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, endangered election worker's safety, and contributed to growing partisan polarization (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. Events like January 6th have brought renewed attention to dangers unfettered misinformation and conspiracy theories on social media. Simultaneously, the advent of generative AI, like ChatGPT, has raised concerns about the proliferation of computer generated misinformation (Pendyala and Tabatabaii, 2023). The sheer volume of content on social media — notwithstanding generative AI's ability to automatically create thousands of posts — makes human content moderation infeasible at scale. As a result, platforms must deploy automated content moderation and misinformation detection systems. Despite the urgency of these challenges, misinformation detection remains an open 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.

Transfer learning is the practice of fine-tuning a language model on one task and then applying it to another (Malte and Ratadiya, 2019). The prevalence of logical fallacies in misinformation suggests that it may be possible to train a model to develop logical inference skills and then apply it to downstream tasks like misinformation detection. In particular, I fine-tune a pre-trained DistilBERT model to detect logical entailment, and then apply this model to the LIAR corpus, a benchmark dataset for fake news detection. I find that more work is needed to understand if natural language understanding can be leveraged to improve misinformation detection, as the fine-tuned model performed roughly the same as a base DistilBERT model with no natural language inference abilities. These results suggest that transfer learning may not be an effective tactic for improving misinformation detection systems [1].

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 misinformation detection benchmark of over 12,000 false and factual statements (Wang, 2017)

[1] For code, see https://github.com/haleyej/logical_fallacy_detection

2.1.1 SNLI Dataset

The SNLI coprus contains examples of logical entailment, contradiction, and neural statements (where one statement neither entails nor contradicts another). The dataset is available for download on online and has already been split into training, testing, and development sets. The 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 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:** 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 contains three different kinds of logical relationships: contradiction, neutral, and entailment. I only used examples of contradictions and entailment. This had two key advantages. First, it simplified the classification task. Secondly, it allowed our model to focus more on the class I was interested in — namely logical contradictions. The dataset was already in a clean format and did not need any preprocessing. Likewise, the dataset was roughly balanced. Due to computational limits, I was not able to use the full training set during fine-tuning. Instead, I randomly sampled 180,000 instances.

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," "half true," "barely true," "mostly true" and "true." Labels were manually assigned by a PolitiFact editor. For simplicity, I recoded these labels into a binary classification task. The labels "false" and "mostly false" where positive instances and all other labels were negative instances. 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 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 classifier 55.9% of instances correctly. 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 achieved 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 performance 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). Extensive work has been done leveraging logically fallacy detection in transfer learning in related domains, namely climate denial (Jin et al., 2022), propaganda detection (Oliinyk et al., 2020), and hyper partisan news detection (Kiesel et al., 2019). This work extends this methodology to a new downstream task — the LIAR benchmark dataset — to further explore it's effectiveness. Furthermore, I hypothesize that training a model on general natural language inference will help it generalize better to unseen text. The nature of fake news is constantly changing, so any feasible misinformation detection system must be able to make inferences about topics, speakers, and rhetorical styles it did not encounter during training.

This project uses multiple fine-tuning on a pre-trained DistilBERT language model [2]. In the first fine-tuning stage, the model will learn to classify logical entailment and contradiction. Then in the second stage, it will learn to classify misinformation. I hypothesize that transfer learning between the logical relationship classification task and misinformation detection task will improve the final model's accuracy over language models with no natural language inference abilities.

Because there are multiple fine-tuning steps, there is a risk than the new model's will loose its ability to precisely model language. To mitigate this, I used a low-rank adaptation of large language models (LoRA) during training, a procedure developed by Hu et al. (2021) to support efficient fine-tuning of LLMs without diverging too much from the original weights[3]. LoRA takes advantage of the

[2]The DistilBERT model was loaded in and fine-tuned using the transformers package. The library is avaliable here: https://huggingface.co/docs/transformers

[3]LoRa was implemented using the PEFT library (PEFT is an acronym for parameter efficient fine-tuning). The library is

fact that most neural networks over-parameterize: Aghajanyan et al. (2020) note that networks have an "intrinsic dimensionality," or the minimum dimensionality required for success on a task. This intrinsic dimensionality is smaller than the size of the actual weight matrix.

LoRA forces fine-tuning to occur at a lower rank. Instead of training the whole model, LoRA trains an adapter for the current layer. The weight matrix is decomposed 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.

Finally, I apply my logical inference model to misinformation detection. I make predictions on the LIAR benchmark test set to assess if transfer learning can aid in detecting false claims.

5 Evaluation & Results

5.1 SNLI Fine-Tuning

Hyperparameter	Value
Epochs	1
Batch size	8
Learning rate	1e-5
Sequence length	512 tokens
Weight decay	0.01
LoRA Rank	8
LoRA Alpha	16
LoRA Dropout	0.1

Table 1: SNLI fine-tuning hyperparameters

Figure 1: Evaluation loss every 5000 steps on SNLI fine-tuning

available here: https://huggingface.co/docs/peft

Due to computational limits, I was only able to fine-tune the DistilBERT model on the SNLI corpus for one epoch. The loss generally decreased after each training step, although it did not fully converge. This indicates that the pre-trained Disitl-BERT model was learning how to correctly detect logical entailment and contradiction and that performance would've likely improved with additional epochs.

The SNLI model was able to correctly classify 74.90% of logical relationships and achieved an F1 score of 74.76%. This represented a significant improvement over a simple Naive-Bayes baseline, which could only correctly categorize 54.19% of instance and had an F1 score of 62.68%. Table 1 details the hyperparameter configuration used to fine-tune the model.

5.2 LIAR Benchmark

Next, I fine-tuned both the logical-inference model and a base DistilBERT model on the LIAR benchmark's training set. The benchmark dataset only included a few thousands instances, so I was able to fine-tune for multiple epochs. 2 details the hyperparameter configuration used during training.

Hyperparameter	Value
Epochs	100
Batch size	32
Learning rate	1e-5
Sequence length	512 tokens
Weight decay	1e-4
LoRA Rank	8
LoRA Alpha	16
LoRA Dropout	0.15

Table 2: LIAR fine-tuning hyperparameters

The base DistilBERT model consistently had a lower loss throughout training. Performance on the test-set was near identical. Similarly, the base DistilBERT model also had a slightly higher accuracy (73.50% vs. 71.63%) and F1 score (64.26% vs. 63.95%). See figures 3 and 4.

Both the DistilBERT base and DistilBERT with SNLI fine-tuning models outperformed a simple Naive Bayes baseline with a bag-of-words representation, although not significantly [4]. The Naive Bayes classifier correctly categorized 64.49% of statements, achieving an F1 score of 57.16%. The

[4]Naive Bayes was implemented using the sklearn package https://scikit-learn.org/stable

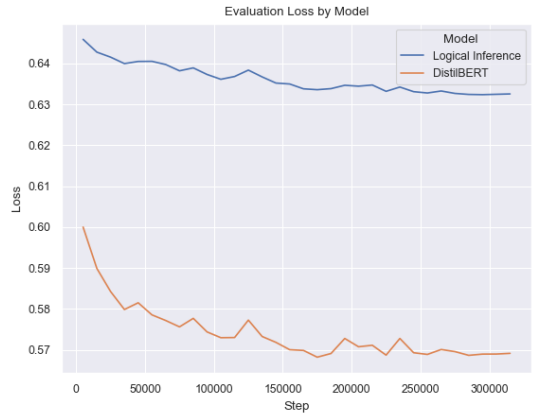

Figure 2: Evaluation loss of logical inference model vs. base DistilBERT model every 5000 steps on LIAR benchmark

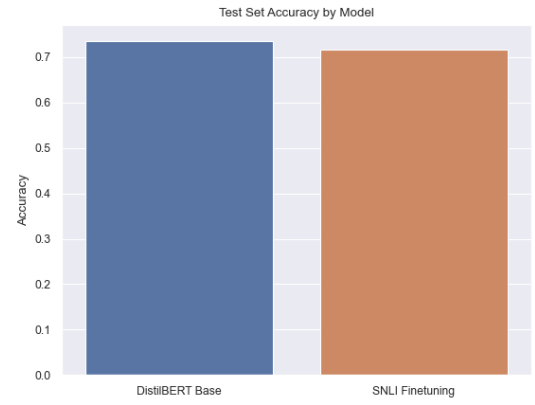

Figure 3: LIAR benchmark accuracy by model

results of all models on the LIAR benchmark are summarized in table 3.

6 Discussion

6.1 SNLI Fine-Tuning

The pre-trained DistilBERT model performed significantly better than a Naive Bayes baseline with a simple bag-of-words representation. In fact, the Naive Bayes classifier only correctly categorized 54.19% of instances. This was only slightly better than random guessing, as the subset of the SNLI dataset I used for training was evenly balanced between examples of entailment and contradiction. Nevertheless, these results were unsurprising, as neural networks are more powerful than statistical models and can learn more complex representations of logical relationships.

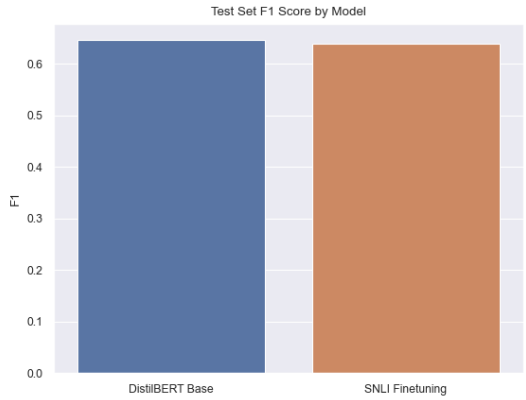

Figure 4: LIAR benchmark F1 score by model

6.2 LIAR Benchmark

The DistilBERT models showed more mixed performance on the LIAR Benchmark. Both models improved upon the Naive Bayes baseline, but results were more modest than on the logical inference task. Likewise, the DistilBERT base model slightly outperformed the DistilBERT model than was pre-trained on the SNLI dataset. Taken together, these results suggest that 1) both DistilBERT models were better at discriminating between misinformation and factual statements than the Naive Bayes baseline and 2) teaching the model logical inference abilities did not improve it's classification abilities.

As previously mentioned, I was only able to fine-tune the DistilBERT model on the SNLI dataset for one epoch due to computational limitations. This was not enough for the loss to converge. Given that transfer learning from natural language understanding to misinformation detection has been used successfully in other work (Oliinyk et al., 2020; Kiesel et al., 2019), I believe this project would've had better results if I would've been able to fine-tune for multiple epochs. Considered in the context of previous liteature, these results suggest that more extensive pre-training is needed to effectively leverage transfer learning from natural language inference for misinformation detection.

7 Conclusion

Misinformation detection is an open problem in natural language processing. Misinformation is highly contextual. Likewise, the content and rhetoric of misinformation is constantly changing, making it difficult for machine learning models to consis-

Model	Accuracy	F1
Naive Bayes	64.49%	57.16%
DistilBERT Base	73.50%	64.26%
DistilBERT with SNLI Fine-Tuning	71.63%	63.95%

Table 3: Summary of results on LIAR benchmark test set

tently achieve high accuracy. Previous work has leveraged transferring learning from natural language inference to improve misinformation detection systems. This is based on the notion that training a model to have general logical inference abilities will help them generalize better. Rather than learning what misinformation looks like in one dataset, the model will be able to asses claims and evidence more generally, allowing it to be deployed on a variety of downstream tasks.

This project contributes to a growing body of literature on misinformation detection, natural language inference, and transferring learning. These results help clarify when transfer learning can be successfully leveraged for misinformation detection. In particular, they suggested that more extensive pre-training is necessary to see meaningful improvements over neural networks without pre-training. More work is needed to fully understand if transfer learning from natural language understanding or other tasks can improve these systems.

8 Other Things I Tried

Initially, I wanted to add a third-fine tuning step on a dataset with examples of different kinds of logical fallacies. I found a dataset that contained examples of 14 different types of logical fallacies (e.g. faulty generalizations, ad hominem, ad populum). I hypothesized that teaching a model to detect different types of logical fallacies would lead to stronger natural language inference abilities than just learning to discriminate between entailment and contradiction.

I wrote code to process and pre-train with this data, but ultimately decided against adding this additional step. First, I was worried too many pre-training steps would cause the model's language abilities to degrade. Secondly, this added a lot of added complexity to the project for something that wouldn't clearly boost performance. I spent ~4 hours trying to develop a system that included 1) SNLI fine-tuning 2) logical fallacy fine-tuning and 3) misinformation detection, but wasn't seeing strong results. Due to time constraints, I decided to abandon this and focus on other components of my project.

9 Future Work

Future work may evaluation misinformation detection systems on datasets taken in different points in time and on different subject matters. This will help asses if logical inference pre-training actually improves the long-term generalizability of misinformation detection systems. Due to time constraints, I was only able to evaluate my models on one misinformation detection benchmark. While it did include statements from several years and on several topics, more evaluation is needed to truly understand the effectiveness of these systems.

Likewise, more pre-training on the SNLI corpus likely would've improved my results. Future work could 1) pre-train for more epochs and 2) pre-train on the full training set, rather than a random sample of 180,000 instances.

Overall this project went very smoothly and I didn't have major difficulties outside of the computational challenges detailed above.

10 Acknowledgement

Special thanks to Zachary Eichenberger, who suggested I use LoRA. Thank you to Bella Karduck and Coulton Theuer for useful feedback on earlier drafts of this project.

References

Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. 2020. Intrinsic dimensionality explains the effectiveness of language model fine-tuning.

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.