Looking at the Overlooked: An Analysis on the Word-Overlap Bias in Natural Language Inference

Sara Rajaee[1], **Yadollah Yaghoobzadeh**[2], and **Mohammad Taher Pilehvar**[3]
[1] University of Amsterdam, Netherlands
[2] University of Tehran, Iran
[3] Tehran Institute for Advanced Studies, Khatam University, Iran
s.rajaee@uva.nl
y.yaghoobzadeh@ut.ac.ir
mp792@cam.ac.uk

Abstract

It has been shown that NLI models are usually biased with respect to the word-overlap between premise and hypothesis; they take this feature as a primary cue for predicting the entailment label. In this paper, we focus on an overlooked aspect of the overlap bias in NLI models: the *reverse* word-overlap bias. Our experimental results demonstrate that current NLI models are highly biased towards the non-entailment label on instances with low overlap, and the existing debiasing methods, which are reportedly successful on existing challenge datasets, are generally ineffective in addressing this category of bias. We investigate the reasons for the emergence of the overlap bias and the role of minority examples in its mitigation. For the former, we find that the word-overlap bias does not stem from pre-training, and for the latter, we observe that in contrast to the accepted assumption, eliminating minority examples does not affect the generalizability of debiasing methods with respect to the overlap bias. All the code and relevant data are available at: https://github.com/sara-rajaee/reverse_bias

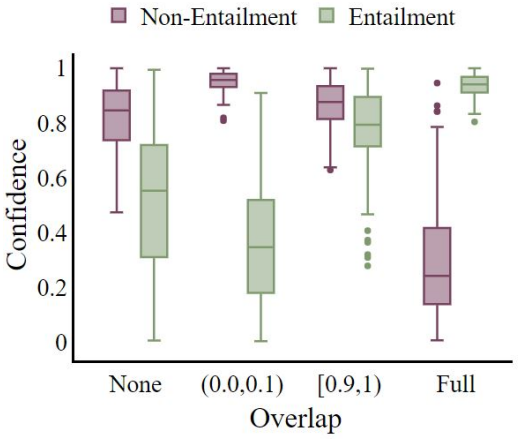


Figure 1: NLI model's confidence on a randomly sampled subset of instances from the SNLI dataset across four different degrees of word overlap between premise and hypothesis. BERT is biased towards the entailment label on instances with full overlap (denoted by the huge confidence gap with the non-entailment label). On the contrary, a *reverse bias* is seen for low and non-overlapping instances, with a significant confidence lead on the non-entailment label.

1 Introduction

Natural Language Inference (NLI) is one of the most commonly used NLP tasks, particularly in the scope of evaluating models for their language understanding capabilities. Since their emergence, pre-trained language models (PLMs) have been highly successful on standard NLI datasets, such as the Multi-Genre Natural Language Inference (Williams et al., 2018, MultiNLI). However, recent analytical studies have revealed that their success is partly due to their reliance on spurious correlations between superficial features of the input texts and gold labels in these datasets (Poliak et al., 2018; Bhargava et al., 2021). As a result, performance usually drops on out-of-distribution datasets where such correlations do not hold. Several proposals have been put forth to enhance the robustness of models to the known and unknown biases and improve performance on the so-called challenging datasets (Stacey et al., 2020; Utama et al., 2020a; Asael et al., 2022).

One of the well-known dataset biases in NLI models is the spurious correlation of the *entailment* label and high word-overlap between premise and hypothesis. A number of challenging sets are designed to showcase the tendency of PLMs to predict entailment for most such cases. HANS (McCoy et al., 2019) is arguably the most widely used dataset in this group. Constructed based on human-made linguistic patterns, the dataset focuses on high-overlapping samples, the non-entailment subset of which is deemed as challenging for NLI models. Most current debiasing methods have considered the word-overlap bias as one of their main

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 10605 - 10616
December 7-11, 2022

targets and have shown substantial improvements on HANS (Mendelson and Belinkov, 2021; Min et al., 2020).

In this paper, we revisit the word-overlap bias in NLI and the effectiveness of existing debiasing techniques. Despite the popularity of this type of bias, we find that some of its aspects are generally ignored in the research community. If we consider word-overlap as a feature with values ranging from no to full overlap, and NLI task with two labels of entailment and non-entailment, we show that there are other kinds of spurious correlation than the popular high word-overlap and entailment. Specifically, as it is shown in Figure 1, we see a clear bias towards non-entailment for the low and no word-overlap values (denoted by the high performance on the non-entailment label, which comes at the price of reduced performance on the entailment class). We will refer to this type of bias as *reverse* word-overlap throughout the paper.

Through a set of experiments, we demonstrate that the overlooked reverse word-overlap bias exists in popular NLI datasets, such as MNLI and SNLI, as well as in the predictions of PLMs. Moreover, our results suggest that while existing debiasing methods can mitigate the overlap bias in NLI models to some extent, they are ineffective in resolving the reverse bias.

Moreover, we analyze how NLI models employ minority instances to enhance their generalization. Focusing on the forgettable debiasing method (Yaghoobzadeh et al., 2021), we realize that eliminating HANS-like examples and the reverse ones do not hurt the generalization noticeably.

In search of the origin of the bias, we employ prompt-based techniques to check whether the bias stems from pre-training. We also verify the robustness of PLMs in a few-shot learning experiment with controlled and balanced training sets. Our results suggest that PLMs do not exhibit any bias towards a specific label. Nevertheless, introducing a few samples triggers the bias toward the entailment label. Furthermore, balancing the training examples with respect to their word-overlap prevents the emergence of bias to some extent.

Our contributions can be summarized as follows:

- We expand our understanding of the word-overlap bias in NLI by revealing an unexplored spurious correlation between low word-overlap and non-entailment.
- We analyze how debiasing methods work for the whole spectrum of word-overlap bias, finding that they generally fail at addressing bias for the low and non-overlapping cases.
- To explore the origin of word-overlap bias in PLMs, we design several new experiments showing that, even when exposed to a few training examples, PLMs get biased towards predicting entailment.

2 Natural Language Inference

In NLI, a model is provided with two input sentences, namely *premise* and *hypothesis*. The task for the model is to predict whether the hypothesis is true (*entailment*), false (*contradiction*), or undetermined (*neutral*) given the premise.

2.1 Bias in NLI Models

Analyzing NLI models have demonstrated that they are sensitive to the shortcuts that appear in the dataset. Several types of bias have been investigated in the literature, including hypothesis-only prediction, spurious correlations between certain words and labels (e.g., negation words and the non-entailment label), sensitivity to the length of hypothesis, and lexical overlap between the premise and hypothesis (Gururangan et al., 2018; Poliak et al., 2018; McCoy et al., 2019; Wu et al., 2022). Relying on these spurious features hampers the language understanding ability of NLI models, leading to poor performance on out-of-distribution datasets where such superficial correlations do not hold (He et al., 2019; McCoy et al., 2019).

Word-Overlap Bias. Among the detected dataset biases, word-overlap is a quite well-studied shortcut in the NLI task (Zhou and Bansal, 2020; Mendelson and Belinkov, 2021). We define word-overlap (wo) as the ratio of words in the hypothesis (h) that are shared with the premise (p), i.e., $\frac{|h \cap p|}{|h|}$. Table 1 shows examples of different degrees of word-overlap.

2.2 Debiasing Methods

Creating high-quality datasets without any spurious features between instances and gold labels is an arduous and expensive process (Gardner et al., 2021a), making it inevitable for a dataset not to have biases to some extent. Therefore, to have a robust model, it is essential to take extra steps for debiasing against dataset artifacts. The past few years

Overlap	Sample	Label
Full (1.0)	P: A little kid in blue is sledding down a snowy hill. H: A little kid in blue sledding.	Entailment
	P: The young lady is giving the old man a hug. H: The young man is giving the old man a hug.	Non-Entailment
$\frac{12}{13} = 0.923$	P: A woman in a blue shirt and green hat looks up at the camera. H: A woman wearing a blue shirt and green hat looks at the camera	Entailment
$\frac{11}{12} = 0.917$	P: Two men in wheelchairs are reaching in the air for a basketball. H: Two women in wheelchairs are reaching in the air for a basketball.	Non-Entailment
$\frac{1}{14} = 0.071$	P: Several young people sit at a table playing poker. H: Youthful Human beings are gathered around a flat surface to play a card game.	Entailment
$\frac{1}{11} = 0.091$	P: A blond woman in a white dress sits in a flowering tree while holding a white bird. H: The woman beats two eggs to make breakfast for her husband.	Non-Entailment
None (0.0)	P: A couple sits in the grass. H: People are outside.	Entailment
	P: An older women tending to a garden. H: The lady is cooking dinner.	Non-Entailment

Table 1: NLI examples with different degrees of word-overlap (between premise and hypothesis), where the overlap is the ratio of hypothesis words that are shared with the premise. The highlighted words are the common (in green) or different (in purple) words (the samples are picked to reflect extreme cases across the word-overlap spectrum).

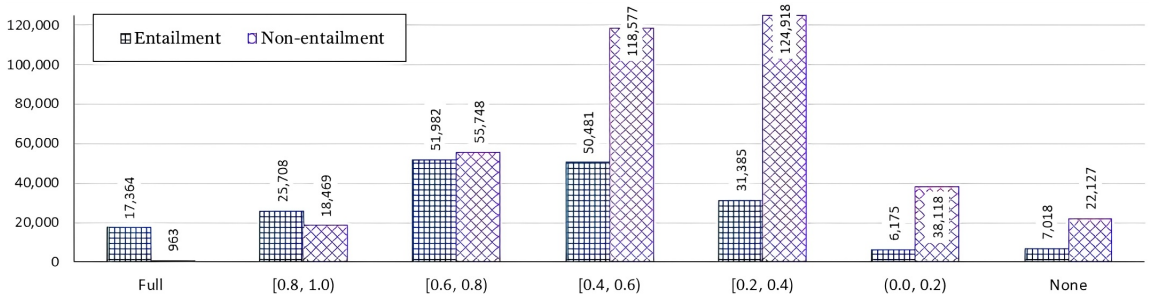


Figure 2: The distribution of instances across word-overlap bins (SNLI dataset).

have seen several debiasing methods (Karimi Mahabadi et al., 2020; Utama et al., 2020a,b; Belinkov et al., 2019). For our experiments, we opted for three different debiasing approaches. We evaluate the effectiveness of these techniques in mitigating the overlap bias and its reverse.

Long-tuning. Tu et al. (2020) have shown that fine-tuning NLI models for more epochs can enhance the generalizability of LMs over challenging datasets. Following their suggestion, we fine-tuned the models for 20 epochs on the MNLI dataset.

Forgettable Examples. Yaghoobzadeh et al. (2021) find minority examples without prior knowledge of the dataset artifacts. In the proposed method, the minority examples are considered samples that have never been learned or learned once and then forgotten by the model. Then, the already trained NLI model is fine-tuned on this subset for a few more epochs. Following the authors' suggestion, to find the forgettable examples, we utilized a simple Siamese Bag of Words (BoW) model where the sentence representations of the premise and hypothesis are the average over their word embeddings.

Product of Experts (PoE). In this method, a weak model is supposed to learn superficial features in the input. The weak learner's output is then used to normalize the main model's predictions on over-confident examples. Following previous studies (Karimi Mahabadi et al., 2020; Sanh et al., 2021), we employed the following combination strategy for taking into account both weak learner

	MNLI-dev	HANS	HANS+	HANS−	WANLI
			BERT		
Baseline	84.2 ±0.3	63.9 ±1.7	98.5 ±1.2	29.3 ±4.6	56.9 ±0.6
Long-tuning	83.4 ±0.8	65.8 ±2.3	99.0 ±0.2	32.6 ±4.4	58.0 ±0.6
$\mathcal{F}_{\text{BoW}}$	82.7 ±0.3	73.8 ±0.5	91.8 ±0.4	55.9 ±1.3	59.0 ±0.3
PoE	80.0 ±0.8	66.9 ±2.2	71.6 ±3.7	62.2 ±2.7	71.6 ±0.7
			RoBERTa		
Baseline	87.2 ±0.2	73.3 ±3.4	98.5 ±1.0	48.2 ±7.8	59.7 ±1.6
Long-tuning	86.9 ±0.3	73.0 ±1.7	97.8 ±1.2	48.2 ±4.2	60.3 ±0.1
$\mathcal{F}_{\text{BoW}}$	85.6 ±0.3	78.9 ±0.6	88.1 ±2.4	69.7 ±2.3	62.0 ±1.4
PoE	84.6 ±0.1	77.0 ±1.5	79.3 ±6.2	71.4 ±3.7	73.4 ±0.1

Table 2: The average accuracy of the baseline models and debiasing methods on the MNLI development (matched) set as the *in-distribution* and WANLI and HANS as the *out-of-distribution* datasets (HANS+ and HANS− are entailment and non-entailment subsets, respectively).

and main model predictions:

$$y = softmax(\log p_w + \log p_m) \quad (1)$$

where p_w and p_m are the outputs of the weak learner and the main model, respectively. The robust model is trained using a cross-entropy loss function based on y. We used TinyBERT (Jiao et al., 2020) as our weak learner.

2.3 Experimental Setup

Datasets. In our experiments, we opted for the Multi-Genre Natural Language Inference dataset (Williams et al., 2018, MNLI) for training the NLI models. The dataset contains 433k training examples. Since the gold labels for the test set are not publicly available, we follow previous work and report results on the *development-matched* (MNLI-dev in the tables). Also, following the convention in previous studies, we merge neutral and contradiction examples into the non-entailment group. As challenging datasets, we considered HANS (McCoy et al., 2019) and WANLI (Liu et al., 2022). In the former dataset, each instance is curated in a way that all words of the hypothesis are also observed in the premise, irrespective of the word order. Previous work has shown that biased NLI models tend to perform poorly on HANS, particularly for the non-entailment class (Yaghoobzadeh et al., 2021). The latter challenging set has employed GPT-3 (Brown et al., 2020) to generate high-quality instances followed by filtering done by human crowd-workers. Quality tests on WANLI indicate that the dataset contains fewer artifacts compared to MNLI.

Models. As for PLMs, we opted for the base version of BERT and RoBERTa (Devlin et al., 2019; Liu et al., 2020) and fine-tuned them for three epochs as our baselines. We trained the models with a learning rate of 2e-5, employing the Adam optimizer for three different random seeds. The batch size was set to 32 with a max length of 128. All the reported results are based on three random seeds.

2.4 Results

Table 2 shows the results for the baseline models (BERT and RoBERTa) and the three debiasing techniques on different datasets. The bias in the baseline model is highlighted by the performance contrast across the entailment (HANS+) and non-entailment (HANS−) subsets. As can be seen, the three debiasing methods are generally effective in softening the biased behavior, reflected by the improved performance on HANS− (and, in turn, HANS), and also WANLI.

3 Reverse Word-Overlap

Considering the word-overlap bias as a spectrum, the existing studies have mainly focused on a small subset of the spectrum, i.e., the case with full word-overlap and its spurious correlation with the entailment label. In this section, we evaluate the performance of NLI models on other areas of the spectrum and with respect to both labels (entailment and non-entailment) to broaden our insights on the robustness of these models considering the word-overlap feature.

Overlap	BERT		RoBERTa	
	Entailment	Non-Entailment	Entailment	Non-Entailment
Full	99.7 ±0.1	13.3 ±1.4	99.7 ±0.1	17.6 ±0.9
[0.8, 1.0)	92.9 ±0.0	83.0 ±1.5	95.9 ±0.6	92.7 ±2.4
[0.6, 0.8)	85.2 ±0.4	86.2 ±1.6	91.5 ±1.4	84.5 ±2.8
[0.4, 0.6)	74.2 ±0.1	91.9 ±1.1	85.8 ±2.4	90.2 ±2.4
[0.2, 0.4)	64.5 ±0.6	95.1 ±0.6	78.5 ±2.8	93.8 ±1.6
(0.0, 0.2)	55.5 ±1.4	96.7 ±0.5	68.6 ±3.3	96.0 ±1.2
None	61.6 ±1.3	95.2 ±0.2	77.2 ±3.4	93.6 ±1.5

Table 3: The accuracy of the two NLI models across different overlap bins and on both subsets. The lowest numbers in each column are underlined.

3.1 Probing Dataset

As for this probing study, we experimented with the SNLI dataset (Bowman et al., 2015), merging the training, development, and test sets to build a unified evaluation set. The set was split into seven bins based on the degree of overlap. The statistics are reported in Figure 2. As an example, the $[0.6, 0.8)$ bin contains samples that have a word overlap (between premise and hypothesis) of greater than (and equal to) 0.6 and less than 0.8.

3.2 Results

Unless specified otherwise, the experimental setup in this experiment is the same as the one reported in Section 2.3. Table 3 reports the results across different word overlap bins for both BERT and RoBERTa and for both labels. As expected, high contrast is observed on the full overlap subset: near-perfect NLI performance on the entailment, while poor performance on non-entailment, suggesting a strong bias towards the entailment label. This is the conventional type of NLI bias that has been usually discussed in previous studies. The HANS challenging dataset is constructed based on the same type of bias. However, surprisingly, the results show that this biased behavior only exists for samples with full overlap. In fact, no notable bias is observed even for the high overlap samples in the $[0.8, 1)$ bin. This observation further narrows down the scope of HANS as a challenging dataset and raises questions on the robustness of models developed based on the dataset.

Reverse bias. Interestingly, the results in Table 3 shed light on another inherent spurious correlation that exists between NLI performance and the degree of word-overlap. Particularly towards the non-overlap extreme, the performance drops on entailment and increases on non-entailment samples. In the (0.0, 0.2) bin, we see the largest gap: 55.5 entailment vs 96.7 non-entailment for the BERT model. We refer to the biased behavior of NLI models on the low word-overlapping samples towards the non-entailment label as the *Reverse bias*.

It is also worth mentioning that based on the proposed results, reverse bias covers a broader range of bins in comparison with the word-overlap bias.

3.3 Effectiveness of Debiasing Methods

Figure 3 shows the performance of the three debiasing methods (described in Section 2.2) across the seven bins in our word-overlap analysis. As can be observed, debiasing methods improve over the baseline on the full-overlap ("Full" in Figure 3) and non-entailment subset, with PoE proving the most effective. The improvement is expected since the results on the challenging dataset, HANS, suggest the same. This, however, comes at the price of reduced performance on the entailment subset, specifically in the BERT model.

As we move toward the non-overlap end of the spectrum ("None" in Figure 3), the performance gap between the entailment and non-entailment labels grows, mainly due to the drop in entailment performance. Interestingly, the experimental results reveal that debiasing methods are clearly ineffective in addressing the reverse bias and perform similarly to the baseline models.

4 Analysis

4.1 Role of Minority Examples

In the context of word-overlap bias, the non-entailment instances that have full overlap (between premise and hypothesis) are usually referred to as *minority* examples. Tu et al. (2020) show that

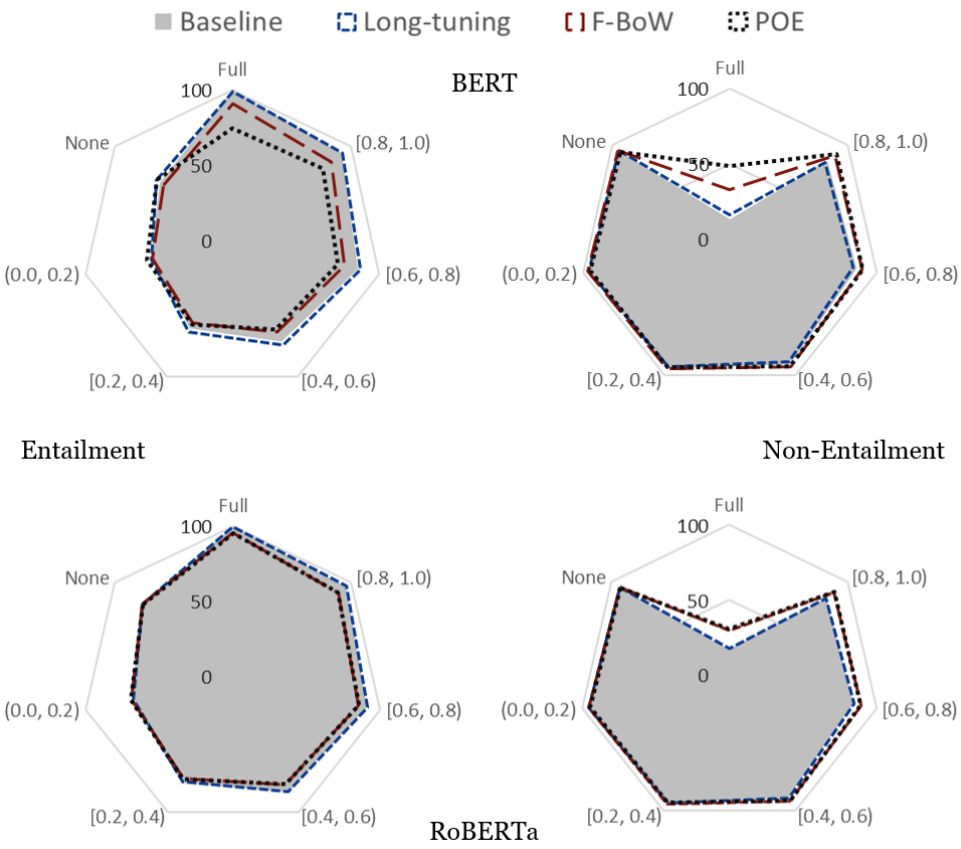


Figure 3: The performance of the baseline and the three debiasing methods across the seven word-overlap bins for both labels and for BERT and RoBERTa. Across the spectrum, the debiasing techniques seem to be effective only on samples with high (particularly full) word-overlap on the non-entailment subset and are either ineffective (or even harmful) towards the other end of the overlapping spectrum and on the entailment subset.

minority examples of the training set play a crucial role in the generalizability of language models, and eliminating them can significantly hurt performance on challenging datasets, such as HANS. Yaghoobzadeh et al. (2021) relate the forgettables with the minority examples by observing the difference in word-overlap distribution in forgettables.

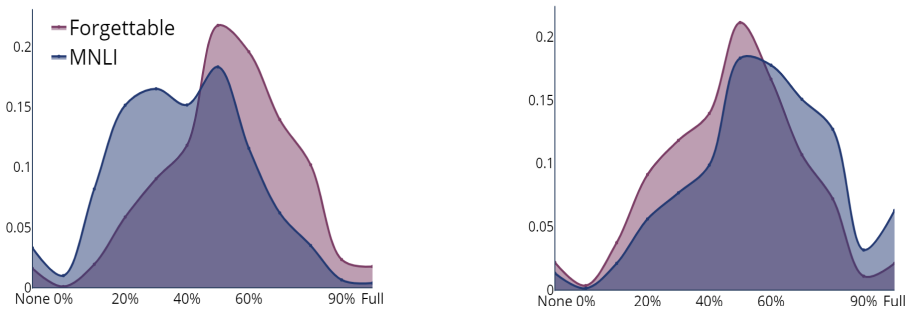
We carry out a set of experiments on the *forgettable* approach, where a subset of the training data is chosen for further fine-tuning of models ($66k$ in our NLI experiments for the $\mathcal{F}_{\text{BoW}}$ method). We extend the forgettable analysis to the low word-overlap or reverse minority examples. We also verify the role played by minority examples in the performance of debiasing methods.

As the first step, we compare the distribution of instances with respect to their overlap in the original training set of MNLI and its forgettable subset. The results are shown in Figure 4. As can be seen, the forgettable subset tends to have better coverage over the minority subset than the original MNLI training set. See the right side of Figure 4(a) and the left side of Figure 4(b).

One can hypothesize that better coverage of minority examples is the reason behind the effectiveness of the forgettable approach. To verify this hypothesis, we eliminate several subsets from $\mathcal{F}_{\text{BoW}}$ and fine-tune the NLI models with the remaining samples. We considered the following four settings:

- **Full − NEnt:** Full overlap between premise and hypothesis with the non-entailment label.
- **None − Ent:** No overlap and entailment label.
- **[0.8, 1.0] − NEnt:** More then 80% overlap and non-entailment label.
- **[0.0, 0.2] − Ent:** Less than 20% overlap and entailment label.

The results are reported in Table 4. Interestingly, we observe that removing HANS-like examples



(a) Non-Entailment (b) Entailment

Figure 4: Normalized distribution of instances with respect to their word-overlap in the original training set of MNLI and the subset identified by $\mathcal{F}_{\text{BOW}}$.

(Full−NEnt), which were hypothesized to play the main role in improving performance on the challenging datasets, does not affect the performance of $\mathcal{F}_{\text{BOW}}$ notably. The observation is consistent even for larger subsets of high-overlapping instances ([0.8, 1]−NEnt). Discarding the reverse group (low-overlapping entailment samples) yields a similar pattern. So, it can be inferred that such samples do not play the primary role in the debiasing methods' effectiveness.

This opens up questions on how NLI models extrapolate to patterns unseen during training and how debiasing methods enhance their generalization over out-of-distribution data. This is particularly interesting in light of observations made by (Tu et al., 2020) that standard training does not enable such extrapolation. We leave further investigations in this area to future work.

4.2 The Origin of Word-Overlap Bias

We conducted another experiment to see if the vulnerability of NLI models to the word-overlap feature and the reverse bias comes from pre-training or from fine-tuning on the task-specific data. To this end, we followed Utama et al. (2021) in evaluating pre-trained models under zero- and few-shot settings. To rule out the impact of fine-tuning and verify if the pre-trained model exhibits similar biases with respect to word-overlap, we evaluated BERT in a zero-shot setting by reformulating the NLI task as a masked language modeling objective. Following previous studies (Schick and Schütze, 2021; Utama et al., 2021), we transformed the NLI examples using the below template:

```
Premise ? [MASK], Hypothesis.
```

where the `[MASK]` token denotes the gold label. We used a simple verbalizer with *yes*, *maybe*, and *no* as mappings to, respectively, the entailment, neutral, and contradiction labels.

The first row of Table 5 shows the results for the zero-shot setting. The similar performance across HANS− and HANS+ shows that the pre-trained BERT model does not exhibit much bias towards a specific label. Therefore, the bias stems from the fine-tuning on the task-specific instances. This is reflected even with as few as 16 samples in the few-shot scenario (where we have fine-tuned the prompt-based model). As the number of training instances increases, the gap between the entailment and non-entailment samples grows.

Balanced data. We also examined the role of class imbalance in the training data on the emergence of word-overlap bias. For this experiment, we defined four categories based on the overlap {Full, [0.5, 1), (0.0, 0.5), and None} and uniformly sampled K instances per label. The bottom block of Table 5 presents the results. It can be inferred that having a balanced training set can reduce the bias to some extent. Finally, the high variance on the HANS subsets suggests that the quality of training examples and word-overlap percentage between the premise and hypothesis can have a significant impact on the bias in NLI systems.

5 Related Work

Dataset biases in NLP. Different categories of bias have been discovered and discussed in NLP datasets. Earlier work has discovered that negative words are correlated with contradiction label in the SNLI dataset (Naik et al., 2018; Gururangan et al., 2018). Hypothesis-only (Gururangan et al., 2018) and word-overlap between hypothesis and premise (McCoy et al., 2019) are other types of biases discussed in the literature of SNLI and MNLI datasets.

	MNLI-dev	HANS	HANS+	HANS−	WANLI	Eliminated
			BERT			
Baseline	84.2 ±0.3	63.9 ±1.7	98.5 ±1.2	29.3 ±4.6	56.9 ±0.6	
$\mathcal{F}_{\text{BoW}}$	82.7 ±0.3	73.8 ±0.5	91.8 ±0.4	55.9 ±1.3	59.0 ±0.3	
Full − NEnt	82.8 ±0.4	71.7 ±0.9	93.2 ±0.4	50.3 ±2.0	59.4 ±0.5	782
[0.8, 1.0] − NEnt	83.2 ±0.2	72.3 ±0.8	93.5 ±1.3	51.1 ±2.9	58.8 ±0.5	6,350
[0.0, 0.2] − Ent	82.9 ±0.4	73.7 ±0.7	91.9 ±0.8	55.4 ±2.1	59.5 ±0.7	1,801
None − Ent	82.8 ±0.5	73.8 ±0.8	92.1 ±1.5	55.5 ±3.1	59.3 ±0.6	482
			RoBERTa			
Baseline	87.2 ±0.2	73.3 ±3.4	98.5 ±1.0	48.2 ±7.8	59.7 ±1.6	
$\mathcal{F}_{\text{BoW}}$	85.6 ±0.3	78.9 ±0.6	88.1 ±2.4	69.7 ±2.3	62.0 ±1.4	
Full − NEnt	86.4 ±0.2	79.1 ±1.3	92.1 ±1.6	66.1 ±4.0	62.2 ±0.9	782
[0.8, 1.0] − NEnt	86.6 ±0.2	78.4 ±1.0	95.9 ±0.8	60.8 ±2.8	61.8 ±0.7	6,350
[0.0, 0.2] − Ent	86.1 ±0.2	79.3 ±1.3	89.8 ±1.2	68.7 ±2.1	62.3 ±0.9	1,801
None − Ent	86.1 ±0.2	79.1 ±1.2	88.6 ±2.1	69.5 ±2.9	62.1 ±0.7	482

Table 4: The performance of $\mathcal{F}_{\text{BoW}}$ after eliminating four different subsets. *Eliminated* denotes the number of eliminated examples in each setting. All the subsets tend to be in the same performance ballpark with respect to the generalizability of the model on the out-of-distribution datasets (WANLI and HANS).

In particular, word overlap has also been investigated in the context of duplicate question detection on the QQP dataset (Zhang et al., 2019). For both NLI and QQP, it has been shown that considerable spurious correlations exist between high word overlap and the entailment/duplicate label. In this word, we focused on the word overlap bias in the NLI dataset and introduced an overlooked aspect of this bias: the correlation between low word overlap and non-entailment class.

Challenging sets. In the past few years, several challenging datasets have been introduced to study the limitations of NLP models and, in particular, pre-trained language models in learning robust features and ignoring dataset biases. Challenging datasets for NLI include HANS (McCoy et al., 2019), ANLI (Williams et al., 2022), MNLI-hard (Gururangan et al., 2018) and Stress-tests (Naik et al., 2018). Similar datasets for other tasks include PAWS (Zhang et al., 2019; Yang et al., 2019), for duplicate question detection, and FEVER-Symmetric (Schuster et al., 2019), for stance detection.

Spurious correlation. Gardner et al. (2021b) argue that for complex language understanding tasks, any simple feature correlation should be considered spurious, e.g., "not" and the contradiction label in NLI. Spurious correlations can also be defined from the viewpoint of generalizability Chang et al. (2021); Yaghoobzadeh et al. (2021). According to this definition, a feature is spurious if it works well only for specific examples. The reverse word overlap feature described in this paper fits well within both definitions. Schwartz and Stanovsky (2022a) review several definitions for spurious correlations.

Debiasing methods. Many studies try to remove the spurious correlations or dataset biases either from the training dataset or the model. Most debiasing approaches filter or down weight those training examples that are either easy or contain spurious correlations (He et al., 2019; Karimi Mahabadi et al., 2020; Utama et al., 2020a; Sanh et al., 2021). Others augment the training set with examples that violate the spurious correlations. A mix of both these approaches has also been investigated by Wu et al. (2022). An alternative approach is to extend the fine-tuning either on all (Tu et al., 2020) or parts of training data (Yaghoobzadeh et al., 2021).

Analysis of debiasing. Given the increasing interest in debiasing methods, there have been concerns about their widespread use. Schwartz and Stanovsky (2022b) argue that excessive balancing prevents the models from learning anything (in particular, important world and commonsense knowledge), making it neither practical nor desired. They suggest abstaining and interacting with the user when the contextual information is not sufficient and also focus on zero- and few-shot learning approaches instead of full fine-tuning. In this paper, we showed that balancing datasets should only be

	Baseline				
	MNLI-dev	HANS	HANS+	HANS−	WANLI
Zero-shot	42.0	55.3	57.5	53.1	58.0
$K = 16$	45.6 ± 1.2	53.6 ± 1.3	73.2 ±16.5	34.4 ±13.9	54.7 ± 2.3
$K = 32$	46.9 ± 0.6	50.8 ± 0.8	98.3 ± 1.2	3.3 ± 2.8	50.1 ± 2.2
$K = 64$	49.6 ± 0.3	50.3 ± 0.3	99.4 ± 0.5	1.1 ± 1.1	48.4 ± 4.3
$K = 128$	52.7 ± 0.9	50.0 ± 0.0	99.9 ± 0.2	0.1 ± 0.2	45.1 ± 0.4
$K = 256$	56.4 ± 0.4	50.7 ± 0.8	98.1 ± 2.2	3.3 ± 3.9	50.3 ± 0.0
$K = 512$	61.4 ± 1.1	50.0 ± 0.1	100 ± 0.0	0.1 ± 0.1	46.2 ± 2.0
	Balanced				
$K = 16$	44.1 ± 0.6	52.5 ± 1.5	95.6 ± 2.6	9.3 ± 5.7	54.3 ± 3.2
$K = 32$	45.7 ± 1.3	51.9 ± 1.1	82.2 ±15.7	21.5 ±13.4	52.0 ± 1.3
$K = 64$	45.2 ± 1.1	52.4 ± 1.1	69.8 ± 6.0	35.1 ± 3.8	54.4 ± 0.3
$K = 128$	48.0 ± 0.1	51.7 ± 0.1	95.7 ± 5.0	7.7 ± 5.2	52.8 ± 3.3
$K = 256$	51.3 ± 1.3	51.2 ± 3.0	84.9 ±15.6	17.5 ±21.5	51.8 ± 3.5
$K = 512$	53.2 ± 0.2	51.3 ± 2.8	86.8 ±10.8	15.8 ±16.5	49.5 ± 1.7

Table 5: Zero-shot and few-shot results of prompt-based fine-tuning for BERT. While no significant bias is seen in the zero-shot setting, only with a few task-specific examples, BERT predictions are biased towards entailment (HANS+ vs. HANS−). Balancing the training set (bottom block) slightly reduces the extent of bias.

taken as a partial solution for eliminating spurious correlations. We also showed that in this context, few-shot learning might not be effective. Mendelson and Belinkov (2021) found that debiasing methods encode more extractable information about the bias in their inner representations. This observation is explained in a concurrent work to ours in terms of the necessity and sufficiency of the biases (Joshi et al., 2022). In this paper and for the word-overlap bias, we showed that our selected debiasing techniques are not robust against if we consider the whole spectrum.

6 Conclusions

In this work, we uncovered an unexplored aspect of the well-known word-overlap bias in the NLI models. We showed a spurious correlation between the low overlap instances and the non-entailment label, namely the *reverse* word-overlap bias. We demonstrated that existing debiasing methods are not effective in mitigating the reverse bias. We found that the generalization power of debiasing methods (the forgettable approach in particular) does not stem from minority examples. We also showed that the word-overlap bias does not seem to come from the pre-training step of PLMs. As future work, we plan to focus on designing new debiasing methods for mitigating the reverse bias for NLI and similar tasks. Also, building specific challenging sets, similar to HANS, for the reverse bias helps to expand this line of research.

7 Acknowledgements

We would like to acknowledge that the idea of reverse bias was initiated in discussion with Alessandro Sordoni (MSR Montreal). Also, we want to thank the anonymous reviewers for their valuable comments, which helped us in improving the paper. Sara Rajaee is funded in part by the Netherlands Organization for Scientific Research (NWO) under project number VI.C.192.080.

8 Limitations

In our experiments, we have focused on two popular PLMs, BERT and RoBERTa. Using more PLMs, with diversity in the objective and architecture and evaluating their robustness is one of the extendable aspects of our work. Moreover, we evaluated three debiasing methods, but this could have been expanded to more. The other susceptible aspect to improvement is creating a more high-quality dataset for analyzing the overlap bias and its reverse. We have used SNLI as our main probing set, a crowdsourcing-based dataset that contains some noisy examples, especially in minority groups.

References

Dimion Asael, Zachary Ziegler, and Yonatan Belinkov. 2022. A generative approach for mitigating structural biases in natural language inference. In *Proceedings of the 11th Joint Conference on Lexical and Compu-*

tational Semantics, pages 186–199, Seattle, Washington. Association for Computational Linguistics.

Yonatan Belinkov, Adam Poliak, Stuart Shieber, Benjamin Van Durme, and Alexander Rush. 2019. Don't take the premise for granted: Mitigating artifacts in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 877–891, Florence, Italy. Association for Computational Linguistics.

Prajjwal Bhargava, Aleksandr Drozd, and Anna Rogers. 2021. Generalization in NLI: Ways (not) to go beyond simple heuristics. In *Proceedings of the Second Workshop on Insights from Negative Results in NLP*, pages 125–135, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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 *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Kai-Wei Chang, He He, Robin Jia, and Sameer Singh. 2021. Robustness and adversarial examples in natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 22–26, Punta Cana, Dominican Republic & Online. Association for Computational Linguistics.

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.

Matt Gardner, William Merrill, Jesse Dodge, Matthew Peters, Alexis Ross, Sameer Singh, and Noah A. Smith. 2021a. Competency problems: On finding and removing artifacts in language data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1801–1813, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Matt Gardner, William Merrill, Jesse Dodge, Matthew E Peters, Alexis Ross, Sameer Singh, and Noah A Smith. 2021b. Competency problems: On finding and removing artifacts in language data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1801–1813.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.

Nitish Joshi, Xiang Pan, and He He. 2022. Are all spurious features in natural language alike? an analysis through a causal lens. In *Proceedings of EMNLP*.

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2020. End-to-end bias mitigation by modelling biases in corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8706–8716, Online. Association for Computational Linguistics.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. Wanli: Worker and ai collaboration for natural language inference dataset creation.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro{bert}a: A robustly optimized {bert} pretraining approach.

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.

Michael Mendelson and Yonatan Belinkov. 2021. Debiasing methods in natural language understanding make bias more accessible. In *Proceedings of the*

2021 Conference on Empirical Methods in Natural Language Processing, pages 1545–1557, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Victor Sanh, Thomas Wolf, Yonatan Belinkov, and Alexander M. Rush. 2021. Learning from others' mistakes: Avoiding dataset biases without modeling them. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. 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 3419–3425, Hong Kong, China. Association for Computational Linguistics.

Roy Schwartz and Gabriel Stanovsky. 2022a. On the limitations of dataset balancing: The lost battle against spurious correlations. *arXiv preprint arXiv:2204.12708*.

Roy Schwartz and Gabriel Stanovsky. 2022b. On the limitations of dataset balancing: The lost battle against spurious correlations. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2182–2194, Seattle, United States. Association for Computational Linguistics.

Joe Stacey, Pasquale Minervini, Haim Dubossarsky, Sebastian Riedel, and Tim Rocktäschel. 2020. Avoiding the Hypothesis-Only Bias in Natural Language Inference via Ensemble Adversarial Training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8281–8291, Online. Association for Computational Linguistics.

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.

Prasetya Utama, Nafise Sadat Moosavi, Victor Sanh, and Iryna Gurevych. 2021. Avoiding inference heuristics in few-shot prompt-based finetuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9063–9074, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020a. Mind the trade-off: Debiasing NLU models without degrading the in-distribution performance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8717–8729, Online. Association for Computational Linguistics.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020b. Towards debiasing NLU models from unknown biases. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7597–7610, Online. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Adina Williams, Tristan Thrush, and Douwe Kiela. 2022. ANLIzing the adversarial natural language inference dataset. In *Proceedings of the Society for Computation in Linguistics 2022*, pages 23–54, online. Association for Computational Linguistics.

Yuxiang Wu, Matt Gardner, Pontus Stenetorp, and Pradeep Dasigi. 2022. Generating data to mitigate spurious correlations in natural language inference datasets. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2660–2676, Dublin, Ireland. Association for Computational Linguistics.

Yadollah Yaghoobzadeh, Soroush Mehri, Remi Tachet des Combes, T. J. Hazen, and Alessandro Sordoni. 2021. Increasing robustness to spurious correlations using forgettable examples. In *Proceedings*

of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 3319–3332, Online. Association for Computational Linguistics.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A Cross-lingual Adversarial Dataset for Paraphrase Identification. In *Proc. of EMNLP*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. 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 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiang Zhou and Mohit Bansal. 2020. Towards robustifying NLI models against lexical dataset biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8759–8771, Online. Association for Computational Linguistics.