How Hard is this Test Set?
NLI Characterization by Exploiting Training Dynamics

Adrian Cosma[1], **Stefan Ruseti**[1], **Mihai Dascalu**[1], **Cornelia Caragea**[2]

[1]National University of Science and Technology POLITEHNICA Bucharest, București, Romania
[2]University of Illinois at Chicago, Chicago, IL

{ioan_adrian.cosma, stefan_ruseti, mihai_dascalu}@upb.ro
cornelia@uic.edu

Abstract

Natural Language Inference (NLI) evaluation is crucial for assessing language understanding models; however, popular datasets suffer from systematic spurious correlations that artificially inflate actual model performance. To address this, we propose a method for the automated creation of a challenging test set without relying on the manual construction of artificial and unrealistic examples. We categorize the test set of popular NLI datasets into three difficulty levels by leveraging methods that exploit training dynamics. This categorization significantly reduces spurious correlation measures, with examples labeled as having the highest difficulty showing markedly decreased performance and encompassing more realistic and diverse linguistic phenomena. When our characterization method is applied to the training set, models trained with only a fraction of the data achieve comparable performance to those trained on the full dataset, surpassing other dataset characterization techniques. Our research addresses limitations in NLI dataset construction, providing a more authentic evaluation of model performance with implications for diverse NLU applications.

1 Introduction

Natural Language Inference (NLI), or textual entailment (Dagan et al., 2009), has emerged as an enduring challenge in the field of Natural Language Processing for evaluating the Natural Language Understanding (NLU) capabilities of models. Persistently, NLI remains a difficult problem, as it implies reasoning across several linguistic phenomena to determine the logical relationship (i.e., entailment, contradiction, or neutral) between two documents - a premise and a hypothesis. The capability to accurately infer relationships between sentences is crucial for a wide range of applications, such as question answering (Demszky et al., 2018), dialogue systems (Welleck et al., 2019), and fact-checking (Thorne et al., 2018; Stab et al., 2018).

Since its inception (Dagan et al., 2005), several large-scale benchmark datasets have been proposed for NLI (Bowman et al., 2015; Williams et al., 2018; Nie et al., 2019); however, in time, the most ubiquitously used are the Stanford Natural Language Inference (SNLI) (Bowman et al., 2015) and the MultiNLI datasets (Williams et al., 2018), which have played a pivotal role in advancing the state of the art (Storks et al., 2019).

However, multiple works (Liu et al., 2020; Gururangan et al., 2018; Tsuchiya, 2018; Poliak et al., 2018; Naik et al., 2018; Glockner et al., 2018) pointed out several critical limitations in these datasets, stemming from systematic annotation errors and spurious correlations that impact both the training and test sets. A critical consequence of these issues is the inflation of model performance, leading to seemingly high results (Naik et al., 2018; Liu et al., 2020) that may not generalize well to real-world scenarios. For example, a widely used RoBERTa model (Liu et al., 2019) trained *solely on the hypothesis* achieves an unreasonable accuracy of 71.7% on SNLI and 61.4% on MultiNLI (random chance being 33%), which strongly points towards systematic errors in dataset construction.

In this work, we aim to address the limitations of existing NLI datasets by proposing an automated construction of a more challenging test set. In contrast to previous approaches, we avoid manually creating artificial examples (Naik et al., 2018); instead, we leverage existing samples from the test set. To accomplish this, we generalize dataset cartography (Swayamdipta et al., 2020) to cluster samples in the test set and characterize them into three categories of increasing difficulty. Our approach leverages 8 measures of training dynamics of each premise-hypothesis pair and is inspired by related works in both NLI (Naik et al., 2018; Geiger et al., 2018; Liu et al., 2020) and approaches tackling the problem of learning with noisy data (Pleiss et al., 2020; Swayamdipta et al., 2020). We show that our

method can isolate examples exhibiting spurious correlations and provide a challenging test set. Furthermore, our method is general, model-agnostic, and easily extensible to other datasets (e.g., for fact-checking (Thorne et al., 2018)). Our experiments show that using the same method on the training set enables the aggressive filtering of uninformative examples during training, reducing data quantity but increasing quality, enabling the model to obtain on-par performance on the NLI stress test proposed by Naik et al. (2018), using only a fraction of data. We make our code publicly available[1].

This work makes the following contributions:

1. We denote spurious correlations in the test sets for two popular NLI datasets - SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018) and a fact-checking dataset, repurposed for NLI: FEVER (Thorne et al., 2018). We show statistically significant correlations between the performance of models and the presence of several measures of spurious correlations across labels.

2. We propose a general method for creating a strong test set for NLI. Using a multitude of training dynamics features of samples in an existing test set, our method automatically characterizes examples in the test set into three increasing difficulty levels, which strongly correlate with decreased model performance. Our method minimizes spurious correlations, providing a more accurate measure of model performance in the real world on NLU tasks. Our method is model-independent and the underlying difficulty splits generalize across models.

3. The same procedure applied to the training data achieves similar performance on the test set while using only 33% of the available data for SNLI and 59% for MultiNLI, surpassing other dataset characterization methods (Pleiss et al., 2020; Swayamdipta et al., 2020), indicating that our approach can be used as a strong method for increasing data quality.

The paper is structured as follows. After emphasizing the shortcomings of existing NLI datasets and presenting various stress tests, we introduce our method for test set characterization. Then, we present the main results, a comparison with a different encoder to argue that our approach is model-

agnostic, and an analysis supporting the viability of our approach as an alternative to training set characterization. The paper ends with conclusions and limitations.

2   Related Work

Across the development of natural language inference and understanding systems, multiple large-scale training and testing datasets have been developed over different linguistic domains. Initially, progress was driven by the addition of SNLI (Bowman et al., 2015), but several other variants have been proposed, such as MultiNLI (Williams et al., 2018), containing multiple domains, SciNLI (Sadat and Caragea, 2022) for scientific question answering, SQuAD (Rajpurkar et al., 2016) and GLUE (Wang et al., 2018) benchmarks for general-purpose NLU. Moreover, many related problems in NLU can be cast as an NLI problem; for instance, the FEVER (Thorne et al., 2018) dataset for fact-checking can be regarded as an NLI problem in terms of identifying the relationship between a statement and supporting evidence.

However, driven by the widespread observations that previous popular NLI datasets contain shortcuts (Tsuchiya, 2018; Gururangan et al., 2018), multiple works (Nie et al., 2019; Glockner et al., 2018; Naik et al., 2018; Geiger et al., 2018; Yanaka et al., 2019; Saha et al., 2020) developed "stress tests" to benchmark specific linguistic phenomena.

For instance, Glockner et al. (2018) proposed a simple test set based on SNLI (Bowman et al., 2015) that involves changing a single word in the premise sentences. In this setting, performance is substantially worse than the original SNLI test set, indicating the presence of spurious correlations in the training dataset construction. Naik et al. (2018) proposed an NLI Stress Test by quantifying the lexical phenomena (e.g., presence of antonyms, numerical reasoning) behind common model errors in MultiNLI (Williams et al., 2018). Their proposed stress test involved constructing artificial examples that exacerbate common sources of model error, showcasing that some models have markedly reduced performance. Nie et al. (2019) proposed a new benchmark called Adversarial NLI (ANLI), which leverages an interactive human-and-model-in-the-loop procedure to collect hard examples for natural language inference, obtaining a challenging stress test for current models.

Geiger et al. (2018) constructed a dataset of artifi-

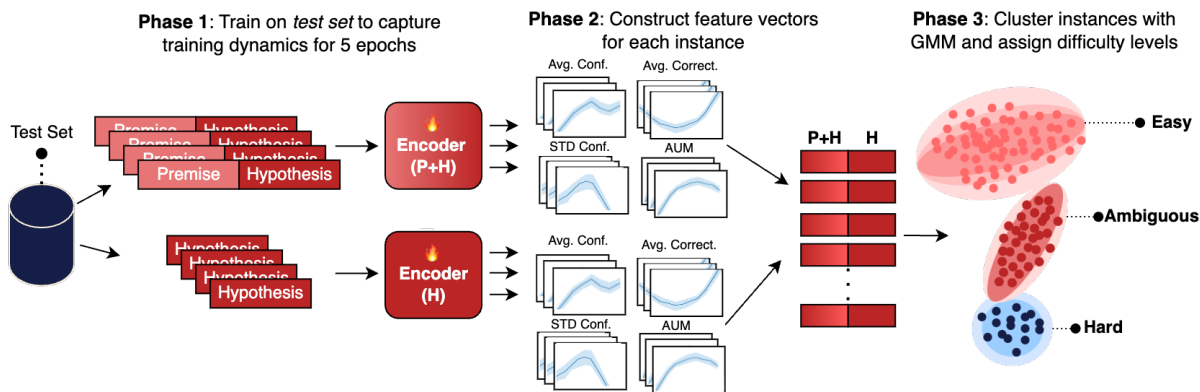---

[1]https://github.com/cosmaadrian/nli-stress-test

Figure 1: Overall diagram of our method to automatically construct a challenging test set for NLI.

cially built sentences based on first-order logic, fixing the sentence structure and only varying words corresponding to parts of speech at predefined positions. Such a dataset comprises unrealistic sentences but provides insight into the (lack-of) expressive power of certain neural architectures. Likewise, Yanaka et al. (2019) proposed the evaluation of monotonicity reasoning for NLI by constructing a dataset through curating and manipulating sentence pairs from the Parallel Meaning Bank (Abzianidze et al., 2017). Saha et al. (2020) identified the lack of conjunctive reasoning examples in current NLI test sets and estimated that around 72% of sentence pairs in SNLI have conjunctions unchanged between premise and hypothesis. The authors proposed CONJNLI, a stress test composed of conjunctive sentence pairs collected automatically from Wikipedia and manually verified.

In contrast to previous work, we propose a method to characterize the test set into multiple difficulty levels by using training dynamics of neural networks (Swayamdipta et al., 2020; Pleiss et al., 2020), thereby isolating easy and spurious examples and keeping only challenging pairs. Our method is general, model-agnostic, utilizes existing dataset samples (avoiding unrealistic artificial sentence pairs), is easily extensible to other datasets, and does not require manual verification of human annotators. Previous approaches (Naik et al., 2018; Saha et al., 2020) aim to develop a stress test for NLI by amplifying spurious correlations and evaluating model performance under various extreme conditions. Our goal is to minimize spurious correlations in existing benchmarks to gain a more realistic sense of performance under challenging real-world examples.

3 Test Set Characterization

Our goal is to generalize the Data Maps proposed by Swayamdipta et al. (2020) to characterize the test set. Swayamdipta et al. (2020) proposed an approach to gauge the contribution of each training sample in a dataset by analyzing training dynamics (variability, average confidence of the gold label, and average correctness) across training for a fixed amount of epochs. After training, each example is split into one of three categories (i.e., *easy-to-learn*, *ambiguous* or *hard-to-learn*) using a fixed percentile threshold on one of the features. For example, instances regarded as *ambiguous* are examples for which the variability across 5 epochs is in the top 33% percentiles, disregarding other measures.

We aim to extend and generalize Data Maps by employing a Gaussian Mixture Model (GMM) (Reynolds, 2009) to learn the best fitting distribution of data difficulty levels, thus avoiding fixed thresholds. Unlike other clustering techniques, such as KMeans, which outputs spherical clusters and disregards cluster variance, we chose a GMM as a more flexible clustering method. Figure 1 showcases the general methodology used in this work. We first characterize the test set by training two separate models with both premise and hypothesis (P+H), and hypothesis only (H); second, we gather 8 measures of training dynamics for each instance and cluster them to obtain three difficulty levels (4 for P+H and 4 for H). We found that difficulty levels simultaneously align with measures of spurious correlations and model performance.

In contrast to the initial approach of Swayamdipta et al. (2020), we include 6 additional features for a more informative characterization across training. In our scenario focused

on NLI in particular, we gather statistics for training dynamics across two types of settings: normal training (i.e., training with P + H) and hypothesis-only (H). Models trained only with the hypothesis have been shown to produce unreasonably high results (Poliak et al., 2018; Liu et al., 2020), mostly due to artifacts in dataset construction. These insights enabled us to gather statistics about such examples and improve data characterization through more diverse features for each instance. Different from Swayamdipta et al. (2020) who focused on characterizing the *training* set for increasing data quality, we aim to construct a more challenging *test set* automatically.

As such, in order to construct a data map of the test set, we trained a model for E epochs on the *test set* using premise and hypothesis and, separately, using only the hypothesis to gather training dynamics for each example in the test set. Let $D_{test} = \{(x, y^*)_i\}_{i=1}^{N}$ be a test dataset containing N instances, where, in our case, x_i is comprised of a premise + hypothesis pair or only a hypothesis. We compute the following measures across training an Encoder model in both scenarios (P + H and H) for each example x_i: confidence ($\hat{\mu}_i$), variability ($\hat{\sigma}_i$), correctness ($\hat{c}_i$) and Area Under Margin (AUM_i).

$$\hat{\mu}_i = \frac{1}{E} \sum_{e=1}^{E} p_{\boldsymbol{\theta}^{(e)}}(y_i^* | \boldsymbol{x_i}) \qquad (1)$$

$$\hat{\sigma}_i = \sqrt{\frac{\sum_{e=1}^{E}(p_{\boldsymbol{\theta}^{(e)}}(y_i^* | \boldsymbol{x_i}) - \hat{\mu}_i)}{E}} \qquad (2)$$

$$\hat{c}_i = \frac{1}{E} \sum_{e=1}^{E} [\text{argmax}(p_{\boldsymbol{\theta}^{(e)}}(x_i)) = y_i^*] \qquad (3)$$

where $p_{\boldsymbol{\theta}^{(e)}}$ corresponds to the model's probability during training at epoch e. Following Swayamdipta et al. (2020), we compute confidence, variability, and correctness concerning the correct label y_i^*. Furthermore, we compute AUM (Pleiss et al., 2020), which was initially proposed to identify mislabeled examples but yielded a similar type of characterization as Data Maps. We include AUM as an additional measure of instance correctness/learnability. Let $z_y^{(e)}(x_i)$ be the logit (pre-softmax) for class y of the model at epoch e, given an instance x_i. The area under margin (AUM or average margin) of x_i is computed as:

$$AUM_i = \frac{1}{E} \sum_{e=1}^{E} (z_{y_i^*}^{(e)}(x_i) - \max_{(y \neq y_i^*)} z_y^{(e)}(x_i)) \quad (4)$$

In all our experiments, we first fine-tune pre-trained RoBERTa models (Liu et al., 2019), followed by DeBERTa (He et al., 2022) models due to their established high performance on a wide set of tasks. In our formulation, any other encoder would yield similar results, as this method is based only on the final classification output and not model internals. However, characterization based on final logits and class confidences is affected by how calibrated the models' predictions are (Guo et al., 2017), as poorly calibrated models have lower logit variance across classes. For our scope, we are interested in identifying and separating spurious correlations in NLI benchmarks and not in benchmarking different classifiers for this task. We explore the impact of the underlying encoder in Section 4.1.

In Table 1, we show the results of our RoBERTa models trained on SNLI, MultiNLI, and FEVER on different configurations of training/testing splits and using both the premise and the hypothesis or only the hypothesis. Our reproduction of results is on par with other works. For completeness, we also show results where the model is trained on the test set, but note that the purpose is only to gather training dynamics and not directly use it as a classifier. The model trained on only the hypothesis obtains 71% accuracy on SNLI and 61% accuracy on MultiNLI, while random chance performance is 33%. These results strongly point toward spurious correlations and annotation artifacts on both datasets (Tsuchiya, 2018; Gururangan et al., 2018; Poliak et al., 2018; Liu et al., 2020). In the case of FEVER, the hypothesis-only model achieved close to random-chance, indicating less spurious correlations found in the hypothesis.

Dataset	Train Split	Test Split	Accuracy (P + H)	Accuracy (H)
SNLI	Train	Test	0.9178	0.7170
	Test	Test	0.9799	0.8764
MultiNLI	Train	Val	0.8773	0.6142
	Val	Val	0.9841	0.8612
FEVER	Train	Dev	0.7702	0.3822
	Dev	Dev	0.9459	0.7137

Table 1: Results for RoBERTa on SNLI, MultiNLI and FEVER under various train/test splits and input types.

After training two classifiers (P + H and H) on each dataset's test set, we construct a feature vector f_i describing the training dynamics of an instance

i by concatenating the training dynamics of each sample in both settings:

$$f_i^{(P+H)} = [\hat{\mu}_i^{(P+H)}, \hat{\sigma}_i^{(P+H)}, \hat{c}_i^{(P+H)}, AUM_i^{(P+H)}]$$
$$f_i^{(H)} = [\hat{\mu}_i^{(H)}, \hat{\sigma}_i^{(H)}, \hat{c}_i^{(H)}, AUM_i^{(H)}]$$
$$f_i = f_i^{(P+H)} \parallel f_i^{(H)}$$
$$(5)$$

Using the feature vectors $\{f_i\}_{i=1}^N$, we cluster the test set using a Gaussian Mixture Model into three clusters. Feature vectors are normalized with standard scaling by subtracting the mean and dividing by the standard deviation of each feature. The clusters are ranked according to the intra-cluster average confidence $\hat{\mu}^{(P+H)}$, and we interpret them as belonging to three difficulty levels, following the terminology introduced by Swayamdipta et al. (2020): *easy*, *ambiguous* and *hard*, in decreasing order of the average intra-cluster confidence $\overline{\hat{\mu}}^{(P+H)}$. See Appendix A for a high-level overview of our algorithm.

Figure 2 depicts the distribution of features across difficulty levels for both datasets. While each type of feature captures different aspects of the learnability of an instance, their combination offers a more diverse view of the learning dynamics during training. In the case of both SNLI and MultiNLI, harder examples have consistently lower average margin and more variability of the correct class. The effect is not as pronounced in a hypothesis-only setting; however, there is a clear delimitation of easy examples for average margin and confidence, indicating potential annotation artifacts. For FEVER, since the dataset has reduced spurious correlations, the identified splits correspond to difficult-to-learn examples, not necessarily examples with annotation artifacts.

In the interest of quantifying the number of spurious correlations found in the test set, we follow Naik et al. (2018) and track several measures that correspond to either shallow statistics between the premise and the hypothesis, or the presence of negations or misspelled words. Table 2 presents the heuristics implemented in our work. We automatically compute each measure, avoiding time-consuming manual annotations.

Training the pretrained models for dataset characterization was performed for 5 epochs each, with a batch size of 32, using the Adam optimizer with a learning rate of 10^{-5} following a linear decay

Name	Explanation
Word Overlap	Number of common words between the premise and hypothesis, normalized by sentence length
Number of Antonyms	Number of antonyms of each of the words in the premise contained in hypothesis, based on WordNet (Fellbaum, 1998), normalized by sentence length.
Length Mismatch	Difference in length between premise and hypothesis, normalized by sentence length
Misspelled Words	Total number of misspelled words using a spellchecker in the premise and hypothesis, normalized by sentence length.
Contains Negation	Boolean flag if either the premise of hypothesis contains a negation word (e.g., *no, not, never, none*)

Table 2: Heuristic measures of spurious correlations used, similar to the categories by Naik et al. (2018).

Dataset	Split	Fraction of total	Accuracy (P + H)	Accuracy (H)
SNLI	Easy	0.70 *(6889 / 9824)*	0.97	0.82
	Amb.	0.17 *(1725 / 9824)*	0.89	0.46
	Hard	0.12 *(1210 / 9824)*	0.56	0.38
MultiNLI	Easy	0.75 *(7381 / 9815)*	0.94	0.67
	Amb.	0.14 *(1420 / 9815)*	0.75	0.43
	Hard	0.10 *(1014 / 9815)*	0.53	0.39
FEVER	Easy	0.50 *(10083 / 19998)*	0.95	0.43
	Amb.	0.24 *(4903 / 19998)*	0.88	0.49
	Hard	0.25 *(5012 / 19998)*	0.29	0.16

Table 3: Performance of RoBERTa models trained on the training set and evaluated on different splits of our stress test.

schedule with warm-up. We used mixed precision for all training runs.

4 Results & Discussion

Table 3 shows the performance of a RoBERTa model trained on each dataset's training set and evaluated on our stress test after characterization using training dynamics. Easier instances have more examples annotated with "contradiction" and "entailment", while harder instances have more examples annotated with "neutral". Performance monotonously degrades upon increasing difficulty levels, reaching 56% accuracy on SNLI-hard and 53% accuracy on MultiNLI-hard. Performance on the *easy* split for both datasets is considerably higher compared to the global accuracy with all splits combined. Furthermore, the accuracy of a model trained using only the hypothesis degrades to almost random chance on harder splits, indicating that the *hard* split has fewer annotation artifacts. Compared to Swayamdipta et al. (2020), the dif-
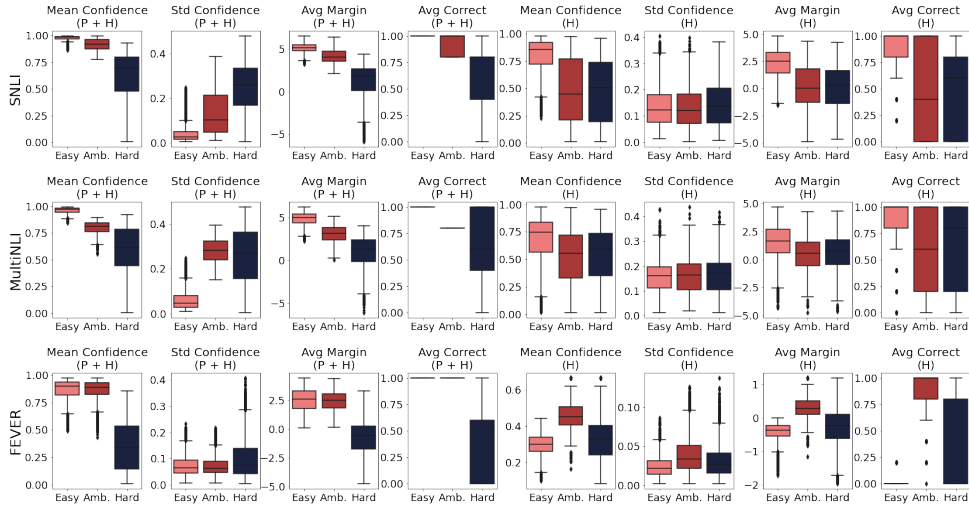
Figure 2: Distributions of feature values across difficulty levels for the test set for SNLI (top), MultiNLI (middle), and FEVER (bottom). In addition to features explored in Data Maps (Swayamdipta et al., 2020), we also incorporated the Average Margin (Pleiss et al., 2020) and included training dynamics across a model trained only on the hypothesis.
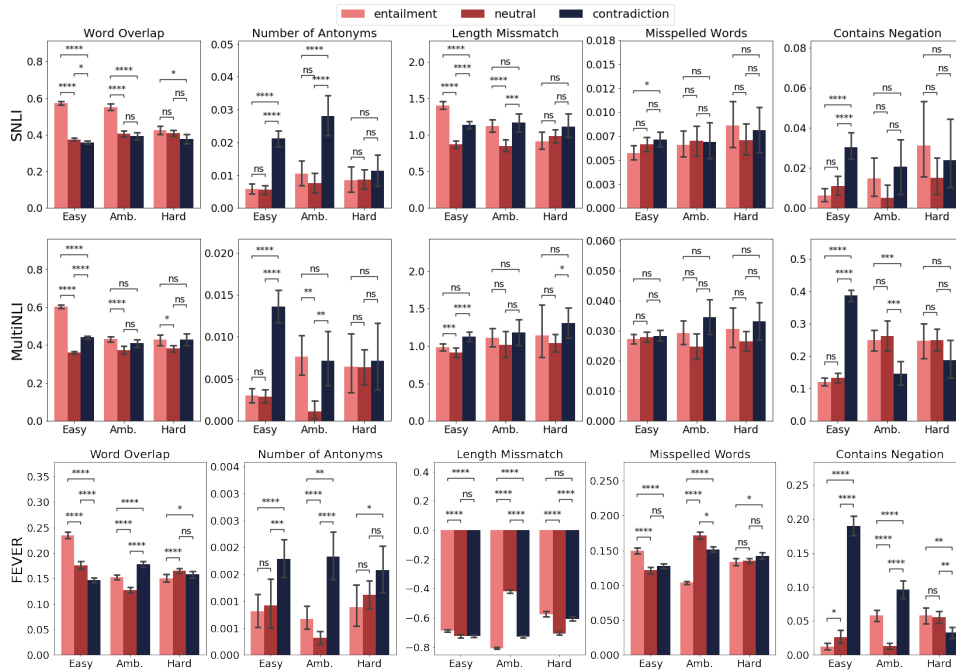

Figure 3: Distributions of the measures of spurious correlations for each level (*easy*, *ambiguous*, *hard*) across the three labels (*entailment*, *neutral*, *contradiction*) for SNLI (top), MultiNLI (middle) and FEVER (bottom).

ficulty levels are not equal in size; the majority (∼70%) of samples belongs to the *easy* category, while only around 10% are characterized as being *hard*. For FEVER, the performance degradation in the hard split is more dramatic: 29% accuracy for *hard* compared to 88% for ambiguous, indicating truly difficult examples for the model.

Even though the *hard* split is relatively small, the subsample is challenging for current models, as it comprises instances with fewer spurious cor-

relations between premise and hypothesis. Fewer annotation artifacts enable fewer "correct" predictions from linguistic patterns present only in the hypothesis. In Figure 4, we show per-class counts relative to the difficulty levels for both datasets. It is the case that easier samples contain more contradictions and entailments, prone to linguistic commonalities (word overlap, presence of antonyms). Thus, the *hard* split has more neutral instances.

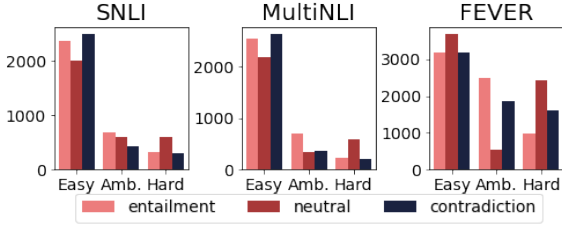Through manual inspection, we found that for

Figure 4: Counts for each class in SNLI, MultiNLI, and FEVER, according to each difficulty level.

SNLI, the *easy* splits contain unrelated sentences which are sometimes annotated incorrectly as Contradiction (e.g., "a woman running in the park" versus "a man cooking at home" - two unrelated sentences annotated as contradicting). The model learns this pattern and incorrectly predicts Contradiction on some Neutral pairs (e.g., "... girls chatting on the stairwell" versus "girls are at school"). For MultiNLI, we found that the easy split usually aligns with simple sentence negations (e.g., "it gets it" versus "it doesn't get it") or paraphrasing ("I guess history repeats itself" versus "history certainly doesn't repeat"). These observations strongly point towards spurious correlations between premise and hypothesis, making the sentences easier to classify correctly. We provide selected examples in the Appendix A. The *ambiguous* and *hard* splits in both datasets contain increasingly more subtle cues, with little overlap in words between premise and hypothesis (e.g., "standing on a tree log" versus "crossing the stream" / "wouldn't have mattered" versus "would have gotten worse"), having more natural and challenging sentence pairs.

Across SNLI, MultiNLI, and FEVER, we tracked the average amount of each measure between difficulty levels and classes (see Figure 3). To rigorously test the difference between the classes at various difficulty levels, we perform a non-parametric two-sided Mann-Whitney-U test (Mann and Whitney, 1947) with Bonferroni correction to test for statistical significant differences[2].

We found no evidence for the presence of spurious correlations (Mann-Whitney's U test $p > .05$) in the *hard* split between the three classes. Some measures are more associated with certain classes. For example, instances annotated with Contradiction have a disproportionate amount of antonyms

between premise and hypothesis in the *easy* and *ambiguous* splits. Similarly, negation is more present in the Contradiction class for easy splits. For instances annotated with Entailment, word overlap is present significantly in *easy* splits. Between SNLI and MultiNLI, MultiNLI has a disproportionately large amount of negations compared to SNLI. For both SNLI and MultiNLI, our method yields little to no significant differences between classes in the *hard* split across the spurious correlation measures. For FEVER, measures such as the presence of negations, number of antonyms, and word overlap are reduced across difficulty levels. Note that FEVER includes a small statement as the premise and a long text extract containing evidence as the hypothesis, which makes the length mismatch negative.

4.1 Impact of the Underlying Encoder

To show that our method is model-agnostic, we further provide a comparison between the dataset characterizations obtained by RoBERTa and DeBERTa. Table 4 showcases the accuracies of the two models on each others' data characterizations. The difficulty splits are maintained cross-model.

Split	Dataset	Target Model	Source: RoBERTa Accuracy	Source: DeBERTa Accuracy
Easy	SNLI	RoBERTa	0.9779	0.9462
		DeBERTa	0.9792	0.9624
	MultiNLI	RoBERTa	0.9470	0.9502
		DeBERTa	0.9545	0.9567
	FEVER	RoBERTa	0.9101	0.9346
		DeBERTa	0.8967	0.9501
Ambiguous	SNLI	RoBERTa	0.8916	0.9802
		DeBERTa	0.9003	0.9881
	MultiNLI	RoBERTa	0.7577	0.9086
		DeBERTa	0.7746	0.9543
	FEVER	RoBERTa	0.9532	0.8697
		DeBERTa	0.9470	0.8697
Hard	SNLI	RoBERTa	0.5678	0.6337
		DeBERTa	0.6446	0.6437
	MultiNLI	RoBERTa	0.5375	0.7497
		DeBERTa	0.6460	0.7585
	FEVER	RoBERTa	0.3444	0.2913
		DeBERTa	0.4089	0.2977

Table 4: Comparison between RoBERTa and DeBERTa accuracy on each difficulty level, across models.

Across datasets and difficulty levels, the performance sharply drops for the *"hard"* split for both models. DeBERTa achieved higher accuracy for *"hard"* set, most likely due to better overall performance compared to RoBERTa. In Figure 5, we show that overall heuristic values for "Contains Negation" are maintained across both models. Extended results for are presented in Appendix A.

Our proposed methodology is general and independent of the underlying encoder model since we process training dynamics computed from raw logit
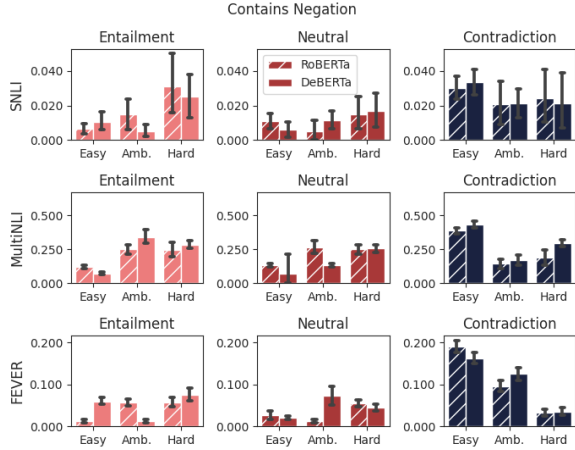
Figure 5: Comparison between the characterizations obtained by RoBERTa and DeBERTa on the "Contains Negation" heuristic measure.

scores. This characterization procedure may be adapted to using Large Language Models (LLMs) (Lee et al., 2023) in a zero-shot classification setting by manipulating the log-likelihood for the tokens of the correct classes. However, using LLMs requires a different approach than the one presented here since the networks are usually used without further training, in an in-context-learning manner (Dong et al., 2022). Furthermore, even if the LLMs are fine-tuned (Hu et al., 2021), it is not straightforward how the logits of each of the three classes are tracked across training. We leave this approach for future work.

4.2 Training Set Characterization

Method	Splits	% SNLI	Stress Test by Naik et al. (2018)						
			Word Overlap	Spelling Error	Numerical Reasoning	Negation	Length Mismatch	Antonym	Overall
DataMaps	Amb.	33%	0.6792	0.7202	0.3745	0.5388	0.7176	0.6237	0.6662
AUM	Amb.	33%	0.6818	0.6904	0.2896	0.4837	0.7134	**0.7417**	0.6416
Ours	Easy	53%	0.6720	0.6713	<u>0.4287</u>	0.4622	0.6870	0.4446	0.6246
	Easy+Amb	86%	0.7231	**0.7500**	**0.4452**	<u>0.6124</u>	**0.7578**	0.6582	<u>0.7083</u>
	Amb	33%	<u>0.7264</u>	<u>0.7414</u>	0.4158	**0.6666**	0.7491	0.6515	**0.7093**
	Amb+Hard	46%	**0.7302**	0.7305	0.3761	0.5533	0.7471	<u>0.6619</u>	0.6860
	Easy+Hard	66%	0.6714	0.6815	0.3803	0.5617	0.6993	0.4971	0.6442
	Hard	13%	0.3191	0.3310	0.3304	0.3184	0.3189	0.0	0.3177
	All	100%	0.7252	0.7559	0.2794	0.5923	0.7771	0.7268	0.7037

Table 5: Results for a RoBERTa model trained on SNLI in various configurations and evaluated on the stress test by Naik et al. (2018) based on MultiNLI. The best results are **bold**, while the second best are <u>underlined</u>.

Our method provides a more challenging test set devoid of shortcuts and spurious correlations. Further, we explore the possibility of using this approach to improve data quality for training NLI models. We employ the same algorithm to characterize the *training sets* for SNLI and MultiNLI and train a RoBERTa model on the different resulting combinations of difficulty levels. Under each

Method	Splits	% MultiNLI	Stress Test by Naik et al. (2018)						
			Word Overlap	Spelling Error	Numerical Reasoning	Negation	Length Mismatch	Antonym	Overall
DataMaps	Amb.	33%	<u>0.7046</u>	0.7938	0.4065	0.5469	0.8197	0.5712	0.7222
AUM	Amb.	33%	0.6428	0.7998	0.2888	<u>0.5638</u>	0.8254	0.5129	0.7115
Ours	Easy	41%	0.6885	0.7978	0.3391	0.5514	0.8210	0.5274	0.7179
	Easy+Amb.	84%	0.6601	<u>0.8274</u>	**0.4777**	**0.5764**	<u>0.8467</u>	**0.6330**	<u>0.7459</u>
	Amb.	43%	0.6283	0.8235	0.4602	0.5578	0.8421	0.5742	0.7337
	Amb+Hard	59%	0.6961	**0.8255**	<u>0.4651</u>	0.5628	**0.8476**	<u>0.6288</u>	**0.7474**
	Easy+Hard	56%	**0.7170**	0.8032	0.2948	0.5607	0.8258	0.5062	0.7237
	Hard	15%	0.4525	0.4966	0.2705	0.4589	0.4914	0.2506	0.4656
	All	100%	0.6701	0.8308	0.5380	0.5654	0.8500	0.6312	0.7511

Table 6: Results for a RoBERTa model trained on MultiNLI in various configurations and evaluated on the stress test proposed by Naik et al. (2018) based on MultiNLI. The best results are **bold**, while the second best are <u>underlined</u>.

configuration, the model is trained for 10 epochs with early stopping on the validation set loss, with a learning rate of 10^{-5} following a linear decay schedule with a warm-up.

We evaluate each model on the stress test proposed by Naik et al. (2018) that is based on MultiNLI. However, we emphasize that the stress test of Naik et al. (2018) is designed to unrealistically amplify spurious correlations to gauge the model performance under various extreme conditions, in contrast to our method, which eliminates linguistic shortcuts while mimicking real-world examples. In Tables 5 and 6, we show the performance on the dataset proposed by Naik et al. (2018) for RoBERTa models trained on SNLI and MultiNLI. The authors provided metadata for each instance that allows fine-grained evaluation under different linguistic reasoning phenomena.

We compared our approach with Data Maps (Swayamdipta et al., 2020) and Area Under Margin (Pleiss et al., 2020), two popular methods for training set characterizations using training dynamics. For Data Maps, we select *ambiguous* examples by keeping the instances where average variability is in the top 66% percentile. For AUM, while the authors did not explicitly propose a threshold for characterizing each instance, we follow a similar approach to Data Maps by considering *ambiguous* examples to have an average margin between the 33% and 66% percentiles. Our method outperforms Data Maps and AUM across the majority of settings and, in some cases, outperforms a model trained on the full dataset while using a smaller amount of data but of higher quality. This indicates that our method is a viable alternative to AUM or Data Maps for increasing dataset quality.

5   Conclusions

Our method highlights significant shortcomings in widely used NLI evaluation datasets (SNLI and MultiNLI) due to spurious correlations in the annotation process. To address these issues, we proposed an automatic method for constructing more challenging test sets, effectively filtering out problematic instances and providing a more realistic measure of model performance. Our approach, which categorizes examples in increasing difficulty levels using a wide range of training dynamics features, enhances evaluation reliability and offers insights into underlying challenges in NLI. Importantly, our methodology is general and model-agnostic, and can be applied across different datasets and models, promising improved evaluation practices in NLP.

Furthermore, we provided evidence that our method can obtain a challenging test set even if the dataset has fewer annotation artifacts; we characterized FEVER, a fact-checking dataset repurposed for NLI, and showed that the identified hard split is a highly challenging subset of the dataset. By aggressively filtering uninformative examples, we show that comparable model performance can be achieved with significantly reduced data requirements. Our work contributes to advancing NLI evaluation standards, fostering the development of more robust NLU models.

Limitations

Our method is unsuitable for automatically identifying mislabeled examples in a dataset. While it does incorporate measures such as Area Under Margin (Pleiss et al., 2020), designed with this purpose, proper manual verification is needed to increase annotation quality.

Acknowledgements

References

Lasha Abzianidze, Johannes Bjerva, Kilian Evang, Hessel Haagsma, Rik van Noord, Pierre Ludmann, Duc-Duy Nguyen, and Johan Bos. 2017. The Parallel Meaning Bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. In *Proceedings of EACL*, pages 242–247, Valencia, Spain. 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 (EMNLP)*. Association for Computational Linguistics.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2009. Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering*, 15(4):i–xvii.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment*, MLCW'05, page 177–190, Berlin, Heidelberg. Springer-Verlag.

Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. Transforming question answering datasets into natural language inference datasets. *CoRR*, abs/1809.09922.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA.

Atticus Geiger, Ignacio Cases, Lauri Karttunen, and Christopher Potts. 2018. Stress-testing neural models of natural language inference with multiply-quantified sentences. *arXiv preprint arXiv:1810.13033*.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

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 NA-ACL*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2022. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

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

Noah Lee, Na Min An, and James Thorne. 2023. Can large language models capture dissenting human voices? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4569–4585.

Tianyu Liu, Xin Zheng, Baobao Chang, and Zhifang Sui. 2020. Hyponli: Exploring the artificial patterns of hypothesis-only bias in natural language inference. *arXiv preprint arXiv:2003.02756*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

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.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*.

Geoff Pleiss, Tianyi Zhang, Ethan Elenberg, and Kilian Q Weinberger. 2020. Identifying mislabeled data using the area under the margin ranking. *Advances in Neural Information Processing Systems*, 33:17044–17056.

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.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of EMNLP*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Douglas A. Reynolds. 2009. Gaussian mixture models. In *Encyclopedia of Biometrics*.

Mobashir Sadat and Cornelia Caragea. 2022. SciNLI: A corpus for natural language inference on scientific text. In *Proceedings of ACL*, pages 7399–7409, Dublin, Ireland. Association for Computational Linguistics.

Swarnadeep Saha, Yixin Nie, and Mohit Bansal. 2020. Conjnli: Natural language inference over conjunctive sentences. *arXiv preprint arXiv:2010.10418*.

Christian Stab, Tristan Miller, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources using attention-based neural networks. *arXiv preprint arXiv:1802.05758*.

Shane Storks, Qiaozi Gao, and Joyce Y Chai. 2019. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. *arXiv preprint arXiv:1904.01172*.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. *arXiv preprint arXiv:2009.10795*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of NA-ACL*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of LREC 2018*, Miyazaki, Japan. European Language Resources Association (ELRA).

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy. 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 NA-ACL*, pages 1112–1122. Association for Computational Linguistics.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019. Help: A dataset for identifying shortcomings of neural models in monotonicity reasoning. *arXiv preprint arXiv:1904.12166*.

A Appendix

A.1 Algorithm

We present the high-level overview of our methodology in Algorithm 1.

Algorithm 1 Pseudo-code for the construction of a stress test based on training dynamics.

Require: D_{test} - the target test set

Train encoder model $M_\theta^{(P+H)}$ on D_{test} using the premise and hypothesis for 5 epochs, tracking training dynamics
Compute $\hat{\mu}^{(P+H)}, \hat{\sigma}^{(P+H)}, \hat{c}^{(P+H)}$, AUM$^{(P+H)}$ (Eqs. 1,2, 3, 4)
Construct training dynamics features $f_i^{(P+H)}$ for each instance i (Eq. 5)

Train encoder model $M_\phi^{(H)}$ on hypothesis for E epochs, tracking training dynamics
Compute $\hat{\mu}^{(H)}, \hat{\sigma}^{(H)}, \hat{c}^{(H)}$, AUM$^{(H)}$ (Eqs. 1,2, 3, 4)
Construct training dynamics features $f_i^{(H)}$ for each instance i (Eq. 5)

Concatenate features vectors $f_i = f_i^{(P+H)} \parallel f_i^{(H)}$

Train a Gaussian Mixture Model with 3 clusters on f_i
Order clusters based on average intra-cluster confidence $\overline{\hat{\mu}}$, considering "easy" (e), "ambiguous" (a) and "hard" (h) having $\overline{\hat{\mu}}^{(e)} > \overline{\hat{\mu}}^{(a)} > \overline{\hat{\mu}}^{(h)}$

Split D_{test} into $D_{test}^{(e)}, D_{test}^{(a)}, D_{test}^{(h)}$, based on examples corresponding to each cluster
return $D_{test}^{(e)}, D_{test}^{(a)}, D_{test}^{(h)}$

A.2 Qualitative Samples

In Table 7, we show qualitative examples from each of the three datasets we experimented on.

A.3 Extended Comparison

Figure 6 depicts extended comparisons between the characterizations of the two models on SNLI, MultiNLI, and FEVER.

Dataset	Difficulty	Premise	Hypothesis	True Label	Model Prediction	Correct?
SNLI	Easy	A brown dog plays in a deep pile of snow.	A brown dog plays in snow	Entailment	Entailment	✔
		Woman running in a park while listening to music.	A man cooking at home.	Contradiction	Contradiction	?
		Two daschunds play with a red ball	A cat in a litter box.	Contradiction	Contradiction	?
		A grim looking man with sunglasses pilots a boat.	The happy pilot flies his plane.	Contradiction	Contradiction	?
	Amb.	An older women tending to a garden.	The lady has a garden.	Entailment	Entailment	✔
		People are hiking up a mountain with no greenery.	The hikers have backpacks.	Neutral	Neutral	✔
		A man in a suit speaking to a seated woman.	A man in a costume speaking to another man.	Contradiction	Contradiction	✔
		A helmeted male airborne on a bike on a dirt road.	The man fell off his bike.	Contradiction	Contradiction	?
	Hard	A couple is taking a break from bicycling.	a couple sit next to their bikes.	Neutral	Contradiction	✗
		Three kids in a forest standing on a tree log.	Children crossing stream in forest.	Neutral	Contradiction	✗
		A car is loaded with items on the top.	The car is a convertible..	Contradiction	Neutral	✗
		3 girls chatting and laughing on the stairwell.	Girls are at school.	Neutral	Contradiction	✗
MultiNLI	Easy	Through a friend who knows the lift boy here.	A friend knows the lift boy here.	Entailment	Entailment	✔
		I guess history repeats itself, Jane.	History certainly doesn't repeat, Jane.	Contradiction	Contradiction	✔
		He says men are here.	He said that the men were not here.	Contradiction	Contradiction	✔
		it gets it	It doesn't get it.	Contradiction	Contradiction	✔
	Amb.	He slowed.	He stopped moving so quickly.	Entailment	Neutral	✗
		uh high humidity	Air with increased water content.	Entailment	Entailment	✔
		I don't know all the answers, fella.	Buddy, I just can't answer all those questions.	Entailment	Neutral	✗
		British action wouldn't have mattered.	If Britain got involved, things would have gotten worse.	Contradiction	Neutral	✗
	Hard	Detroit Pistons they're not as good as they were last year	Detroit Pistons played better last year	Entailment	Entailment	✔
		The White House denies this.	The White House, off the record, knows it to be true.	Neutral	Contradiction	✗
		I'm not interested in tactics, Al.	Al is very interested in tactics.	Neutral	Contradiction	✗
		The four Javis children? asked Severn.	You have to ask Severn about the four Jarvis children.	Contradiction	Entailment	✗

Table 7: Selected qualitative examples from SNLI and MultiNLI for each split. In some cases, easy instances for SNLI are mislabeled neutral pairs while for MultiNLI easy instances are simple negations and paraphrasing. The hard split contains sentence pairs with more subtle linguistic cues.
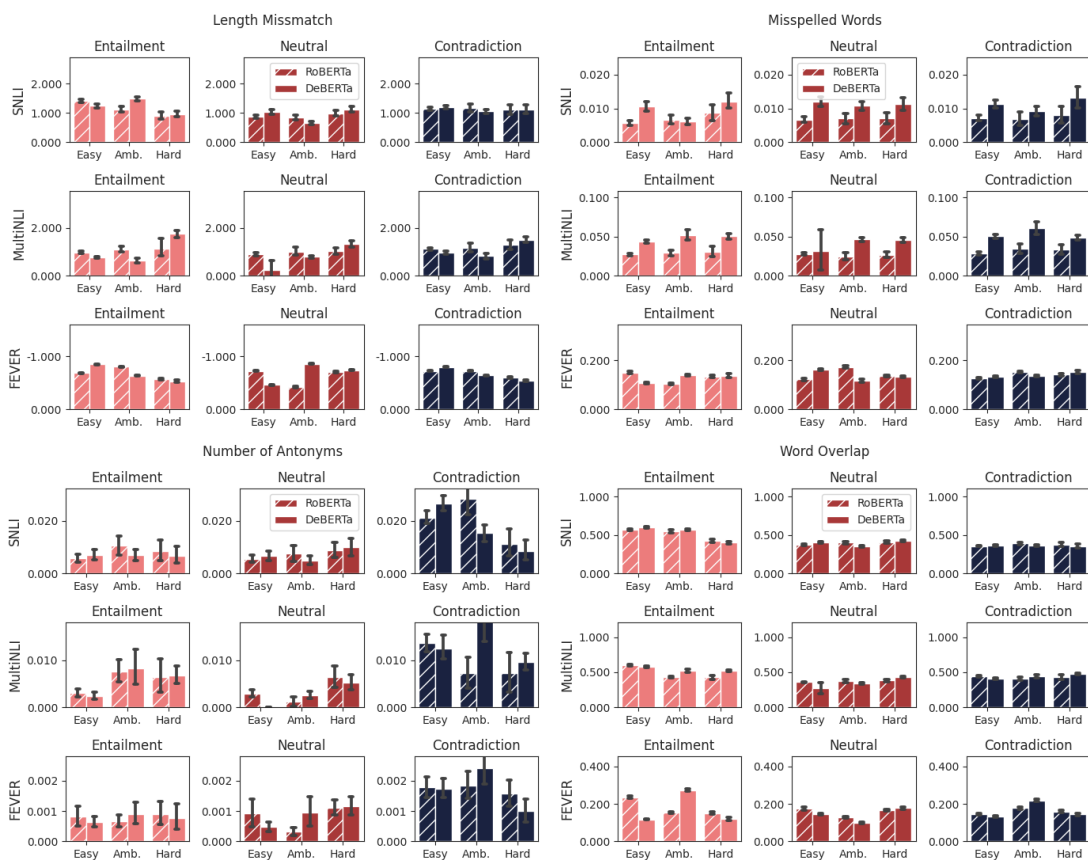

Figure 6: Extended comparison between characterizations given by RoBERTa and DeBERTa across the three datasets for the proposed heuristics.