# Investigating the impact of Training Data Selection on Calibration and Selective Prediction

## Problem Statement

Given is a training dataset and its Data Map which tells us the distribution of training instances based on the model's predictions (i.e. which instances are easy and which ones are hard). Study the impact of Training data selection (from various regions of the Data Map) on model calibration and Selective Prediction performance in both In-Distribution Generalization and Out-Of-Distribution Generalization settings.

| Datasets Used | |
|---|---|
| SNLI | https://nlp.stanford.edu/projects/snli/ |
| SWAG | https://arxiv.org/abs/1808.05326 |
| Commonsense QA | https://www.tau-nlp.org/commonsenseqa |
| Abductive NLI | https://arxiv.org/abs/1908.05739 |
| Social IQA | https://leaderboard.allenai.org/socialiqa/submissions/get-started |

## Introduction

Models designed for tasks ranging from Question Answering to Predicting Sentiments tend to perform better when trained and tested on the same distribution. However, in a practical setup, a Question Answering model may encounter questions that diverge from the model's training dataset. As an example, a sentiment analyzer that was trained on movie reviews might run into a restaurant review in a test set. In such a situation, the model may not perform to the same level as would have been expected were the model given a movie review. Although an ideal system would come up with the right answers to all such corner cases, such perfection is not attainable given limited training data. Our objective is to improve the model's generalization. In this process, we will examine the impact of training data selection from various regions of the Data Map on the model's Selective Prediction performance and accuracy.

### Dataset Descriptions

CommonsenseQA - A dataset for commonsense question answering. To capture common sense beyond associations, we extract from ConceptNet (Speer et al., 2017) multiple target concepts that have the same semantic relation to a single source concept. Crowd-workers are asked to author multiple-choice questions that mention the source concept and discriminate in turn between each of the target concepts.

SNLI - The Stanford Natural Language Inference (SNLI) corpus (version 1.0) is a collection of 570k human-written English sentence pairs manually labeled for balanced classification with the labels *entailment*, *contradiction*, and *neutral*. We aim for it to serve both as a benchmark for evaluating representational systems for text, especially including those induced by representation-learning methods, as well as a resource for developing NLP models of any kind.

SWAG - A dataset with 113k multiple choice questions about a rich spectrum of grounded situations. To address the recurring challenges of the annotation artifacts and human biases found in many existing datasets, we propose Adversarial Filtering (AF), a novel procedure that constructs a de-biased dataset by iteratively training an ensemble of stylistic classifiers, and using them to filter the data.

Abductive NLI - The dataset ART consists of over 20k commonsense narrative contexts and 200k explanations. Based on this dataset, we have a task : Abductive NLI, a multiple-choice question answering task for choosing the more likely explanation.

Social IQa - Social Interaction QA, a new question-answering benchmark for testing social common sense intelligence. Contrary to many prior benchmarks that focus on physical or taxonomic knowledge, Social IQa focuses on reasoning about people's actions and their social implications.

# Methodology

---

For our project, we chose to work with the bert-base-uncased model for each of the datasets. Additionally, we used the following parameters for each dataset both when generating the datamaps and when calculating the accuracy for the validation sets against the various subsets of the data.: 5 epochs and a learning rate of 5e-5.

To begin, we trained and validated each of the datasets on a subset of each dataset's training data. Each subset consisted of 5k instances from the original training dataset. With this subset of data in hand, we ran the model pertaining to the dataset using the 5k instances as both the training and validation datasets. We were able to generate 5 epoch's worth of accuracy measures which we then used to calculate the variance and average confidence of each sample across the 5 epochs.

From this data, we calculated the data maps shown below. As can be seen, the data maps follow the bell curve distribution noted by the authors of Dataset Cartography. The vast majority of the data points in our sets belong to the high confidence region with some variation in the variance.

Due to this imbalance, we opted to segment our data into 4 subsets: easy, ambiguous, hard, and mixed(combination of the aforementioned segments). The regions were generated by sorting the train dataset generated for the data map by confidence. Therefore, easy would consist of the highest confidence instances, hard would consist of the lowest confidence regions, and finally mixed would be a combination of easy, hard, and ambiguous. For each segmentation of the original training dataset, half the data or 2.5k instances were used. The breakdown of each region is as follows:

- Baseline consists of 5k instances from the training dataset using the full validation set as the dev file for the model. This is the set against which all segments are measured against.
- Easy consists of the first 2.5k instances with the highest confidence score from the original dataset of 5k instances.
- Ambiguous consists of the middle 2.5k instances from the original dataset when sorted by confidence score.
- Hard consists of the last 2.5k instances from the original dataset when sorted by confidence score. I.e. the least confident instances.
- Mixed consists of 2.5k instances in which 80% or 2k instances belong to the ambiguous subset and 10% comes from the easiest 250 instances based on confidence score and another 10% (250 instances) comes from the hard subset. The total size is 2.5k for this subset.

With the segments created, we trained each segment against the full validation set. The final results of each segment were then fed into a python script for calculating the accuracy and selective prediction performance based on the results from the validation dataset. Those results are discussed further in the following section.

# Results

The premise of Dataset Cartography is that using segmentation can improve the accuracy of a model while at the same time using less data. In the paper's findings, the performance of a model increased when training on the ambiguous segmentation of the original dataset. In the paper, the authors focused on a couple of different datasets and worked to fine-tune each model. Additionally, they used Roberta rather than Bert.

Our findings are more mixed than the authors of Dataset Cartography. However, it is important to note that through the course of this project all training instances were limited to 5k. Additionally, the dev datasets for all but the SWAG and SNLI datasets consisted of less than 2k instances. In

the case of the SWAG dataset, the validation (dev) dataset consisted of over 20k instances. This may explain why the accuracy for both the SWAG and SNLI datasets were comparatively so much higher than the other datasets; There was just much more data to train on. Furthermore, the inclusion of multiple choice bert models may have further eroded the accuracy with such a small amount of training data.

Regardless, the results do show some interesting trends that seem to favor the initial premise of Swayamdipta et al. The performance of the mixed segmentation for all but the SNLI dataset proved to be higher than the baseline, in some cases noticeably so. What makes this result more interesting is that the ambiguous segmentation, of which mixed is largely formed from, does not perform nearly as well nor as consistently as the mixed. The major difference between these two segmentations is that the mixed set includes 10% extremes from either end of the spectrum: easy and hard. This seems to indicate that there is an advantage to including some extremes along with the ambiguous confidence instances.

Just as in Dataset Cartography, the hard segmentation did not perform particularly well. However, what surprised us was that the ambiguous segmentation did not perform as well as expected. Once again, this may be attributed to the low amount of data for training, but this was one unexpected difference between our experiments and findings reported by Swayamdipta et al.

In regard to the selective prediction performance, we found the overall performance to be inconsistent and does not align with the performance shown for the mixed segmentation. This may be due to the low amount of training data. However, one interesting thing we noted is the wide differences in selective prediction between segmentations in the same dataset. At this time we do not have a hypothesis, but it is worth exploring. It may be that these swings would vanish with more data.

| **Accuracy** | | | | | |
|---|---|---|---|---|---|
| **Dataset** | **Baseline** | **Easy** | **Ambiguous** | **Hard** | **Mixed** |
| SNLI | 78.84 | 74.02 | 75.86 | 64.04 | 26.99 |
| SWAG | 67.42 | 69.36 | 68.99 | 64.38 | 69.56 |
| Common Sense QA | 38.9 | 54.35 | 34.07 | 39.89 | 49.39 |
| Abductive NLI | 55.03 | 54.18 | 54.37 | 52.22 | 55.09 |
| Social IQA | 47.95 | 47.70 | 50.15 | 42.68 | 49.39 |

| **Selective Prediction** | | | | | |
|---|---|---|---|---|---|
| **Dataset** | **Baseline** | **Easy** | **Ambiguous** | **Hard** | **Mixed** |

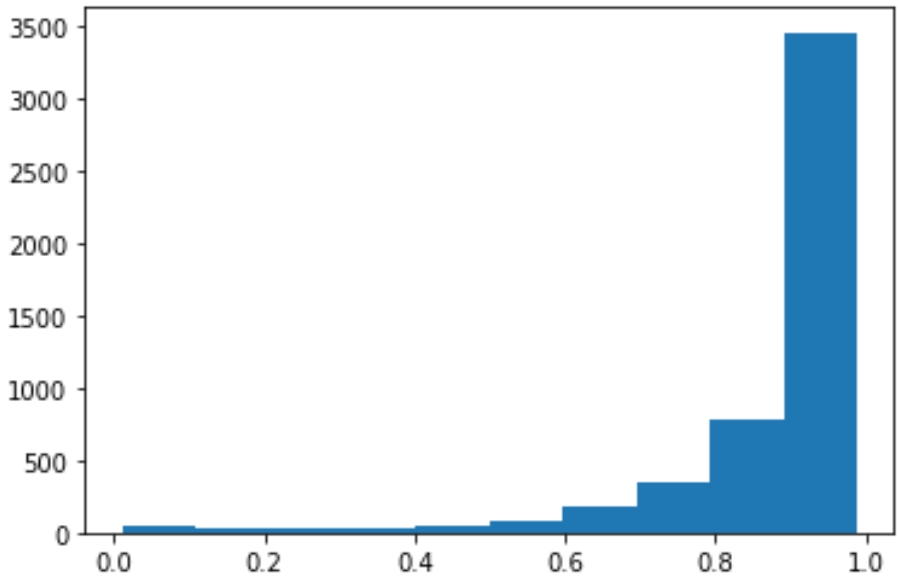
| | | | | | |
|---|---|---|---|---|---|
| SNLI | 11.19 | 14.69 | 14.49 | 26.99 | 15.13 |
| SWAG | 17.50 | 17.09 | 16.43 | 21.21 | 15.94 |
| Common Sense QA | 51.78 | 34.71 | 61.30 | 53.45 | 39.37 |
| Abductive NLI | 40.16 | 42.67 | 42.13 | 46.40 | 42.22 |
| Social IQA | 45.72 | 46.22 | 44.81 | 53.61 | 45.15 |

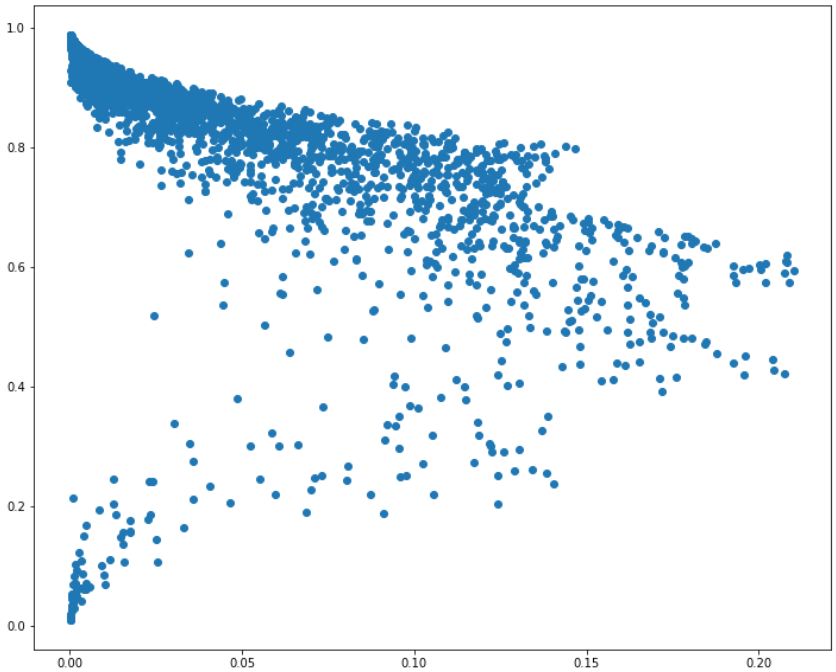
# Data Maps

---

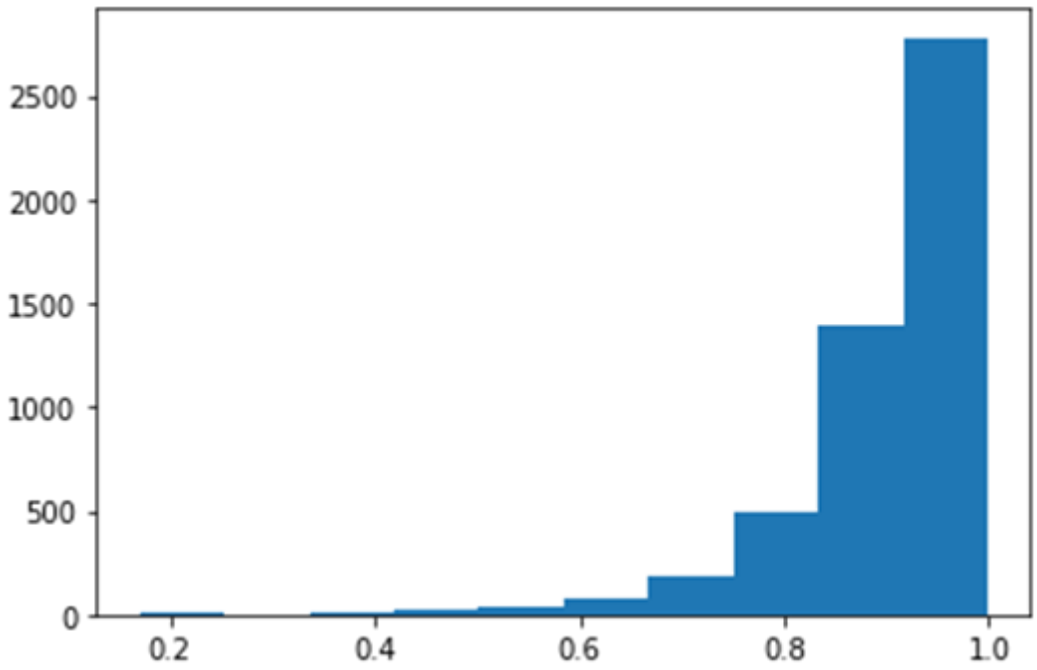
This section shows the histogram plots of confidence estimates for each dataset and their corresponding Data Maps. Data Maps are a scatter plot between variance and confidence. The regions higher up in the Y-axis indicate the “easy-to-learn” segment since it has a higher confidence level. The instances that fall in the middle are “ambiguous” while the ones that lie at the bottom are “hard-to-learn” instances since their confidence estimates are quite poor.

## SNLI Dataset

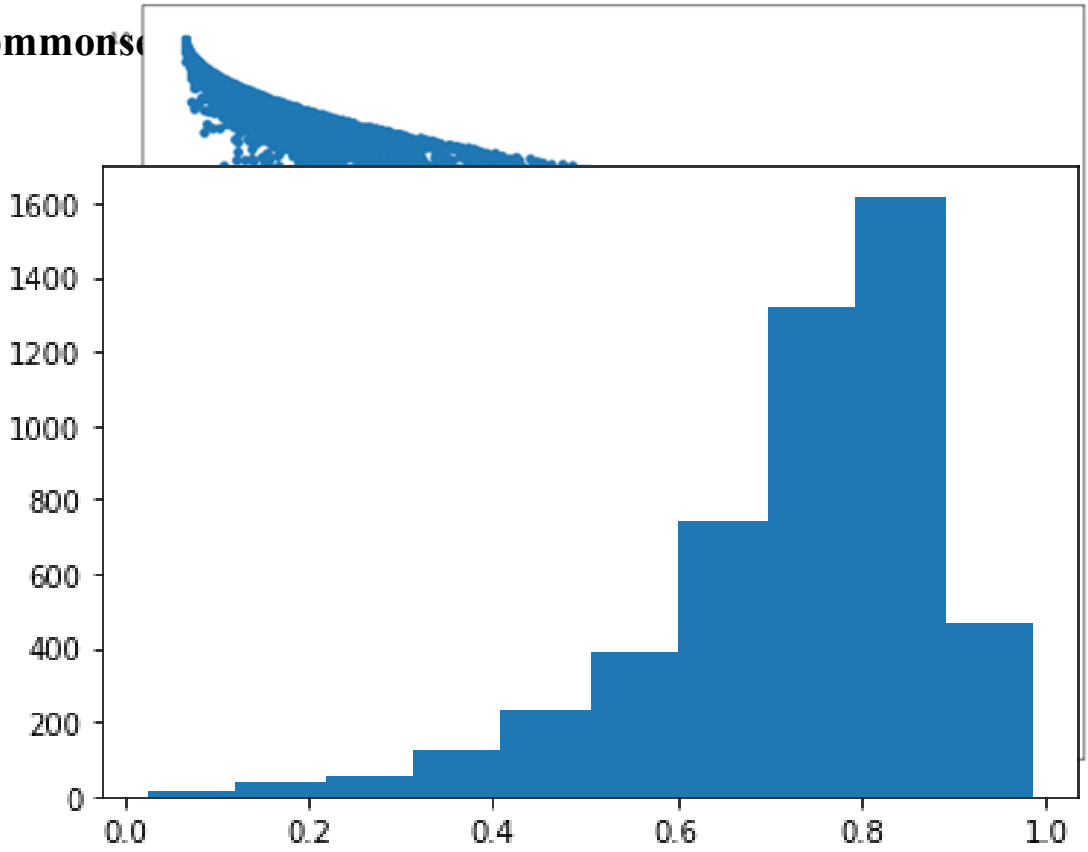




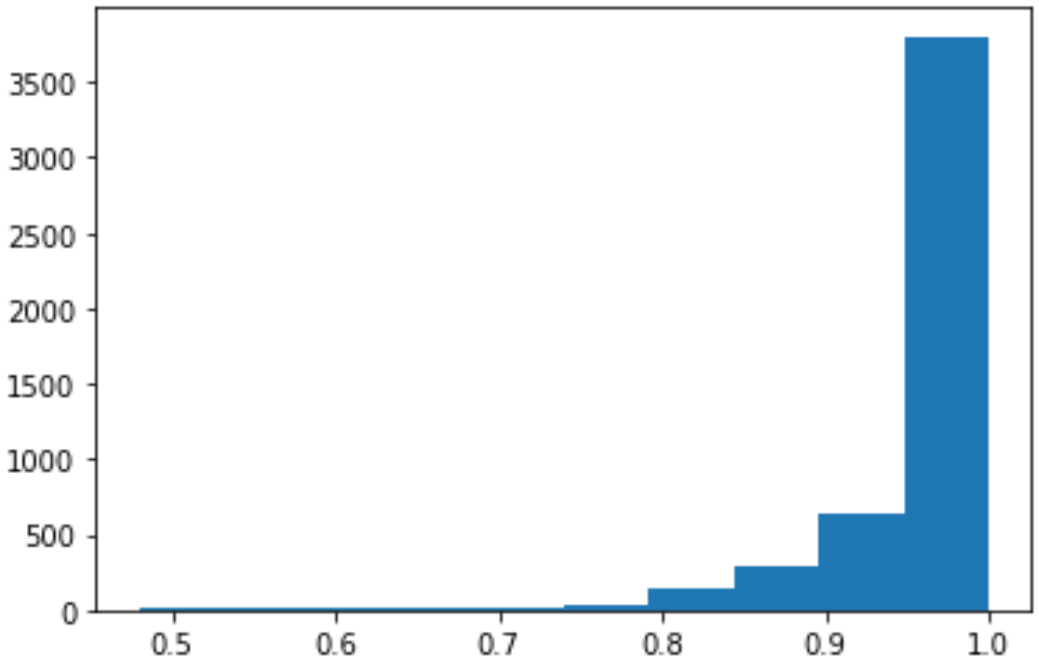

**SWAG Dataset**

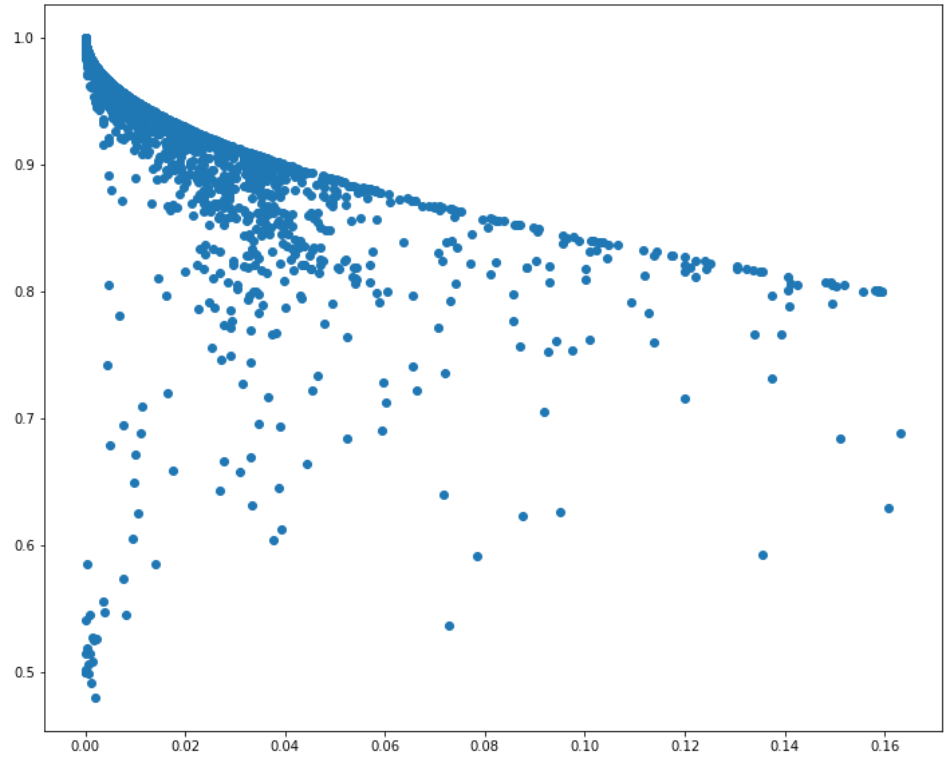


**Commonse**

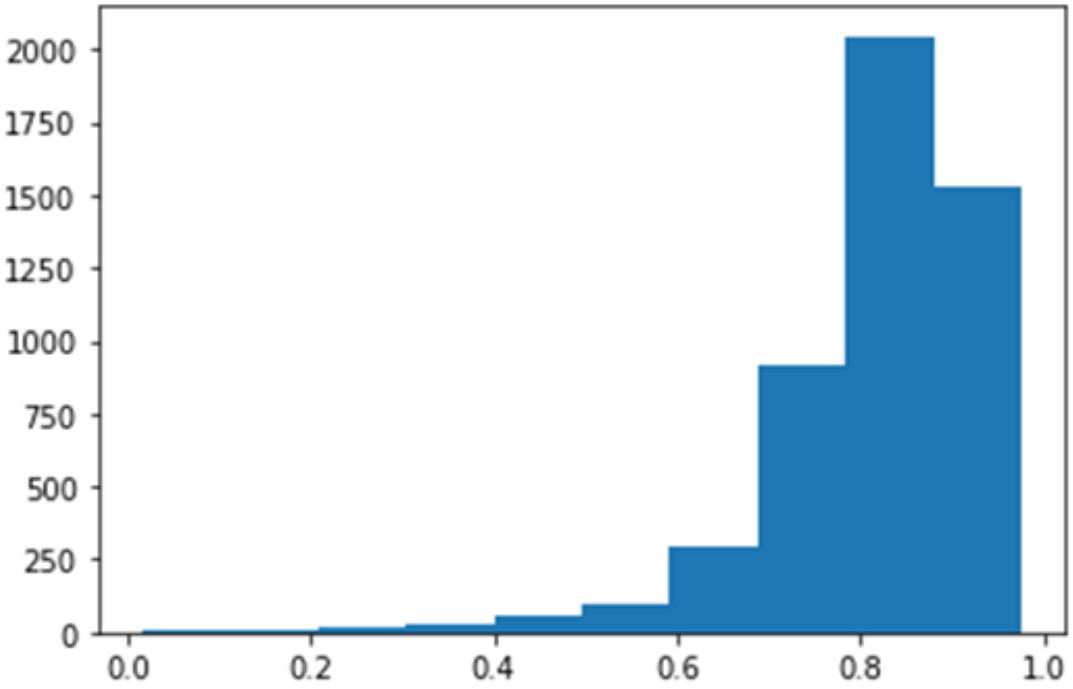


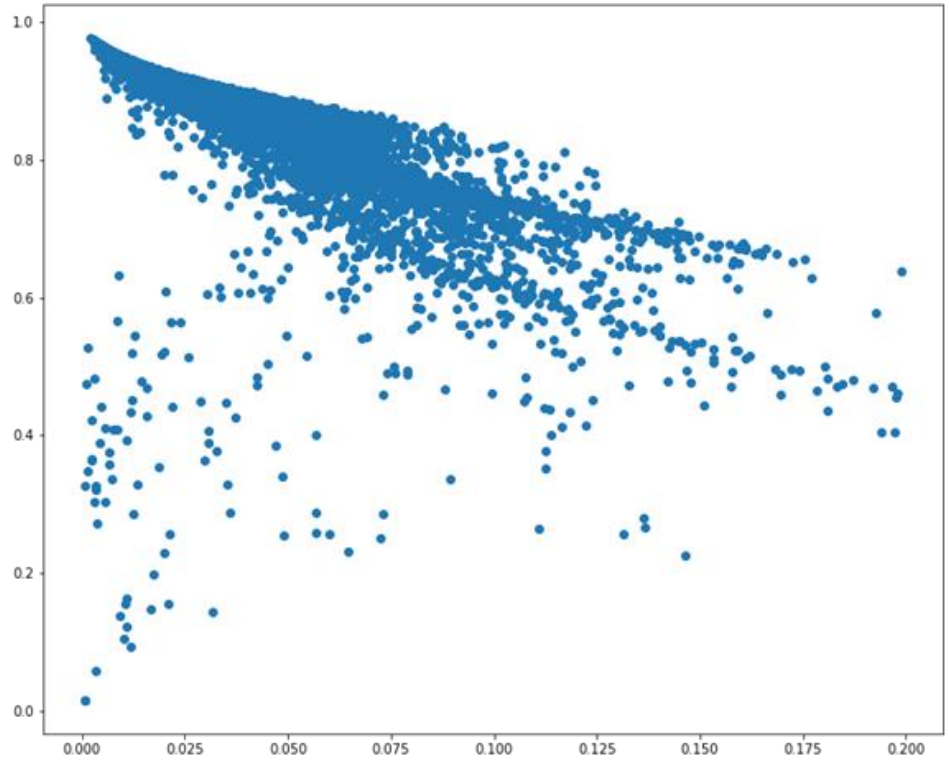

## Abductive NLI Dataset




## Social IQA Dataset


2000
1750
1500
1250
1000
750
500
250
0
0.0
0.2
0.4
0.6
0.8
1.0


1.0
0.8
0.6
0.4
0.2
0.0
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200

# References


- Swayamdipta, S., Schwartz, R., Lourie, N., Wang, Y., Hajishirzi, H., Smith, N. A., & Choi, Y. (2020). Dataset cartography: Mapping and diagnosing datasets with training dynamics. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. https://doi.org/10.18653/v1/2020.emnlp-main.746
- Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.*
- Neeraj Varshney, Swaroop Mishra, and Chitta Baral. 2020. It's better to say" i can't answer" than answering incorrectly: Towards safety critical nlp systems. arXiv preprint arXiv:2008.09371.
- Ji Xin, Jimmy ., Raphael Tang, Yaoliang Yu (Aug 02 2021). The Art of Abstention: Selective Prediction and Error Regularization for Natural Language Processing, *The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Underline Science Inc. Available from: https://underline.io/25966-the-art-of-abstention-selective-prediction-and-error-regularization-for-natural-language-processing*
- Siddhant Garg and Alessandro Moschitti. 2021. Will this Question be Answered? Question Filtering via Answer Model Distillation for Efficient Question Answering. *In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 7329–7346.*