Issues with Entailment-based Zero-shot Text Classification

Tingting Ma[1,2]*, **Jin-Ge Yao**[2], **Chin-Yew Lin**[2], **Tiejun Zhao**[1]
[1]Harbin Institute of Technology, Harbin, China
[2]Microsoft Research Asia
hittingtingma@gmail.com
{jinge.yao, cyl}@microsoft.com, tjzhao@hit.edu.cn

Abstract

The general format of natural language inference (NLI) makes it tempting to be used for zero-shot text classification by casting any target label into a sentence of hypothesis and verifying whether or not it could be entailed by the input, aiming at generic classification applicable on any specified label space. In this opinion piece, we point out a few overlooked issues that are yet to be discussed in this line of work. We observe huge variance across different classification datasets amongst standard BERT-based NLI models and surprisingly find that pre-trained BERT without any fine-tuning can yield competitive performance against BERT fine-tuned for NLI. With the concern that these models heavily rely on spurious lexical patterns for prediction, we also experiment with preliminary approaches for more robust NLI, but the results are in general negative. Our observations reveal implicit but challenging difficulties in entailment-based zero-shot text classification.

1 Introduction

Natural language inference (NLI, Bowman et al., 2015), also known as recognizing *textual entailment* (RTE, Condoravdi et al., 2003; Dagan et al., 2005), is normally formatted as the task of determining whether or not a **premise** sentence semantically entails a **hypothesis** sentence. The generality of the task format has aroused some recent studies to apply NLI models for various downstream applications (Poliak et al., 2018), and more recently text classification (Yin et al., 2019, 2020), making them generally-applicable solutions along with all those similar attempts to build a universal framework for various NLP tasks (Kumar et al., 2016; Raffel et al., 2020, *inter alia*). Text classification is then reduced to textual entailment by setting the input sentence as the premise and simultaneously casting the candidate label into a hypothesis sentence using pre-defined templates or lexical definitions from WordNet. Once we have any pre-trained NLI models at hand, *zero-shot text classification* under any specified label space is enabled for free without the need to collect annotated data. With contextualized representation based on pre-trained language models such as BERT (Devlin et al., 2019), NLI performance has been drastically improved. Promising empirical results have been shown on various text classification benchmarks that vary across topic classification, emotion classification, and situation classification, outperforming earlier standard approaches (Chang et al., 2008) or simple scoring schemes derived from distributional similarity (Mikolov et al., 2013).

However, such generality is conceptually contradictory with the specificity of text classification in many practical scenarios. In this opinion piece, we conduct extended analysis on the recent attempts (Yin et al., 2019) and point out some implicit issues under entailment-based zero-shot text classification that are overlooked in this line of work. We experiment with additional classification datasets and observe huge variance across them amongst standard BERT-based NLI models. More surprisingly, we find that raw BERT models without fine-tuning can sometimes yield more competitive results. We also experiment with preliminary approaches for improving the robustness of NLI models, but only to find negative results in general. Our observations reveal implicit but massive difficulties in building a successful general-purpose zero-shot text classifier based on text entailment models.

2 Our Investigation and Implied Issues

We attempt at re-examining the earlier study (Yin et al., 2019) with extended analysis to help estab-

*Work during internship at Microsoft Research Asia.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Short Papers), pages 786–796
August 1–6, 2021.

lish a better understanding of zero-shot text classification based on textual entailment. Our focus is to check **how well the models pre-trained for NLI could generalize to the prediction of unseen categories**, which is the major target of zero-shot classification. We did not study the setting that test set also include labels seen in training, commonly phrased as *generalized zero-shot learning* (Xian et al., 2018) and referred to as the *label-partially-unseen* setting by Yin et al. (2019). That setting strongly assumes that a bunch of in-domain data for a number of classes are available already. [1]

2.1 Basic setup

2.1.1 Text classification datasets

As an attempt to study zero-shot text classification in conceptually different and diverse aspects, Yin et al. (2019) experimented with three instances:

Topic classification : The Yahoo! Answers dataset from Zhang et al. (2015) with 10 categories.

Emotion classification : The Unify Emotion dataset (Bostan and Klinger, 2018) with 9 emotion types and a `none` label if no emotion applies.

Situation classification : The Situation Typing dataset (Mayhew et al., 2019) with 11 situation types and instances and an extra type `none`.

Additionally, we extend our experiments with the test sets from the following datasets:

Snips : A popular dataset[2] for *intent detection* collected from the Snips personal voice assistant (Coucke et al., 2018), with seven intent labels.

AG's news : To further study the models on *topic classification* in a different genre, we additionally use the English news data from (Zhang et al., 2015) that consists of four types of articles: `World`, `Sports`, `Business`, `Sci/Tech`.

SST-2 : The Stanford Sentiment Treebank dataset[3] processed by Socher et al. (2013) for *sentiment polarity classification* with binary labels (`positive` and `negative`).

[1] Another reason for not studying on this setting is that the split of development set and test set in (Yin et al., 2019) contain the same label space, which is flawed to be used for any claim on the performance of "unseen labels".

[2] https://github.com/snipsco/snips-nlu

[3] For SST-2 we follow Zhang et al. (2021) and Gao et al. (2021) to use the development set from GLUE for testing.

2.1.2 Experimented systems

To study entailment-based approaches, we use the models released by Yin et al. (2019) which are `bert-base-uncased` models pretrained on GLUE RTE (Dagan et al., 2005; Wang et al., 2019b), MNLI (Williams et al., 2018), and FEVER (Thorne et al., 2018), respectively. We reuse the same scheme for mapping labels into hypotheses using templates and WordNet definition for all datasets[4], as well as the same mechanism for producing final predictions. We leave more implementation details to the Appendix.

We keep reporting results from these baselines following Yin et al. (2019) for reference:

- **Majority**: Output the most frequent label.
- **Word2Vec**: Using the average word embeddings to vectorize input and labels, output label with maximum cosine similarity.
- **ESA**: Representing the text and label in the Wikipedia concept vector space. Using the implementation[5] from Chang et al. (2008).

Moreover, due to the obvious variance in performance for models trained on different NLI datasets, we are also tempted to check how much the performance might degrade when given no NLI data at all for fine-tuning. This corresponds to naively using a raw BERT model which has been pre-trained for *next sentence prediction* (NSP). For consistency, we use the same premises and hypotheses as the delegate for label names and templates to formulate the sentence pair classification. Since NSP is not predicting for a directional semantic entailment, we also try a variant with all pairs reversed, i.e., setting all hypothesis sentences ahead of premises as input, denoted as NSP(Reverse).

2.2 Results and further analysis

The results from all systems on different datasets are displayed in Table 1, including an additional group for MNLI results as we found an even better run overall in our experiments. There are some interesting observations emerge from our extended experiments and analysis.

[4] For newly introduced datasets we follow the similar strategy to prepare for the hypothesis templates.

[5] https://github.com/CogComp/cogcomp-nlp/tree/master/dataless-classifier

	Topic (Yahoo)	Emotion	Situation	AG's News	SST-2	Snips
Majority	10.0	5.9	11.0	25.0	50.9	17.7
ESA	28.6	8.0	26.0	73.3	55.5	63.4
Word2Vec	35.7	6.9	15.6	44.1	53.7	63.6
RTE (Yin et al., 2019)	43.8	12.6	**37.2**	56.7	52.5	56.4
FEVER (Yin et al., 2019)	40.1	**24.7**	21.0	**78.3**	71.7	69.4
MNLI (Yin et al., 2019)	37.9	22.3	15.4	72.4	67.5	77.6
MNLI (our best overall run)	49.1	19.9	14.5	77.7	67.5	77.6
NSP (Reverse)	**53.1**	16.1	19.9	**78.3**	**79.7**	**81.3**
NSP	50.6	16.5	25.8	72.1	73.9	73.4

Table 1: Text classification results. We report label-wise weighted F1 for emotion and situtation datasets, and accuracy for the others. Reported results from (Yin et al., 2019) have been reproduced from their released models.

2.2.1 How much have NLI data contributed?

The big difference from various NLI datasets drives us to try a raw BERT without fine-tuning on any NLI data, i.e., merely relying on NSP pre-training for sentence pair classification. The results are shown at the bottom two rows in Table 1, which turn out to be surprisingly strong, especially on topic classification, intent classification, and binary sentiment classification.

We conjecture that the raw BERT model has already acquired certain ability of topic distinction and sentiment polarity due to the construction of positive and negative sentence pairs in NSP pre-training to detect pairwise coherence. In this way, NSP could serve as a non-trivial, strong alternative baseline for zero-shot text classification scenarios where the target labels are semantically more concrete (e.g., topics) or more frequently appeared (e.g., words expressing sentiment). In such scenarios, fine-tuning on limited NLI data could weaken the semantic coherence acquired from the raw BERT pre-trained on generic-domain corpora, especially now that fine-tuned models have utilized many spurious lexical cooccurrence features as shown in many similar sentence pair classification models (Feng et al., 2019; Niven and Kao, 2019), possibly due to the inherent lexical bias from the current NLI datasets collected from crowd workers. [6] Readers who are curious about more details on this problem can refer to our qualitative analysis in the Appendix which could hopefully help establish a slightly better sense on the behavioral difference introduced by NLI fine-tuning.

On the other hand, fine-tuning on NLI data might seem to be marginally helpful for more abstract cases such as emotion and situation typing, but the performance metrics are in fact pathetically disappointing across all systems.

2.2.2 How stable are these NLI models?

Apart from the obvious difference caused by different training data, there underlies a more serious concern: the *discrepancy* between the training task (NLI) and the target usage (classification). The gap in task formatting (and henceforth data distribution) naturally raises a question: *do NLI models with similar in-domain performance generalize similarly for text classification?*

We train NLI models on the largest MNLI dataset with varied hyperparameter settings and random seeds, and keep models achieving similarly strong in-domain generalization performance as measured by the early-stopping dev set performance. Results are listed in Table 2, where the absolute differences between the worst and the best are large, especially on classifying topic or intent. We observe even worse trends on other smaller NLI datasets (see Appendix). These results are consistent with recent studies within the scope of NLI reporting that BERT instances which achieve similar performance metrics on standard NLI datasets could have huge variance in out-of-distribution generalization or linguistic stress testing (McCoy et al., 2020; Zhou et al., 2020; Geiger et al., 2020), while providing another instance of the underspecification problem in modern machine learning (D'Amour et al., 2020).

As a verification, we also try to tune the models for different development sets that better characterize the generalization behavior for zero-shot

[6] Some readers might guess that other NLI datasets collected via a more careful process (Jiang and de Marneffe, 2019; Eisenschlos et al., 2021) might partially mitigate the bias appearing from crowdsourced annotation, but this does not mean that such better intended datasets can be free from statistically biased lexical distributions with coincidental cooccurrences that could be utilized by our strong data-fitting models during fine-tuning (Geirhos et al., 2020; Du et al., 2021). Our additional results described in the Appendix do not seem to be promising on this direction towards better NLI data.

Dataset	Average	Std	Min	Max
MNLI dev set	90.5	0.3	90.0	90.8
Yahoo	39.0	10.5	26.9	50.2
Emotion	18.1	2.0	15.7	20.5
Situation	16.2	1.5	14.5	18.7
AGNews	63.7	11.0	50.0	77.7
SST-2	68.6	2.0	66.1	70.9
Snips	74.1	3.9	68.4	77.6

Table 2: Results of five runs of BERT fine-tuned on MNLI and tested on classification datasets

classification. We reorganize the splitted development set and the test set of the topic classification datasets (Yahoo and AG's News) to make sure they do not have overlapped classes.[7] The new results are shown in Table 3, where we can clearly see more stable generalization performance. This observation necessitates that a certain amount of annotated data for targeted classification already existed, making NLI models difficult to apply in practice. Results in this part reveals that text classification via NLI is asking for out-of-distribution generalization, a property that current NLI models rarely have, henceforth susceptible to huge *instability*.

Dataset	Average	Std	Min	Max
Yahoo-dev	52.7	2.6	49.1	56.2
Yahoo-test	48.1	2.7	44.2	51.7
AGNews-dev	79.0	6.9	72.1	89.1
AGNews-test	73.8	3.8	69.6	77.4

Table 3: Results of five runs for training BERT on MNLI with model selection via target domain dev set

2.2.3 Is more robust NLI helpful?

Previous studies have raised concerns on that the current NLI models heavily rely on spurious lexical overlap patterns (Sanchez et al., 2018; Naik et al., 2018; McCoy et al., 2019, *inter alia*). For analytical purposes, we randomly permute the tokens of each input instance to see how much the predictions might change. Results shown in Table 4 suggest that shuffling the input tokens does not affect the model performance by much, which is consistent with similar recent findings (Gupta et al., 2021; Sinha et al., 2021). This reveals a concern that all these models might just predict with shallow lexical patterns that may not be robust for more semantically abstractive input instances.

There have been a few recent attempts trying to remove the shallow overlap bias for NLI model

[7]Details are described in the Appendix.

Model	Yahoo	AGNews	SST-2
NSP(Reverse)	-5.1 / 67.2	+0.4 / 82.7	-13.5 / 75.9
RTE	-2.0 / 77.5	+0.3 / 90.0	+0.6 / 94.5
FEVER	-7.2 / 64.6	+0.5 / 90.6	-9.5 / 82.3
MNLI	+1.6 / 54.8	+2.7 / 84.9	-6.4 / 84.4
Random	- / 10.0	- / 25.0	- / 50.0

Table 4: Results of shuffling perturbation. In each cell: the change of accuracy after input shuffling, followed by the percentage of examples where the predictions do not change. All these results are reported as the average score of five different random shuffles.

training. We experiment with three schemes on the MNLI data to see whether they could lead to better generalization of zero-shot classification: (1) *Data augmentation* with syntactic transformations (Min et al., 2020)[8], denoted as *DA*, (2) *Instance reweighting* following Clark et al. (2019) that reweights each example with one minus the probability a bias-only model assigns the correct label, denoted as *RW*, and (3) The *bias product* method (Clark et al., 2019) that ensembles a bias-only model via a product of experts, denoted as *BP*, which is essentially the same as its concurrent work via fitting the residual of the biased models (He et al., 2019). There exist additional solutions with richer details such as multi-task learning (Tu et al., 2020) where proper auxiliary tasks could be identified to improve robustness. We plan to explore more in this line in our more extensive future study.

The results are shown in Table 5 . All the three debiasing methods improve the NLI performance on the HANS dataset (McCoy et al., 2019) for robustness testing, indicating that the debiased models overcome the word overlap heuristics to some extend. In general, we do not observe any real improvement other than the neglectable gains on emotion and situation datasets where the original performance is pathetically low.

	HANS	Yahoo	Emo.	Situ.	AG	SST	Snips
MNLI	53.0	**49.1**	19.9	14.5	**77.7**	67.5	**77.6**
w/ DA	**67.3**	47.3	18.0	16.3	74.3	**73.1**	76.6
w/ RW	64.5	43.4	21.8	**23.5**	71.7	68.6	71.6
w/ BP	65.4	48.5	**23.0**	22.3	75.6	69.8	72.7

Table 5: Results of NLI debiasing based on MNLI

[8]We directly use the data released at https://github.com/Aatlantise/syntactic-augmentation-nli

3 Conclusion and Discussion

We investigate entailment-based zero-shot text classification further with extended analysis, uncovering the following overlooked issues:

- Raw BERT models trained for next sentence prediction are surprisingly strong baselines and NLI fine-tuning does not bring performance gain on many classification datasets.
- Large variance on different classification scenarios and instability to different runs, still requiring annotated data (at least used for validation) to stablize generalization performance.
- NLI models usually rely heavily on shallow lexical patterns, which hampers generalization as required by text classification, and currently more robust NLI methods might not help.

Our observations reveal *implicit but massive difficulties in building a usable zero-shot text classifier based on text entailment models*. Given the difficulty of NLI data collection that aims at out-of-domain generalization or transfer learning (Bowman et al., 2020), we question the feasibility of this setup in the current progress of language technology. Before significant progress in language understanding and reasoning, it seems more promising to consider alternative schemes built on explicit external knowledge (Zellers and Choi, 2017; Rios and Kavuluru, 2018; Zhang et al., 2019) or more crafted usage of pre-trained models that hopefully have captured more comprehensive semantic coverage and better compositionality from large corpora or grounded texts (Meng et al., 2020; Brown et al., 2020; Radford et al., 2021).

This study also implies *the huge difficulty for benchmarking zero-shot text classification without any further restriction on the task setting*. The three datasets used by Yin et al. (2019) were originally intended for diverse coverage but are not sufficient to draw consistent conclusions as we have shown. We suggest future studies on zero-shot text classification either conduct experiments over even more diverse classification scenarios to verify any claimed generality, or directly focus on more specific task settings and verify claims within a smaller but clearer scope such as zero-shot intent classification or zero-shot situation typing for more reliable results with less instability, and perhaps based on more carefully curated data (Rogers, 2021).

Acknowledgments

We thank all the anonymous reviewers for their helpful comments on our submitted draft. The empirical studies conducted in this work were mostly based on the open-source repositories on GitHub from other papers as described earlier. We thank their original authors for sharing their implementation and we also publicly release our experimental scripts on GitHub[9].

References

Laura-Ana-Maria Bostan and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119, Santa Fe, New Mexico, USA.

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*, pages 632–642, Lisbon, Portugal.

Samuel R. Bowman, Jennimaria Palomaki, Livio Baldini Soares, and Emily Pitler. 2020. New protocols and negative results for textual entailment data collection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8203–8214, Online.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *NeurIPS 2020*.

Ming-Wei Chang, Lev Ratinov, Dan Roth, and Vivek Srikumar. 2008. Importance of semantic representation: Dataless classification. In *AAAI'08*.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4069–4082, Hong Kong, China. Association for Computational Linguistics.

[9] https://github.com/mtt1998/issues-nli

Cleo Condoravdi, Dick Crouch, Valeria de Paiva, Reinhard Stolle, and Daniel G. Bobrow. 2003. Entailment, intensionality and text understanding. In *Proceedings of the HLT-NAACL 2003 Workshop on Text Meaning*, pages 38–45.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges, Evaluating Predictive Uncertainty, Visual Object Classification and Recognizing Textual Entailment, First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April 11-13, 2005, Revised Selected Papers*, pages 177–190. Springer.

Alexander D'Amour, Katherine A. Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yi-An Ma, Cory Y. McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. 2020. Underspecification presents challenges for credibility in modern machine learning. *CoRR*, abs/2011.03395.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mengnan Du, Varun Manjunatha, Rajiv Jain, Ruchi Deshpande, Franck Dernoncourt, Jiuxiang Gu, Tong Sun, and Xia Hu. 2021. Towards interpreting and mitigating shortcut learning behavior of NLU models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 915–929, Online. Association for Computational Linguistics.

Julian Eisenschlos, Bhuwan Dhingra, Jannis Bulian, Benjamin Börschinger, and Jordan Boyd-Graber. 2021. Fool me twice: Entailment from Wikipedia gamification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 352–365, Online. Association for Computational Linguistics.

Shi Feng, Eric Wallace, and Jordan Boyd-Graber. 2019. Misleading failures of partial-input baselines. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5533–5538, Florence, Italy. Association for Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *ACL-IJCNLP 2021*.

Atticus Geiger, Kyle Richardson, and Christopher Potts. 2020. Neural natural language inference models partially embed theories of lexical entailment and negation. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 163–173, Online. Association for Computational Linguistics.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Ashim Gupta, Giorgi Kvernadze, and Vivek Srikumar. 2021. BERT & family eat word salad: Experiments with text understanding. In *35th AAAI Conference on Artificial Intelligence (AAAI-21)*.

He He, Sheng Zha, and Haohan Wang. 2019. Unlearn dataset bias in natural language inference by fitting the residual. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 132–142, Hong Kong, China. Association for Computational Linguistics.

Nanjiang Jiang and Marie-Catherine de Marneffe. 2019. Evaluating BERT for natural language inference: A case study on the CommitmentBank. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6086–6091, Hong Kong, China.

Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher. 2016. Ask me anything: Dynamic memory networks for natural language processing. In *Proceedings of The 33rd International Conference on Machine Learning (ICML)*, pages 1378–1387.

Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The CommitmentBank: Investigating projection in naturally occurring discourse. *Proceedings of Sinn und Bedeutung*, 23(2):107–124.

Stephen Mayhew, Tatiana Tsygankova, Francesca Marini, Zihan Wang, Jane Lee, Xiaodong Yu, Xingyu Fu, Weijia Shi, Zian Zhao, Wenpeng Yin, Karthikeyan K, Jamaal Hay, Michael Shur, Jennifer Sheffield, and Dan Roth. 2019. University of Pennsylvania LoReHLT 2019 Submission. Technical report.

R. Thomas McCoy, Junghyun Min, and Tal Linzen. 2020. BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 217–227, Online. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Chenyan Xiong, Heng Ji, Chao Zhang, and Jiawei Han. 2020. Text classification using label names only: A language model self-training approach. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9006–9017, Online. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26, pages 3111–3119. Curran Associates, Inc.

Junghyun Min, R. Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. Syntactic data augmentation increases robustness to inference heuristics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2339–2352, Online. Association for Computational Linguistics.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018. Collecting diverse natural language inference problems for sentence representation evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 67–81, Brussels, Belgium. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krüger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *OpenAI Technical Report*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Anthony Rios and Ramakanth Kavuluru. 2018. Few-shot and zero-shot multi-label learning for structured label spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3132–3142, Brussels, Belgium. Association for Computational Linguistics.

Anna Rogers. 2021. Changing the world by changing the data. In *ACL-IJCNLP 2021*.

Ivan Sanchez, Jeff Mitchell, and Sebastian Riedel. 2018. Behavior analysis of NLI models: Uncovering the influence of three factors on robustness. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1975–1985, New Orleans, Louisiana.

Koustuv Sinha, Prasanna Parthasarathi, Joelle Pineau, and Adina Williams. 2021. Unnatural language inference. *arXiv preprint arXiv:2101.00010*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana.

Lifu Tu, Garima Lalwani, Spandana Gella, and He He. 2020. An empirical study on robustness to spurious correlations using pre-trained language models. *Transactions of the Association for Computational Linguistics*, 8:621–633.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019a. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *NeurIPS 2019*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *ICLR*.

Adina Williams, Nikita Nangia, and Samuel 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, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. 2018. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, Hong Kong, China.

Wenpeng Yin, Nazneen Fatema Rajani, Dragomir Radev, Richard Socher, and Caiming Xiong. 2020. Universal natural language processing with limited annotations: Try few-shot textual entailment as a start. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8229–8239, Online.

Rowan Zellers and Yejin Choi. 2017. Zero-shot activity recognition with verb attribute induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 946–958, Copenhagen, Denmark. Association for Computational Linguistics.

Jingqing Zhang, Piyawat Lertvittayakumjorn, and Yike Guo. 2019. Integrating semantic knowledge to tackle zero-shot text classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1031–1040, Minneapolis, Minnesota.

Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q Weinberger, and Yoav Artzi. 2021. Revisiting few-sample BERT fine-tuning. In *International Conference on Learning Representations*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *NeurIPS 2015*.

Xiang Zhou, Yixin Nie, Hao Tan, and Mohit Bansal. 2020. The curse of performance instability in analysis datasets: Consequences, source, and suggestions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8215–8228, Online. Association for Computational Linguistics.

A Appendix

A.1 Additional Experimental Details

Templates for generating hypothesis For Yahoo, Emotion, and Situation datasets, we followed Yin et al. (2019) and just explored the label names and WordNet definition accompanied with a template[10] to convert labels to hypotheses for entailment-based models. When applying NSP, we only used label names to generate hypotheses as we did not observe real improvement from using WordNet definitions in our preliminary experiments. For AGNews, SST-2, and Snips, we simply used the label names to fill the templates. The templates we used are given in Table A.1.

Other implementation details For all experiments, we train BERT models by using *bert-base-uncased* version and code from the HuggingFace library (Wolf et al., 2019). We used the same prediction strategy as Yin et al. (2019): we pick the label with the maximal probability in single-label scenarios while choosing all the labels with "next sentence" decision in multi-label cases for both NSP and NSP(Reverse) baselines.

Label spaces of classification The labels of each dataset we used are listed in Table A.2.

[10] https://github.com/yinwenpeng/BenchmarkingZeroShot

Dataset	Template	Label to words mapping
Yahoo	It is related with [LABEL] .	[Sports]: sports, [Society & Culture]: society or culture, etc.
Emotion	This person feels [LABEL] .	[sadness]: sad, [anger]: angry, [guilt]: guilty, etc.
Situation	The people there need [LABLE] .	[shelter]: shelter, [utilities]: utilities, etc.
AGNews	It is related with [LABEL] .	[Sci/Tec]: technology, [Business]: business, etc.
SST-2	The movie is [LABEL] .	[positive]: great , [negative]: terrible
Snips	I want to [LABEL] .	[RateBook] : rate a book, [SearchCreativeWork]: search creative work, etc.

Table A.1: Templates used for each dataset. For Topic Emotion and Situation dataset, we also use the WordNet definitions following Yin et al. (2019)

Dataset	Labels
Yahoo	Society & Culture, Science & Mathematics, Health, Education & Reference, Computers & Internet, Sports, Business & Finance, Entertainment & Music, Family & Relationships, Politics & Government
Emotion	sadness, joy, anger, disgust, fear, surprise, shame, guilt, love, none
Situation	search, evacuate, infrastructure, utilities, water, shelter, medical assistance, food, crimeviolence, terrorism, regime change, none
AGNews	World, Sports, Business, Sci/Tech.
SST-2	Positive, Negative
Snips	RateBook, SearchScreeningEvent, PlayMusic, GetWeather, SearchCreativeWork, AddToPlaylist, BookRestaurant,

Table A.2: The label names of the evaluation datasets.

Additional results on CommitmentBank We finetune BERT on the CommitmentBank dataset (de Marneffe et al., 2019; Wang et al., 2019a) converted into the NLI format (Jiang and de Marneffe, 2019), denoted as CB. Following Wang et al. (2019a), we also try to pretrain BERT on MNLI dataset before finetuning on CommitmentBank, called MNLI+CB. In our experiments, we found both two models trained on CB did not show a better performance compared to model trained on other NLI datasets, especially on Yahoo and AGNews (19.9% accuracy on Yahoo for CB and 17.8% accuracy on Yahoo for MNLI+CB). This indicates that the finetuned BERT models may still focus on features that are beneficial for NLI performance, while losing the topic discriminability.

A.2 Qualitative Analysis

Table 1 shows that NSP(reverse) achieves better performance than NSP on several datasets. This could be related to the templates we used for generating previous or next sentences. For example, for the input *"play the god that failed on vimeo"* with label "PlayMusic", NSP(Reverse) predicts "PlayMusic" while NSP predicts "AddToPlaylist". It is a more natural expression for *"I want to play music. play the god that failed on vimeo"* than *"play the god that failed on vimeo. I want to play music"*. Among the entailment models, We find the RTE-based model performs best on situation dataset. The main class of situation dataset is the "none" label. As shown in Figure A.1, we find RTE-based model performs best on "none" label. Actually, if we calculate the average number of prediction labels each instance, we find NSP, NSP(Reverse), and FEVER's average prediction label number per instance is about 6.2 to 8.3, while RTE and MNLI's average number is about 1, which is closer to the average number of gold labels per instance. The implies NSP is not good at identifying the "none" label since the condition of predicting "entailment" (a premise entails its hypothesis) is more strict than predicting a "next sentence" label. For SST-2, we observe that all three entailment models tend to mislabel sentences with "negative" label as "positive". This may be attributed to the label word distribution in NLI datasets. We find the keyword "great" for positive label is much more frequently occurred than the keyword "terrible" for negative label in all the three NLI datasets.

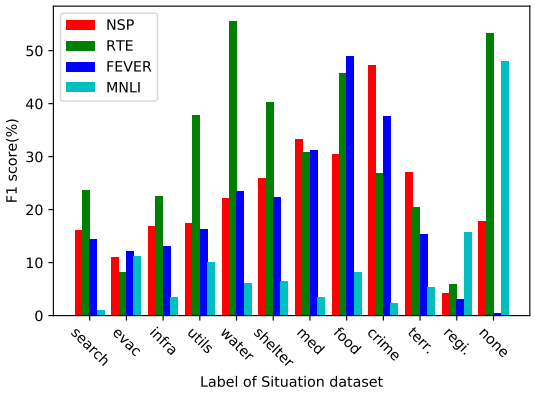

Figure A.1: F1 score of each label in Situation dataset

Case study To get a better understanding of NLI models' behavior, we carry out a case study on

SNIPS. We use Integrated Gradient (Sundararajan et al., 2017) method to attribute the entailment class's output score of BERT model to per input token [11]. Several examples are shown in Table A.3.

We found the NLI models sometimes rely on spurious patterns to do prediction. In the first example, the model finetuned on FEVER assigns a high negative attribution score to the word "*zero*" and makes a wrong prediction. However, if we replace "*zero*" with other numbers, the model changes its prediction and can correctly predicts the "RateBook" label. These examples reflect model trained on FEVER dataset learns the spurious correlations between "not entailment" label and the occurrence of word "*zero*"[12]. These superficial patterns may not be the models' main behaviour for prediction, it still leaks the model's fragility and could be an important factor to the model's failure in zero-shot scenario.

The other two groups of cases show another problem: current NLI models only predict "*entailment*" label when the premise *entails* its hypothesis, this problem definition is just different from the zero-shot test tasks. For example, in the last group, model trained on MNLI outputs a low probability for entailment since "*restaurant*" can not be directly inferred from premise sentence. If we change "*restaurant*" into "*place*", the model confidently predicts "*entailment*" .

Error cases We also show some additional examples in Table A.4, from which we might naturally conjecture that the entailment models could rely on spurious lexical features for prediction.

Impact of template choice How to properly choose templates is another issue when utlizing NLI for zero-shot classification. As shown in Table A.5, different templates that seem meaningful to human might have large performance variance on SST-2.

A.3 Details for Stability Experiments

Details for training settings For MNLI dataset, we merge the *neutral* and *contradiction* labels into *not-entailment* label following Yin et al. (2019). We choose hyperparameters randomly for different runs: we choose learning rate from $\{2e^{-5}, 3e^{-5}, 5e^{-5}\}$, training epochs from $\{3, 4, 5\}$ and randomly set the random seed.

Results for training on RTE As shown in Table A.6, the performance of different runs has large variance on both RTE dev and text classification datasets due to its small size.

Reorganize dev and test sets for Yahoo and AGNews We reorganize the Yahoo development set provided by Yin et al. (2019) and divide test set as follows: For the dev set, the instances with label in set {"Society & Culture", "Health", "Computers & Internet", "Business & Finance", "Family & Relationships"} are preserved, we call this new dev set as **Yahoo-dev** . For the original test set, we only select instances with the label which doesn't appear in the dev set as our new test set, denoted as **Yahoo-test**. During the NLI model training, we select the checkpoint by the performance on Yahoo-dev, and we report the variance of five different runs trained on MNLI. We also conduct experiments on AGNews in the same way. We use {"World", "Sports"} as seen labels and select 1800 instances per seen label randomly in train data as our new development set. In the same way, we get dev set : **AGNews-dev** and our test set **AGNews-test**.

A.4 Details of Robust NLI models

Details for training settings For all the models, we use the same set of hyperparameters: We train all the models with batch size of 64, the Adam optimizer with the initial learning rate of $2e^{-5}$ and finetune the BERT model for 3 epochs. The maximum sequence length is limited to 128.

For DA (data augmentation) method, we use the most effective strategy which is called *inversion with a transformed hypothesis* in Min et al. (2020) . For the bias model used in Reweight and BiasProduct, we use the feature based word overlap bias model[13] in Clark et al. (2019).

Detailed results on HANS Table A.7 shows detailed results for the base BERT model and each robust strategy on the HANS dataset (McCoy et al., 2019) that diagnose each of the three heuristics (the Lexical Overlap Heuristic, the Subsequence Heuristic, and the Constituent Heuristic).

[11] we use inputs which replace all tokens with pad token except for [SEP] and [CLS] as baseline of the attribution method.

[12] There are 407 premise and hypothesis pairs which contain word "*zero*" with a REFUTES label, while 122 pairs with a SUPPORTS label.

[13] https://github.com/chrisc36/debias

Model	Input text with label as hypothesis	Predicted	Gold-Std.
FEVER	**Original :** [CLS] rate current essay a zero [SEP] i want to rate a book . [SEP] (0.140) **Variation:** [CLS] rate current essay a one [SEP] i want to rate a book . [SEP] (0.627) **Variation:** [CLS] rate current essay a five [SEP] i want to rate a book . [SEP] (0.758)	SearchScreeningEvent (0.203) RateBook (0.627) RateBook (0.758)	RateBook
RTE	**Original :** [CLS] for the current saga i rate 2 of 6 stars [SEP] i want to rate a book . [SEP] (0.001) **Variation:** [CLS] for the current novel i rate 2 of 6 stars [SEP] i want to rate a book . [SEP] (0.925) **Variation:** [CLS] for the current essay i rate 2 of 6 stars [SEP] i want to rate a book . [SEP] (0.029)	AddToPlaylist (0.001) RateBook (0.925) SearchCreativeWork (0.043)	RateBook
MNLI	**Original :** [CLS] make me a reservation in tn somewhere nearby for a party of 4 [SEP] i want to book a restaurant . [SEP] (0.012) **Variation:** [CLS] make me a reservation in tn somewhere nearby for a party of 4 [SEP] i want to book a place . [SEP] (0.918) **Variation:** [CLS] make me a reservation in tn somewhere nearby for eating [SEP] i want to book a restaurant . [SEP] (0.797)	AddToPlaylist (0.017) - BookRestaurant (0.797)	BookRestaurant

Table A.3: Examples for visualization of attribution score. Each example is followed by the model's prediction probability for entailment class. "Predict" column shows the model's predicted class with its entailment probability for the input premise text and "Gold-Std." column displays the true labels. The red color represents negative attribution score and the blue color represents positive score for entailment class. Better viewed in color.

Text with Gold-standard and Predicted labels
• Gold-standard: Computers&Internet • Prediction: Entertainment&Music (MNLI, RTE), Computers&Internet (FEVER) *Is it possible to rip the **music** from PS2 games ? No i dont think thats possible because your computer cant understand the data format your ps2 games . Ive also never heard of that being done so id have to say no .*
• Gold-standard: Education&Reference • Prediction:Family&Relationships(RTE,FEVER,MNLI) *Who or which company would do the best **family** history and genealogy research for me in Utah ? I know if you go to the Mormon Church , they can provide tons of answers about your genealogy , and probably suggest a company or person who would do the work for you .*
• Gold-standard: BookRestaurant • Prediction: RateBook (RTE,FEVER,MNLI) ***book** a bakery for lebanese on january 11th 2032*
• Gold-standard: BookRestaurant • Prediction: RateBook(RTE,FEVER,MNLI) ***book** a highly **rated** place in in in seven years at a pub*
• Gold-standard: Negative • Prediction: Positive (RTE,FEVER,MNLI) *outer-space buffs might **love** this film , but others will find its **pleasures** intermittent .*

Table A.4: Error cases of the entailment models which may rely on spurious lexical features to make prediction. Bolded tokens indicate those cue words that may mislead the NLI models.

Template	NSP	RTE	MNLI	FEVER
The movie is great/terrible.	79.7	52.5	67.5	71.7
The movie is good/bad.	78.9	52.6	75.8	78.3
The person feels good/bad.	69.3	63.5	78.3	82.9

Table A.5: Accuracy on SST-2 dev set using different templates

Dataset	Average	Std	Min	Max
RTE Dev set	69.0	2.2	66.1	70.8
Yahoo	20.6	7.4	11.2	28.6
Emotion	3.8	0.4	3.5	4.4
Situation	23.0	4.5	16.9	28.3
AGNews	31.1	15.5	9.1	46.4
SST-2	67.0	3.9	63.9	72.0
Snips	67.5	2.5	64.3	71.3

Table A.6: Results of five runs of BERT fine-tuned on RTE and tested on classification datasets

		Entailment			*Non-entailment*		
	Overall	*L*	*S*	*C*	*L*	*S*	*C*
MNLI	53.0	99.5	99.8	97.2	2.7	1.6	17.2
w/ DA	67.3	81.2	94.6	96.6	86.8	23.7	20.7
w/ RW	64.5	69.8	80.6	78.5	53.1	40.2	65.0
w/ BP	65.4	71.4	77.4	84.6	61.0	40.7	57.2

Table A.7: HANS accuracy of BERT pretrained on MNLI and different debiasing methods, broken down by the heuristic that the example is diagnostic of and by its gold label. *L* represents for Lexical Overlap Heuristic, *S* represents for Subsequence Heuristic, and *C* represents for the Constituent Heuristic.