Towards Fair Text Classification: Expert-Ensemble Debiasing And Adversarial Challenge Sets in NLI

Ayush Kumar (ak45898)
akumar49@utexas.edu

Abhishek Paul (ap62752)
abhishek.paul@utexas.edu

Abstract

We provide a thorough framework for increasing the fairness and reliability of machine learning models in the field of natural language inference (NLP) in this paper. Considering the existence of bias in NLP models and corresponding issues, we used two methods (adversarial challenge sets and ensemble-based debiasing) as key techniques to effectively control the biases we found within our analysis. Since we tackled the final project in a team of two, our team decided to independently test two different approaches to fixing issues we found within the code and data provided to us. The use of adversarial challenge sets is motivated to train the model on data that is intentionally created to fool the NLP model. This can help with robustness by increasing data diversity and making our model resistant to the biases found within the original dataset. The use of ensemble-based debiasing is motivated by the demand for accurate and inclusive models. Bias in NLP models can occur from a variety of sources, making it critical to handle this issue comprehensively. Ensemble-based debiasing incorporates the collective intelligence of several expert models to provide predictions that are less prone to bias and error. Our experiment is divided into steps as follows: expert model(s) selection, training the main model, merging predictions using soft debiased labels, training with debiased labels, and final evaluation. We train the SNLI dataset on the electra-small-discriminator model as our main model and show its performance through assessment and analysis.

1 Introduction

Natural Language Understanding (NLU) is a critical problem in artificial intelligence and machine learning. It involves a wide range of tasks including text classification, sentiment analysis, question answering, and more. Effective NLU models are required for machines capable of comprehending and producing human-like language, making them beneficial to several applications.

Traditional methods frequently fail to successfully eliminate prejudice, particularly in complicated NLP jobs where biases might be firmly embedded. One of the pre-trained models implemented for text classification tasks is the Electra-Small-Discriminator model. Electra models have proved their capability for numerous NLU tasks by exploiting the power of transformers. We examine the performance of the Electra-Small-Discriminator model when used for text classification on the SNLI data set in this research.

The Stanford Natural Language Inference (SNLI) data set has been widely used to evaluate NLU models. It is made up of sentence pairs that have been labeled with one of three categories: ”entailment,” ”neutral,” or ”contradiction.” The SNLI data set is an appropriate option for text classification problems because it provides an excellent testing ground for models that identify the logical relationships between sentences.

To achieve better results with the model and data sets provided, we describe two thorough methods for improving the fairness and robustness of machine learning models via adversarial challenge sets and ensemble-based debiasing.

Our research also found using adversarial challenge sets can be used to remove the biases within the training data sets by training the models on data sets with different perturbations. This inhibits the model from simply exploiting the biases in the data set to achieve high accuracy as the adversarial set purposefully challenges those biases. Our research provides insights into the Electra-Small-Discriminator model with ensemble-based debiasing for increasing classification performance. Ensemble-based debiasing entails merging the predictions of many expert models to provide more robust and less biased collective predictions.

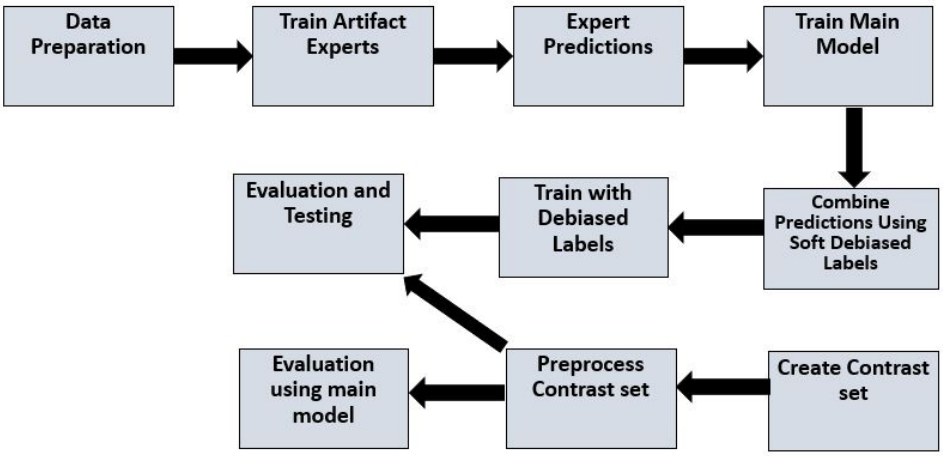

Figure 1: Overview of our Proposed Methodology.

2 Adversarial Challenge Sets

One common problem among recent NLP inference models is that they can achieve high accuracy simply by exploiting data set biases, without needing to get a deep understanding of the language semantics. Two common biases that are intrinsic to most models are:

1. **Word Overlap**: If the premise and the hypothesis sentence have a high word-overlap, then the sentence pair is very likely to be entailment.

2. **Negation bias**: If the hypothesis sentence contains negations words (like not) that often are used to generate contradiction pairs, then the hypothesis is likely to be classified as a contradiction

One way to investigate if our model is susceptible to these biases is by evaluating the model on different NLI "stress tests" (Naik et al., 2018). This is done by augmenting the premise hypothesis pairs with the tautologies "true is true" and "false is not true" which challenge these biases without changing the meaning of the sentence pairs. An illustration of this augmentation can be seen in table 3.

After training the model for one epoch, it was evaluated on SNLI validation set, the negation, and the word overlap nli stress. The accuracies on the data sets are shown in Figure 2 and the distribution of the golden and predicted labels for both stress sets are shown in tables 1 and 2:

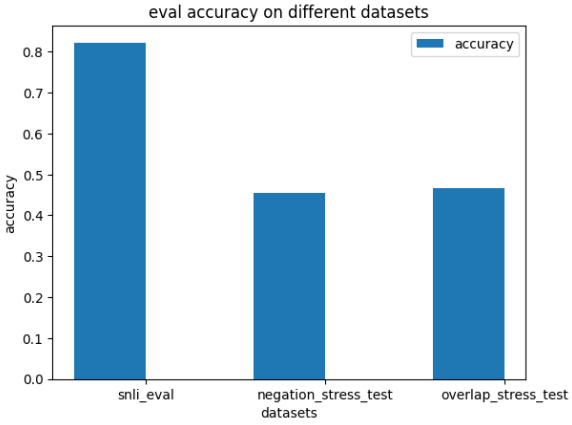

Figure 2: Performance of model on eval and stress datasets

Table 1: Table: Performance of overlap set on default model

	Overlap Set	
	Golden Labels	**Predicted Labels**
Entailment	2773	657
Neutral	2497	5785
Contradiction	2595	1423

Table 2: Table: Performance of negation set on default model

	Negation Set	
	Golden Labels	**Predicted Labels**
Entailment	2773	515
Neutral	2595	6258
Contradiction	2497	3027

From the figure and tables above we can see the introduction of the tautologies into the hypothesis highlights the internal biases of the model as the model struggles to identify proper entailment. In the overlap case, the number of neutral predictions increases significantly as the overall word overlap score decreases. For the negation stress set, the number of neutral and contradiction predictions increases as the introduction of contradictory words (not and false) makes a contradiction prediction more likely.

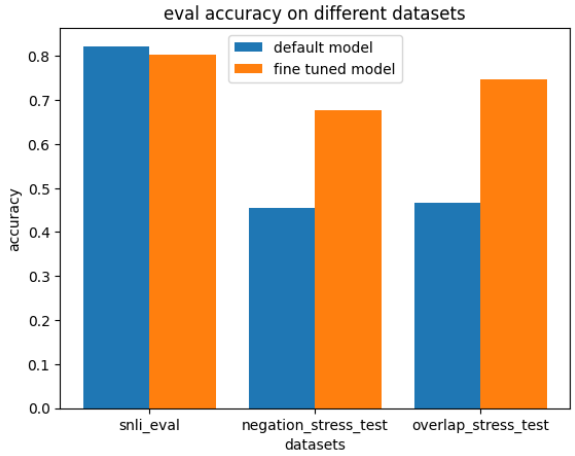

Figure 3: Performance of inoculated model on eval and stress datasets

2.1 Adversarial Data Augmentation

In order to debias the model of the biases seen above, we will implement fine tuning by inoculation (Liu et al., 2019). Given a model trained on the original dataset, we will further train and continue to learn from a small number of examples from the challenge dataset. If the problem lies in biases in the original dataset, then the model should be able to learn from this small subset and perform well on both the original and challenge validation out dataset. If the problem lies somewhere else in the model, then the model's performance will remain unchanged and deteriorate.

In order to implement fine-tuning by inoculation, we took a small subset of the two challenge sets (about 1000 sentence pairs each) and trained our model on the data, making sure to check that the validation accuracy in the original data set does not decrease. The training terminates as soon as there is a degradation of the validation accuracy of the SNLI data.

After inoculating the model against the challenge data sets we evaluated the model on the held-out original and challenge data. The change in accuracies is shown in figure 3.

By learning from the challenge data sets, the model overcame the word overlap and negation biases, that were initially present. This showed that the issue did not lie in the model itself but in the biases that existed within its original training data and the SNLI data set. By augmenting and increasing the diversity of the data, the model could not simply exploit those biases and had to learn to overcome them.

3 Ensemble-Based Approach

We proposed an Ensemble-based debiasing approach in which some artifact experts are trained to learn the correlations. Afterwards, the main model is trained to learn the residual of that model (He et al., 2019). Ensemble-based debiasing ensures the robustness and fairness of the "Electra-small-discriminator" model. This strategy integrates the findings of many models to eliminate biases in data and correlations. It will:

- Effectively addresses biases and correlations present in the training data.
- Make accurate model predictions

Weighted combinations of artifact experts are used in this approach to fine-tune the impact of each expert on the final model predictions. The same dataset is used to train each model.

3.1 Choose and Train Artifact Experts

We start by deciding the architecture and structure of partially trained models. 1st expert model "mDeBERTa-v3-base-xnli-multilingual" (mod, a) is a multilingual model and capable of performing natural language inference (NLI) on 100 languages. This aspect makes it appropriate for multilingual zero-shot classification. Microsoft pre-trained the underlying model on the CC100 multilingual dataset, which contains 100 languages. The model was then fine-tuned using the XNLI and multilingual-NLI-26lang-2mil7 datasets. Both databases include over 2.7 million hypothesis-premise pairs in 27 languages spoken by over 4 billion people in general (mod, a).
Our second expert model "nli-deberta-v3-large" (mod, b) was trained on the SNLI and MultiNLI datasets. For a given sentence pair, it will output three scores corresponding to the labels: contradiction, entailment, and neutral (mod, b). Al-

Category	Premise	Hypothesis
Word Overlap	There was confusion at that moment in the FAA	The FAA was confused at that time and true is true
Negation	There was confusion at that moment in the FAA	The FAA was confused at that and false is not true

Table 3: Shows the bias category, premise, and hypothesis augmentation in order to test model bias in both word overlap and negation

though we performed the experiment of the "bert-base-nil" model it did not provide good results so we discarded it. These models should capture specific biases, correlations, or problematic patterns in data. We choose and train the models from hugging face.

3.2 Obtaining Expert Predictions

We gather predictions for our expert models on an evaluation dataset after training the artifact experts. Each expert model will provide us with a set of predictions.

3.3 Training 'electra-small-discriminator' model

In this step, We train our main model on the dataset.

3.4 Combine Predictions using Soft Debiased Labels

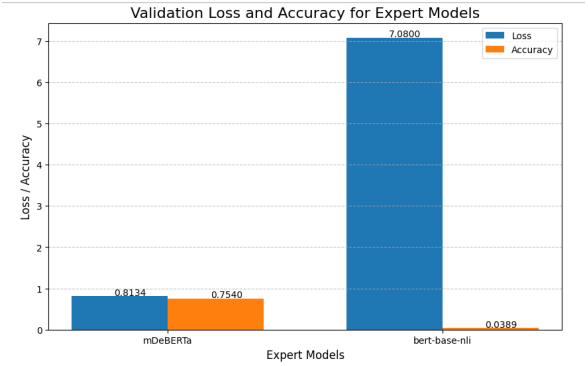

Figure 4: Comparison between mDeBERTa and bert-base-nil Model

We integrate our predictions of the artifact experts and the main model after training them both. This prediction can be combined in two ways. In the first way, we can learn the residual by subtracting the artifact experts' predictions from the main model's predictions. However, this is only for binary classification. The second option is Soft Debiased Labels, in which a weight is assigned to each class (0, 1, and 2 in our case) and has a weight assigned to the main model's predicted as well as weights assigned to each expert's predictions. We use the second method after considering the multi-class classification. We decided the weight of each model based on its performance on the evaluation set before training. Assign higher weights to models that perform better on the validation set, assuming that better performance indicates less bias and high weight. Figure 4 and 5 show the evaluation result to decide the weight associated with each expert model.

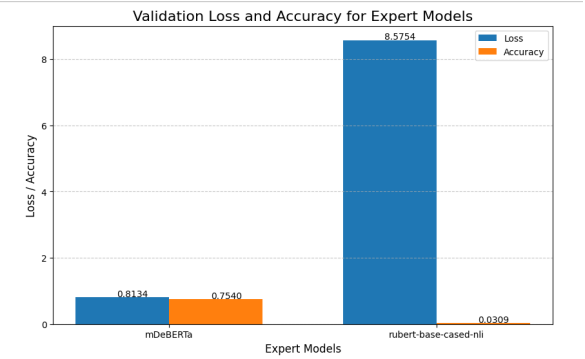

Figure 5: Comparison between mDeBERTa and nli-deberta Model

3.5 Train with Debiased Labels

In the last step, we train our main model using the debiased labels after step 4. The artifact experts' predictions should have minimal impact on the training of the main model. We will modify the loss function here to use the calculated soft debiased labels instead of the original labels.

4 Result and Evaluation

We train our model to see the overall evaluation result. The training process consists of multiple steps and 3 epochs. The output shows the step number, training loss, validation loss, and accuracy at specific intervals. Training begins with an initial training loss of 0.9105, a validation loss

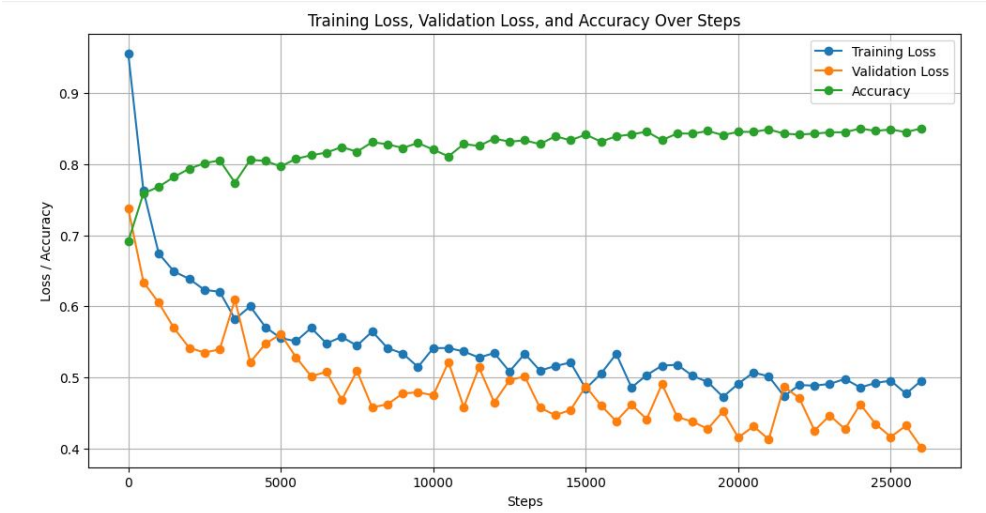

Figure 6: Evaluation on Test result after training on 20,000 train sample and step size is 500 with 3 epochs

of 0.6913, and an accuracy of 0.7257. As training progresses, the training loss decreases, indicating that the model is learning. See the figure 6 to understand the trend in the losses and accuracy. If we keenly observe the validation loss, it decreases, which suggests that the model is generalizing well to the validation data but at some point, it fluctuates which shows not good model training. The accuracy increases over time but is lower than desired. After seeing the evaluation result of the main model using the "Electra-small-discriminator" model. We process our proposed methodology. We can see the evaluation of the expert model in Figure 4 and 5 and conclude that model 1 is better than the other so we give 0.7 weight to expert model 1 and 0.3 weight to expert model 2. Also, we assign 0.7 weight to our main model to generate a combined label.

[100]:

	premise1	hypothesis1	label1	premise2	hypothesis2	label2
0	Some people in blue shirts are standing up at ...	The people are wearing blue.	0	A man with a cigarette in his mouth is riding ...	A man has a cigarette while riding a harley.	1
1	A Mormon missionary walks his bicycle past a M...	The missionary is wearing a helmet.	1	The little boy is wearing white shorts.	Boy has on white shirts.	2
2	Woman running in jeans in a forest on a trail.	A woman is running in a forest.	0	A man is standing with his arms folded looking...	There is only one bed in the room.	2
3	A woman wearing a red and white apron is stand...	A man is standing next to a wall.	2	A player kicks the soccer ball from the corner...	a person playing a sport.	0
4	A brunette man sitting next to a picnic table ...	man at picnic table	0	A little girl wearing a yellow dress moves fra...	Everything is calm at the stadium.	2
...	...	...	...	...	...	...
95	The cylinder's pattern could be described as r...	The cylinder is part of a piping system.	1	Girls in the aftermath of a rainstorm continue...	Girls hitchhiking .	1
96	A man is walking a white dog on a leash in a p...	A dog is wandering looking for it's owner.	2	A dog is running on a pathway.	The dog is standing still on the path.	2
97	A girl going into a wooden building.	Lisa is going to church.	1	A spotted dog stands on his hind legs to catch...	On the beach a dog play the ball.	1
98	A person with a helmet on is airborne and upsi...	A person with a metal helmet.	1	Three women smiling and sitting down.	3 people smile and sit down	0
99	Two men and a young lady with red hair play Wi...	No one is playing video games.	2	People standing outside of an Irish Pub.	People are standing outside a pub.	0

100 rows × 6 columns

Figure 7: Some Example of Contrast set we created

4.1 Dataset Selection

We used the SNLI data set (Stanford Natural Language Inference), which is extensively used for natural language analysis: but it has certain flaws and limitations:

1. **Limited Diversity:** The data set focuses exclusively on sentence pairings reflecting simple textual entailment relationships (for example, "entailment," "contradiction," or "neutral"). It may not effectively represent the complexity and diversity of real-world natural language processing tasks.

2. **Lack of Out-of-Distribution Data:** Because the data set is primarily intended for in-distribution tasks, it may struggle to handle out-of-distribution or adversarial situations. When faced with unexpected inputs, models trained on SNLI might fail.

3. **Biases and Stereotypes:** SNLI, like many NLP data sets, may have biases and stereotypes in its training data. This can result in biased model predictions and a lack of fairness and equity in natural language processing tasks.

The dataset has 700+ missing and a null label indicating -1. So we remove the extra class -1 from the data. Here is the distribution of the SNLI dataset 8.

4.2 Creation and Evaluation of Contrast set

Contrast sets can be used to evaluate natural language inference models. They are made up of examples that are designed to test models and demonstrate their shortcomings. The contrast set is often made up of similar samples with differing labels or entailment connections. These pairs

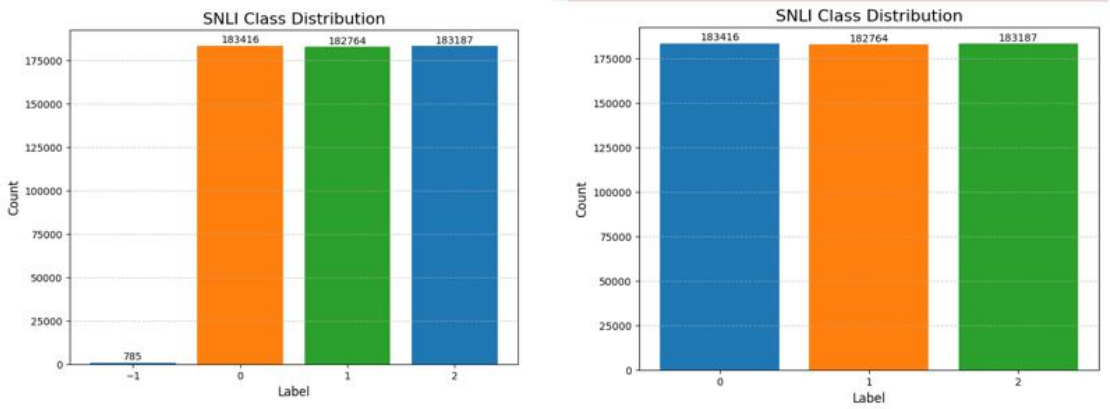

Figure 8: Distribution of class before and after removal of null labels

can be manually picked or produced automatically. Refer to Figure 7 to see the example of the created contrast set. In order to generate a contrast set, first create a data set of pairs of examples, each pair consisting of two premises and two hypotheses. These pairs are carefully chosen to ensure that premise 1 and premise 2 do not overlap in each pair. The algorithm chooses two random samples from the original data set for each pair, extracts their premises, hypotheses, and labels, and adds them to the comparison set. This produces a data set with twice the number of examples as the original dataset. Both premise1_hypothesis1 and premise2_hypothesis2 are labeled, and the text pairs are tokenized and readied for model input. By comparing predictions on different examples, this contrasting set can be used to evaluate and debias classification of text models.

5 Conclusion

Overall, ensemble-based debiasing and adversarial challenge sets both showed that we can improve the outcomes and eliminate many biases with these two methods.

Ensemble-based learning on the SNLI data set was a promising strategy for improving natural language inference model performance and robustness. We may effectively limit potential biases and increase classification accuracy by combining the predictions of artifact experts and a well-trained core model. This approach takes advantage of the capabilities of various models, making it a valuable solution for difficult NLP tasks.

For adversarial challenges, we showed that using a small data set of curated examples that challenge the model's existing biases and then having the model learn on those can help it eliminate biases and perform better in the wild.

References

a. Hugging face cross-encoder (nli-deberta-base). Accessed on November 3, 2023.

b. Hugging face cross-encoder (nli-deberta-base). Accessed on November 3, 2023.

Nelson F. Liu, Roy Schwartz, and Noah A. Smith. 2019. Inoculation by fine-tuning: A method for analyzing challenge datasets.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference.