CS 584 Final: An analysis of roBERTa variants and their performance on Kaggle's "Contradictory, My Dear Watson" NLI task

BENNETT WOODS, Stevens Institute of Technology, USA

Additional Key Words and Phrases: NLP, NLI, roBERTa, kaggle

ACM Reference Format:
Bennett Woods. 2025. CS 584 Final: An analysis of roBERTa variants and their performance on Kaggle's "Contradictory, My Dear Watson" NLI task. 1, 1 (May 2025), 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Background and Motivation

Natural Language Inference (NLI) is a fundamental task in natural language understanding, requiring models to determine whether a hypothesis is entailed by, contradicts, or is neutral with respect to a given premise. Despite substantial advances in transformer-based models, their behavior across domains and tasks remains underexplored. This project investigates the comparative performance of several RoBERTa variants fine-tuned on the "Contradictory, My Dear Watson" (CMDW) Kaggle dataset, a multilingual NLI challenge with diverse linguistic inputs and label noise.

1.1 Kaggle Challenge

The focus of this paper is the Kaggle challenge "Contradictory, My Dear Watson" (CMDW). In it, Kaggle users are tasked with submitting models that are capable of determining the relationship between two sentences: entailment, neutral, and contradiction. Furthermore, the dataset is in 15 different languages.

2 Related Work

2.1 NLI

Natural Language Inference (NLI) is a core task in natural language understanding and has been widely benchmarked through datasets such as SNLI [1] and MultiNLI [8], which form the basis of evaluation suites like GLUE. To support multilingual evaluation, the XNLI corpus [3] extends MultiNLI across 15 languages and has driven work in cross-lingual generalization. BERT [4] and its multilingual variant (mBERT) demonstrated promising zero-shot transfer abilities on XNLI, but their performance is often limited in low-resource or structurally divergent languages. XLM-R [2], which scales up multilingual pretraining, has become a strong baseline, outperforming mBERT across high- and low-resource languages.

2.2 roBERTa

RoBERTa [6], a robustly optimized variant of BERT, achieves state-of-the-art results on English NLI tasks by training longer and removing the next-sentence prediction objective. It serves as a strong monolingual baseline in contrast to cross-lingual models like XLM-R. Fine-tuning RoBERTa on large-scale NLI datasets (e.g., MNLI or ANLI) has been

Author's Contact Information: Bennett Woods, Stevens Institute of Technology, Hoboken, New Jersey, USA, bwoods@stevens.edu.

Manuscript submitted to ACM

shown to improve generalization, especially when used as an intermediate task [7]. This study builds on those findings by comparing RoBERTa-base, XLM-R, and a RoBERTa model fine-tuned on a large NLI corpus within the multilingual, noisy environment posed by the "Contradictory, My Dear Watson" dataset.

3 Approach

This project approaches Natural Language Inference (NLI) as a multilingual generalization challenge. Rather than building a novel architecture, we aim to analyze how different pretrained Transformer models perform when fine-tuned on a multilingual and noisy NLI dataset. Specifically, we evaluate three models: (1) `roberta-base`, a widely used monolingual English model; (2) `xlm-roberta-base`, a multilingual model pretrained on 100+ languages; and (3) a version of `roberta-base` fine-tuned on a large English NLI corpus (MNLI). my objective is to compare their accuracy, generalization, and error characteristics when trained on the "Contradictory, My Dear Watson" dataset from Kaggle, which features data in 15 languages and presents label noise and lexical variation.

My hypothesis is that multilingual pretraining (XLM-R) or task-specific pre-finetuning (MNLI-finetuned RoBERTa) may provide generalization benefits over training a standard English model from scratch. This setup allows us to examine trade-offs between model size, pretraining regime, and transfer effectiveness. We control for hyperparameters, training infrastructure, and data splits to isolate the effects of model choice. All models were fine-tuned using HuggingFace's `Trainer` API with identical configurations and evaluated using accuracy and leaderboard score. This analysis contributes to understanding whether pretrained multilingual or task-specialized models offer an advantage on real-world NLI tasks with diverse linguistic inputs.

4 Experimental Design

my experiments were conducted on the Kaggle "Contradictory, My Dear Watson" dataset, a multilingual NLI corpus covering 15 languages. Since the Kaggle test set lacks ground truth labels, we partitioned the provided training set into a 90/10 train-validation split for hyperparameter tuning and evaluation. All models were trained using HuggingFace's `transformers` [9] and `datasets` [5] libraries, with tokenization handled by `tokenizers`. Fine-tuning and evaluation were performed using the `Trainer` API to ensure consistency across all model runs.

We compared three models: `roberta-base`, `xlm-roberta-base`, and a publicly available `roberta-large` model fine-tuned on the MNLI corpus. Each model was fine-tuned on a Google Colab Pro instance with A100 GPUs. We performed a small hyperparameter search over learning rate $\{1e{-}5, 2e{-}5\}$ and weight decay $\{0.0, 0.1\}$, using early stopping based on validation loss. All models were trained with a batch size of 16, up to 3 epochs. The best checkpoint for each model was uploaded to the HuggingFace Hub for later use.

Predictions for the Kaggle competition were generated in a separate Kaggle notebook using P100 GPUs, loading each final model checkpoint from HuggingFace. The main evaluation metric was classification accuracy, measured on the held-out validation set and reflected in the public leaderboard scores. No additional preprocessing, label smoothing, or data augmentation techniques were applied.

5 Experimental Results

Table 2 summarizes the leaderboard accuracy of all three models on the "Contradictory, My Dear Watson" task. The MNLI-finetuned `roberta-large` model achieved the highest score (73.0%), followed by `xlm-roberta-base` (70.7%) and then the baseline `roberta-base` trained from scratch (64.3%). These results support my hypothesis that domain-adaptive pretraining (in this case, on a large English NLI dataset) improves generalization on multilingual NLI tasks. Surprisingly,

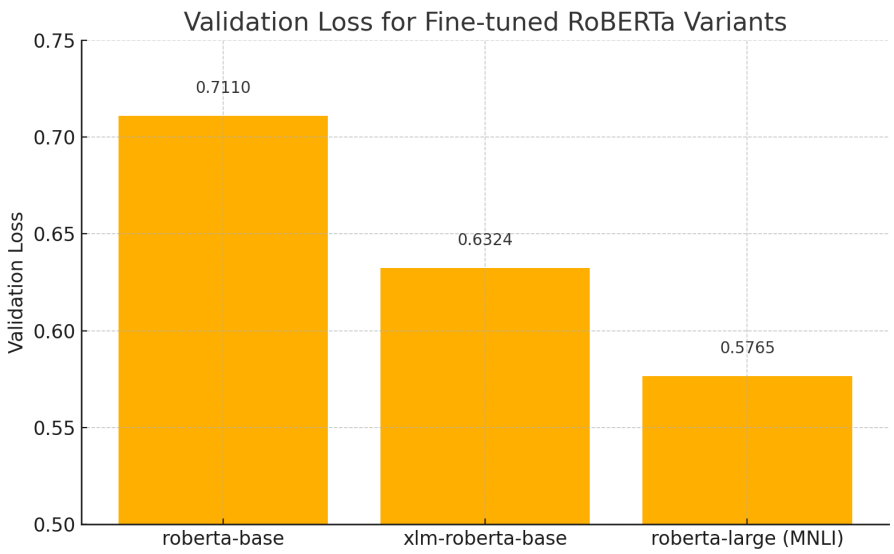

Fig. 1. Validation loss for each fine-tuned model. Lower is better. The MNLI-finetuned RoBERTa achieves the best validation performance.

Table 1. Model Parameter Comparison

Model	Parameters
`roberta-base`	125M
`xlm-roberta-base`	270M
`roberta-large (MNLI)`	355M

despite being monolingual, the MNLI-finetuned model outperformed the multilingual XLM-R variant, suggesting that task alignment may offer more benefit than language coverage in this case.

Training dynamics revealed notable differences in model behavior. The MNLI-finetuned model converged in just one epoch and was the only model to stop early, indicating it was already well-prepared for the NLI task. The baseline RoBERTa model required two epochs before early stopping, while the XLM-R model trained for the full five epochs, likely due to its larger cross-lingual capacity. Validation loss curves indicated mild overfitting in the monolingual models, whereas XLM-R was more stable across epochs. Although we did not perform language-wise error analysis, future work could investigate whether performance differences across languages contributed to these rankings, particularly given the known challenges of cross-lingual transfer [2, 3].

Table 2. Leaderboard Accuracy on the CMDW Kaggle Task

Model	Leaderboard Accuracy
`roberta-large (MNLI-finetuned)`	0.730
`xlm-roberta-base`	0.707
`roberta-base`	0.643

6 Limitations.

This study used a small hyperparameter search space and did not conduct detailed language-wise evaluation due to time constraints. Evaluation relied on Kaggle leaderboard scores, which reflect a mixture of language-specific test distributions. Further probing is needed to assess robustness and cross-lingual consistency.

7 Conclusion

This project compared three Transformer-based models—`roberta-base`, `xlm-roberta-base`, and an MNLI-finetuned `roberta-large`—on the multilingual NLI task presented by the "Contradictory, My Dear Watson" Kaggle competition. my results show that task-specific pretraining on NLI data significantly boosts performance, even in multilingual settings, outperforming both multilingual and general-purpose baselines. These findings suggest that domain alignment may be more impactful than multilingual pretraining when dealing with label noise and semantic variability in NLI tasks.

8 Future Work

Future work could extend this analysis along several directions. First, a detailed breakdown of performance across individual languages would provide insight into cross-lingual generalization gaps. Second, training strategies such as intermediate task fine-tuning (e.g., via ANLI or XNLI), language-specific adapters, or multi-task learning could be investigated to improve multilingual robustness. Finally, qualitative error analysis and adversarial testing may help reveal systematic weaknesses in how these models handle semantic ambiguity, contributing to a deeper understanding of NLI behavior across linguistic and cultural boundaries.

9 Supplementary Material

Here are the links to the various resources I used:

- HuggingFace model (MNLI-finetuned RoBERTa-large): https://huggingface.co/Woodsii/FacebookAI_roberta-large-mnli
- HuggingFace model (XLM-roberta-base fine-tuned on CMDW): https://huggingface.co/Woodsii/xlm-roberta-base
- HuggingFace model (roberta-base fine-tuned on CMDW): https://huggingface.co/Woodsii/roberta-basee

References

[1] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 632–642.

[2] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 8440–8451.

[3] Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2475–2485.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4171–4186.

[5] Quentin Lhoest, Luke Vilnis, Kushal Patil, Julien Chaumond, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Teven Le Scao, Sylvain Drame, Julien Plu, et al. 2021. Datasets: A Community Library for Natural Language Processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 175–184.

[6] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. In *arXiv preprint arXiv:1907.11692*.

[7] Jason Phang, Thibault Fevry, and Samuel R Bowman. 2018. Sentence encoders on STILTs: Supplementary training on intermediate labeled-data tasks. In *arXiv preprint arXiv:1811.01088*.

[8] Adina Williams, Nikita Nangia, and Samuel R Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1112–1122.

[9] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 38–45.