods for logical fallacy identification. One stream involves neural language models like GPT-3 and Codex, while the other stream focuses on neuro-symbolic methods such as reasoning as a soft logic problem.
2. Data augmentation for more data and lesser known logical fallacies.
3. Getting extra statements for lesser known logical fallacies through AI.
4. Able to get acceptable accuracy in multiple datasets
5. Combining datasets to make a more general and robust model

Reference

[1]Sourati, Z., Ilievski, F., Sandlin, H., & Mermoud, A. (2023).*Case-Based Reasoning With Language Models for Classification of Logical Fallacies*. https://doi.org/10.24963/ijcai.2023/576

[2] Zhivar Sourati, Vishnu Priya Prasanna Venkatesh, Darshan Deshpande, Himanshu Rawlani, Filip Ilievski, Hông- n Sandlin, Alain Mermoud,*Robust and explainable identification of logical fallacies in natural language* arguments,Knowledge-Based Systems, Volume 266, 2023, 110418, ISSN 0950-7051,https://doi.org/10.1016/j.knosys.2023.110418.(https://www.sciencedirect.com/science/article/pii/S0950705123001685)

[3] Hruschka TMJ, Appel M (2023) *Learning about informal fallacies and the detection of fake news: An experimental intervention*. PLoS ONE 18(3): e0283238. (https://doi.org/10.1371/journal.pone.0283238)

[4] Diana, N., Eagle, M., Stamper, J., Koedinger, K.R. (2017). *Teaching Informal Logical Fallacy Identification with a Cognitive Tutor*. In: André, E., Baker, R., Hu, X., Rodrigo, M., du Boulay, B. (eds) Artificial Intelligence in Education. AIED 2017. Lecture Notes in Computer Science(), vol 10331. Springer, Cham. https://doi.org/10.1007/978-3-319-61425-0_74

[5] Scovanner, Paul, Saad Ali, and Mubarak Shah. "A 3-dimensional sift descriptor and its application to action recognition." Proceedings of the 15th ACM international conference on Multimedia. 2007

[6] Sourati, Z., Ilievski, F., Sandlin, H., & Mermoud, A. (2023a, January 27). Case-Based Reasoning with Language Models for Classification of Logical Fallacies. arXiv.org. https://arxiv.org/abs/2301.11879

[7] A. Payandeh, D. Pluth, J. Hosier, X. Xiao, and V. K. Gurbani, "How susceptible are LLMs to Logical Fallacies?," arXiv.org, Aug. 18, 2023. https://arxiv.org/abs/2308.09853

[8] Muhammad Zulfikar Faishol Ali, Abdurrachman Faridi, Zulfa Sakhiyya (2021); *Performance of Critical Thinking and Existence of Logical Fallacies in Indonesian Varsity English Debate 2020 in Jakarta*; International Journal of Scientific and Research Publications (IJSRP) 11(1) (ISSN: 2250-3153), DOI: http://dx.doi.org/10.29322/IJSRP.11.01.2021.p10980

Thank You