Leveraging Natural Language Understanding for Misinformation Detection

Haley Johnson
haleyej@umich.edu

1 Introduction

In the last decade years, Americans have witnessed a growing epidemic of fake news. Misinformation has eroded public confidence in U.S. elections and endangered election workers (West et al., 2022). Most notably, misinformation fueled former President Donald Trump's "big lie" (West et al., 2022) that the 2020 presidential election had been stolen, ultimately leading to the riot at the U.S. capitol on January 6. Recent events such January 6th have brought renewed attention to automatic misinformation detection, and the advent of generative AI, such as ChatGPT, has raised new concerns about the proliferation of computer generated misinformation (Pendyala and Tabatabaii, 2023). Despite this, misinformation detection remains a largely unsolved problem in Natural Language Processing (Pendyala and Tabatabaii, 2023)

1.1 Project Goals

Previous work has shown that misinformation is stylistically distinct from factual media (Potthast et al., 2018). Style-based text classification and deception detection have achieved some succession at automatic misinformation detection, but there is still room for significant improvement (Potthast et al., 2018; Chen et al., 2015). One promising area of research that may have applications to misinformation detection is natural language inference.

A common natural language inference task is detecting if a hypothesis is entailed in a given premise. Of particular interest for misinformation research is logical fallacy detection. A logical fallacy is an argument "that appears correct and may even be extremely persuasive, but which proves upon closer inspection to be logically invalid" (McClurg, 2010). Logical fallacies commonly appear in misinformation (Jin et al., 2022) — the presence of a fallacy in one part of a text can signal that the rest of it is logically or factually unsound, or at the very least relies on dubious reasoning.

This project investigates if natural language understanding can be leveraged to detect fake news. In particular, I use the LIAR corpus, a benchmark dataset for fake news detection that contains 12.8K short statements from PoltiFact.com (Wang, 2017). Additionally, I aim to improve upon existing methods for detecting logical fallacies and evaluating reasoning ability in natural language. These improvements can be leveraged for other NLP tasks, such as classifying hate speech or propoganda.

2 Data

2.1 Logical Reasoning Data

This project leverage two datasets to evaluate logical reasoning in large language models: the Stanford Natural Language Understanding (SNLI) Corpus, a dataset of 500,000 logical relationships (Bowman et al., 2015), and a dataset of 1,700 logical fallacies developed by (Jin et al., 2022).

The SNLI coprus contains examples of logical entailment, contradiction, neural statements (where one statement does entail or imply another), and paradoxes. The dataset is available for download on online and has already been split into training, testing, and development sets.

The logical fallacy dataset was scrapped from educational websites and include 13 types of reasoning errors. It is available for download on github. The original authors have already separated the data into testing, training and development sets. The test set will be used in the evaluation portion of this project.

2.1.1 SNLI Dataset

The SNLI corpus was built using image captions from the photo sharing site Flickr. MTurk workers were presented with a premise — the caption — and asked to generate a hypothesis that conflicted with, was entailed in, or was neither conflicted nor

Data Source	Purpose	Number of Examples
Stanford Natural Language Inference Coprus	Train the model to learn a general understanding of logical inference	570,000
Logical Fallacies	Explicitly train the model to understand different kinds of logical fallacies, with the hope it will develop a richer understanding logical inference	1,849
LIAR	Determine if logical inference abilities help large language models learn to detect misinformation	12,836

Table 1: Summary of project data sources

entailed by the original premise (Bowman et al., 2015). The corpus is one of the largest in natural language inference and commonly used as a benchmark in the field.

For instance, one example in the dataset is:

- **Premise:** A person on a horse jumps over a broken down airplane
- **Hypothesis:** A person is training his horse for a competition
- **Label:** Neutral

Another example is:

- **Premise:** Six soccer players on field with player in red uniform in the air and ball airborne
- **Hypothesis:** Six people are playing basketball
- **Label:** Contradiction

The dataset is evenly split between examples of entailment, contradiction, and neutral statements. Training examples were already in a tidy format and only needed to be tokenized.

2.1.2 Logical Fallacy Dataset:

This dataset is compromised of short statements and their associated logical. These statements are typically 1-4 sentences. The dataset contains examples of 13 kinds of logical fallacies - the distribution of classes and examples are shown in Table 2.

The logical fallacy dataset was scrapped from several websites. As a result, some light preprocessing was needed. Some training examples were in the format "{logical statement} is an example of: ". I modified the text so only the fallacious statement was in the dataset.

Likewise, some examples note who is making a statement. For instance, one example in the dataset is: "Board Member: “If this company is going to maximize its profits in the coming year, we need to fully exploit all of our available resources.”Resources Director: “Not so fast. Our employees are one of our most valued resources, and we have a strict policy against exploiting our workers.”" I removed the beginning portion so the final dataset only included the fallacy.

2.2 LIAR Misinformation Benchmark:

The LIAR dataset contains 12,836 short statement that are annotated for "truthfulness, subject, context/venue, speaker, state, party, and prior history" (Wang, 2017). It includes statements from democrats and republicans and remarks made in speeches, official statements and social media posts. Statements are categorized as "pants on fire," "false," "barely true" and "mostly true." Labels were manually assigned by a PoltiFact editor.The dataset is freely avaliable to download online and has been split into train, test, and validation sets by the original creators.

One training example is:

- **Statement:** “The last quarter, it was just announced, our gross domestic product was below zero. Who ever heard of this? Its never below zero.”
- **Speaker:** Donald Trump
- **Context:** presidential announcement speech
- **Label:** Pants on Fire
- **Justification:** According to Bureau of Economic Analysis and National Bureau of Economic Research, the growth in the gross domestic product has been below zero 42 times

Logical Fallacy	Example	Percent of Examples
Faulty Generalization	All four year olds talk too much	17.25%
Ad Hominem	Only a selfish, non-caring person would believe that this is ok	12.17%
False Causality	People who eat yogurt have healthy guts. If I eat yogurt I will never get sick	9.41%
Ad Populum	Officer, I was only driving as fast as everyone around me. I'm sure I wasn't speeding	8.54%
Circular Reasoning	The Cardinals are the best football team because they're better than all the other teams	7.25%
Appeal to Emotion	Same-sex marriage must be prohibited, or the family structure as we know it will collapse	7.03%
Fallacy of Logic	We should stop using hairspray because it is snowing in New York	6.54%
Fallacy of Relevance	Finish your dinner. There are starving children in Africa	6.17%
Intentional	Real Cubans understand presumption of guilt	6.06%
False Dilema	You can buy the new Sensitivity perfume, or you can smell bad all day	5.95%
Fallacy of Credibility	My uncle is a mechanic and he says you shouldn't spank children. He says it's ineffective	5.79%
Fallacy of Extension	Mayor Blake wants to create more bicycle lanes in Lowell. Why is he forcing us to give up our cars and bike everywhere?	5.73%
Equivocation	According to the Supreme Court, we have a right to abortion. Therefore, it is right to have an abortion	2.11%

Table 2: Class distribution in logical fallacies dataset

over 68 years. Thats a lot more than "never." We rate his claim Pants on Fire!

3 Related Work

3.1 Natural Language Understanding

Jin et al. (2022) adopt a "structure aware approach" for detecting logical fallacies about climate change. Their approach, which draws inspiration from philosophy and the study of logic, holds that the *structure* of the argument, rather than the content, is the best way to identify faulty reasoning. The author's selected several pre-trained baseline models and fine-tuned them on data about logical fallacies, achieving 57% accuracy. Similar to this project, the authors were working with a small dataset of 1,700 logical fallacy examples and 1,000 fallacious claims about climate change.

Tian et al. (2021) fine-tune three state of the art language models (BERT, RoBERTa, Xl-Net) on the SNLI corpus. They find that all three models perform better than random guessing, but are worse than human annotators. RoBeRTa performed the best, achieving 68.3% accuracy, while BERT achieved 55.9% accuracy. All three model's performance significantly degraded when more irrelevant information was introduced. This indicates that these models may struggling to pick out relevant information when evaluating claims in longer or more complex texts.

Notably, the authors found that training the models to detect paradoxical claims significantly improved first-order logical reasoning and that the resulting models were less likely to pick up spurious correlations in the data. Therefore, fine-tuning a model on SNLI and then performing additional fine-tuning on (Jin et al., 2022)'s logical fallacy corpus may result in a more robust model that's better able to evaluate complex arguments.

Other work has compared different machine learning architectures. Bowman et al. (2015) curated a dataset of 570k natural language understanding examples to evaluate the accuracy of "rule-based systems, simple linear classifiers, and neural network-based models." In particular,

the authors implement a Long Short-Term Memory (LSTM) based neural model, which accurately classified 77.6% of textual entailment examples. However, much simpler rule-based systems were able to achieve 71.9% accuracy on the same dataset. Advancements in neural networks, namely transformed-based architectures and attention, have lead to significant gains on other NLP tasks and could further improvements in natural language inference.

3.2 Misinformation Detection

The creators of the LIAR corpus leverage achieve 27% accuracy on the dataset using a convolutions neural network (Wang, 2017). Interestingly, convolutions networks outperformed bidirectional Long Short-Term Memory based models. Additionally, the authors experimented with models that used both the text and its associated metadata, such as the speaker or the topic the statement was about, to make predictions, but found that text-metadata hybrid models did not significantly outperform text-only models (27.4% accuracy vs. 27%) (Wang, 2017). Similar to Bowman et al. (2015), leveraging transformed-based architectures and attention may offer additional accuracy gains.

Potthast et al. (2018) adopts a style-based approach to misinformation detection. Using Unmasking a "meta learning approach originally devised for authorship verification" (Potthast et al., 2018), they were able to accurately classify 75% of claims in a dataset of statements fact checked by journalists at Buzzfeed News. This suggests that hyperpartisan media has a distinctive writing style that helps discern it from mainstream media. This is promising, because leveraging natural language inference for misinformation detection presumes that there is something rhetorically distinctive between fake news and real news. While the authors do not investigate which textual features drive stylistic differences between articles in their corpus, it is reasonable to suggest that these difference may extend to the article's argumentative structure, use of evidence, or other rhetorical flourishes that are important for logical fallacy detection.

4 Methodology

This project will use transfer learning to evaluate logical reasoning in news articles. Transfer learning is a "a technique where a neural network is fine-tuned on a specific task after being pre-trained on a general task" (Malte and Ratadiya, 2019). Transfer learning offers several advantages. First, extensive work has been done on logical fallacy detection in other domains, such as climate denial (Jin et al., 2022), propaganda detection (Oliinyk et al., 2020), and hyper partisan news detection (Kiesel et al., 2019). This work suggests that natural language understanding abilities may help LLMs in downstream tasks like misinformation detection.

Likewise, training a model to develop general language understanding and inference abilities, rather than misinformation-detection specific abilities, will hopefully allow the language model to generalize better to unseen text. The nature of fake news is constantly changing, so any effective automatic misinformation detection method should be able to generalize to topics, speakers, and rhetoric it has not encountered in training.

In particular, this project will use multiple fine-tuning on a pretrained BERT language model. First, the model will be fine-tuned to learn to detect logical reasoning. Then, it will be fine-tuned to identify logical fallacies. I'll use low-rank adaptation of large language models (LoRA), a procedure developed by Hu et al. (2021) to support efficient fine-tuning of LLMs without diverging too much from the original models. Rather than directly modifying the original weights, LoRa decomposes the weight matrix into two smaller matrices A and B where the rank of A and B is lower than the original matrix (Hu et al., 2021). This reduces the number of trainable parameters, making the pre-training stage less computationally intensive.

By combining logical reasoning and logical fallacy detection, I believe I can improve upon existing methods for evaluating reasoning abilities in large language models. Finally, I'll use the resulting model to make predictions on the LIAR benchmark. This will allow us to examine if logical fallacy detection can improve upon existing methods for fake news detection.

5 Evaluation Results

5.1 Evaluation Metrics

All three datasets use in this project have test sets created by the original authors. I'll use these to evaluate my model. Ultimately, the goal of training the model to recognize logical entailment and logical fallacies is to improve its accuracy on misinformation detection. As a result, it's performance

Baseline	Performance
Most Frequent Class	20.81%
Naive Bayes	18.23%

Table 3: Baseline performance on fake news detection

on the LIAR benchmark dataset is the most important metric. The LIAR dataset is reasonably balanced across classes, so I will report the accuracy and F1 score. If my project is successful, my method should outperform the results achieved by Wang (2017) in the original paper that proposed the LIAR benchmark.

5.2 Baselines

I evaluated LIAR fake news corpus on two simple baselines. These baselines serve as a point to compare more sophisticated architectures against. Simply guessing the most frequent class in the dataset — half true — achieved 20.81% accuracy on the test set. Interestingly, a Naive Bayes baseline perform *worse* than just guessing the majority of class. A Guassian Naive Bayes model using a bag-of-words representation only correctly classified 18.23% of statements. This poor performance underscores the difficulty of detecting misinformation, as well the significant room for improvement on this task.

Additionally, I am trying to asses if a second fine-tuning step focused on logical errors will improve the model's performance. Therfore, in my final project I will pre-train one language model soley on the SNLI dataset and another language model on just the logical fallacy dataset, then asses if the model trained using both datasets achieves higher accuracy.

6 Work Plan

This is a rough timeline for the remainder of the semester

- **Week of April 8:** Implement LoRa to pre-train BERT model, generate plots and tables that detail training procedure
- **Week of April 15:** Apply model to fake news detection task — create plots that show results
- **Week of April 23:** Write, wrap up any remaining tasks report
- **April 26:** Project report due

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference.

Yimin Chen, Nadia K. Conroy, and Victoria L. Rubin. 2015. News in an online world: The need for an "automatic crap detector". *Proceedings of the Association for Information Science and Technology*, 52(1):1–4.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schoelkopf. 2022. Logical fallacy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7180–7198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. SemEval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Aditya Malte and Pratik Ratadiya. 2019. Evolution of transfer learning in natural language processing. *ArXiv*, abs/1910.07370.

Andrew Jay McClurg. 2010. Logical fallacies and the supreme court: A critical analysis of justice rehnquist's decisions in criminal procedure cases.

Vitaliia-Anna Oliinyk, Victoria Vysotska, Yevhen Burov, Khrystyna Mykich, and Vítor Basto Fernandes. 2020. Propaganda detection in text data based on nlp and machine learning. In *MoMLeT+DS*.

Vishnu S. Pendyala and Foroozan Sadat Akhavan Tabatabaii. 2023. Spectral analysis perspective of why misinformation containment is still an unsolved problem. In *2023 IEEE Conference on Artificial Intelligence (CAI)*, pages 210–213.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A Stylometric Inquiry into Hyperpartisan and Fake News. In *56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pages 231–240. Association for Computational Linguistics.

Jidong Tian, Yitian Li, Wenqing Chen, Liqiang Xiao, Hao He 0007, and Yaohui Jin. 2021. Diagnosing the first-order logical reasoning ability through logicnli. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP*

2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021, pages 3738–3747. Association for Computational Linguistics.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Darrell M. West, Elaine Kamarck, Marvin Kalb, William A. Galston, and Michael Hais Morley Winograd. 2022. Misinformation is eroding the public's confidence in democracy.