ETH zürich

Creating a dataset for Fallacy Detection

Gauthier Boeshertz, Alvaro Cauderan, Franz Nowak

1 Introduction

The information age has led to a vast increase in misinformation and bad reasoning. News media and social networks struggle to keep up with and mark these as such. Similar to automatic spelling and grammar correction tools and fact checkers, automatic fallacy detection could help improve online discourse.

[1] introduced the task of fallacy classification by training a classifier to differentiate different types of fallacies. We extend this work by creating a larger fallacy dataset with counterexamples and train a binary classifier on it to detect whether an argument is a fallacy or not.

2 Dataset

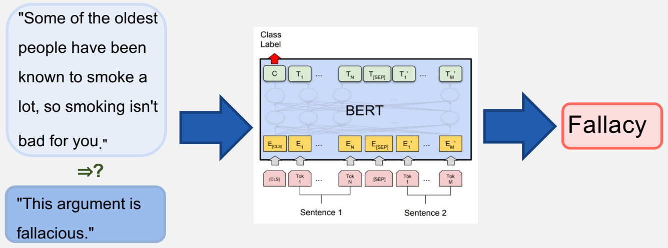
- [1] collected two datasets, called `Logic` and `LogicClimate`, which contain fallacies mostly from online teaching resources, and climate related fallacies from an online climate related news website, respectively.
- In order to use these datasets to train our binary fallacy classifier, we extended them by adding non-fallacy examples we sourced from a set of Kialo discussions created by [2]. This resulted in two extra datasets, `LogicValid` and `LogicClimateValid`. Kialo is an online debating platform so it contains arguments from real world exchanges, and we used users' ratings of arguments as a proxy to select the most logically sound arguments.
- Finally, we generated two extra datasets by scaping arguments ourselves from Kialo using user responses as proxy binary fallacy labels for then cleaning the data by hand. `Kialo` consists of the Kialo fallacies, while `KialoValid` contains the corresponding counterexamples.

Dataset name	number of statements
LOGIC	2,449
LOGICVALID	2,200
LOGICCLIMATE	1,079
LOGICCLIMATEVALID	721
KIALO	848
KIALOVALID	3,067

Table 1: Datasets and counts of statements comprising our final dataset

3 Modeling

- We use Natural Language Inference (NLI), see [3], to classify whether the hypothesis "this argument is fallacious", follows from the premise (the potentially fallacious claim).



- We also tested fallacy classification performance on a hold-out distribution related to climate change to test the model's generalisation capacity.

	P	R	F1	Acc
BERT	**72**	**67**	**61**	**62**
Electra	71	65	57	58
Roberta	71	65	58	59
deBERTa	70	63	55	57

Table 3: Comparison of different models on the climate test set

4 Results

- We split the dataset into train, validation, and hold-out test set. The best model achieves 89.4% accuracy on the test set.

	P	R	F1	Acc
Bart-MNLI	51	51	51	53
Roberta-MNLI	52	52	52	54
Bert	87	85	86	87
Bert + SA P	86	84	85	86
Bert + Hypo	89	85	86	87
Bert + SA P + Hypo	87	85	86	87
Electra	88	86	87	87
Roberta	**90.0**	**88.1**	**88.8**	**89.4**
deBERTa	89.9	87.9	88.5	89.0

Table 2: Comparison of different models on the combination of all the data

5 Summary

1. We introduce the task of **fallacy detection**, i.e. the binary classification of arguments into fallacies and non-fallacies.
2. We aggregated a dataset of fallacies and counterexamples extending the datasets from [1] and collecting our own data from Kialo.
3. Finally, we use state of the art preprocessing and modeling techniques to show how this data can be used to train high accuracy classifiers for fallacy detection (89% accuracy).

References

1. Logical Fallacy Detection, Jin et al, 2022, https://arxiv.org/abs/2202.13758
2. GraphNLI: A Graph-based Natural Language Inference Model for Polarity Prediction in Online Debates, Agarwal et al, 2022, https://arxiv.org/abs/2202.08175
3. Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach, Yin et al 2019 https://arxiv.org/abs/1909.00161