Carnegie Mellon University
School of Computer Science

Authors: Atharva Kulkarni, Raghav Kapoor, Anubha Kabra

Language Technologies Institute

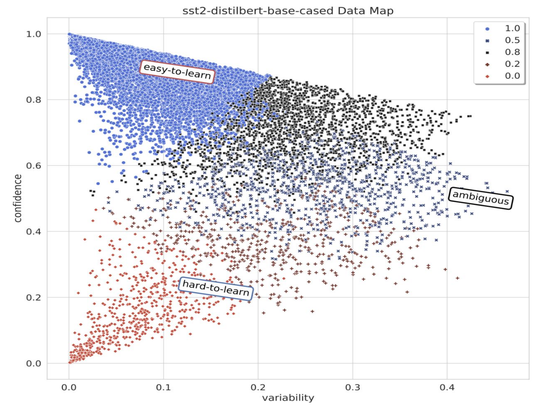
Towards Detecting, Augmenting, and Sampling, Efficient Data points for Robust Text Classification

Problem Statement

Machine learning models mostly utilize all the training data without assessing its quality. Many times, the datasets contain mislabelled points along with ambiguous ones. Thus, we propose a quality assessment technique along with intelligent augmentation for robust text classification. Our approach is two folds:

1. Identify samples that are easy to learn, ambiguous, and hard to learn using the model's training dynamics.
2. Mixup-based data augmentation technique that generates synthetic samples by intelligently combining easy and ambiguous samples based on entropy.

Data Maps

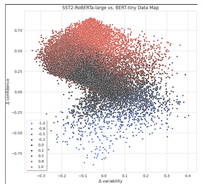


Inspired by the works of Swayamdipta et al. (2020), we first categorize our data into easy-to-learn, hard-to-learn, and ambiguous samples.
Learning with the pruned dataset containing easy-to-learn and hard-to-learn examples leads to better performance than training on the entire dataset.

Methodology/Workflow

Original sample in SST2 dataset i was perplexed to watch it unfold with an aston-ishing lack of passion or uniqueness .	-ve
Perturbed sample in $\{(T_{\text{DelStop}}, 2), (T_{\text{SubSyn}}, 2)\}$ i was perplexed ~~to~~ watch it unfold with an astound-ing absence of passion or uniqueness .	+ve
Original sample in SST2 dataset this is pretty dicey material .	-ve
Perturbed sample in $\{(T_{\text{Dup}}, 2), (T_{\text{SubSyn}}, 2)\}$ this becomes pretty pretty dicey material . .	+ve

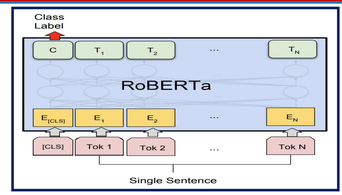
Baseline Dataset



Categorize into easy/hard/ambiguous

1. Easy + Easy
2. Easy + Ambiguous
3. Ambiguous + Ambiguous

Apply Intelligent Mixup Technique



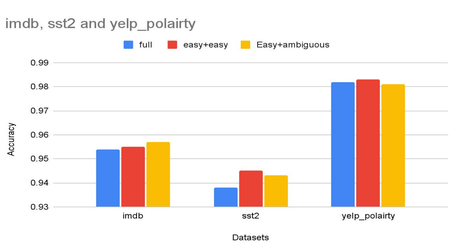
Run Experiments

Original sample in SST2 dataset i was perplexed to watch it unfold with an aston-ishing lack of passion or uniqueness .	✓
Perturbed sample in $\{(T_{\text{DelStop}}, 2), (T_{\text{SubSyn}}, 2)\}$ i was perplexed ~~to~~ watch it unfold with an astound-ing absence of passion or uniqueness .	✓
Original sample in SST2 dataset this is pretty dicey material .	✓
Perturbed sample in $\{(T_{\text{Dup}}, 2), (T_{\text{SubSyn}}, 2)\}$ this becomes pretty pretty dicey material . .	✓

Robust Results

Results

Easy+Ambiguous performs best for IMDB, while Easy+Easy performs best for SST2 and Yelp polarity datasets.



- Choosing a smaller subset of points efficiently gives similar results to full dataset.
- Our intuition is that improvement will be much more in datasets which are originally noisy or having less data points, like low resources language data.
- Running the full dataset with roberta gives us the following accuracies -
  - IMDB - 95.4 %
  - SST2 - 93.8 %
  - YELP - 98.2 %

Mixup Technique

We plan to generate synthetic samples by interpolating specific combinations of examples, namely, (easy-to-learn, easy-to-learn), (easy-to-learn, ambiguous), and (ambiguous, ambiguous). We ignore the hard-to-learn examples as they often constitute mislabeled instances.
We select the samples based on their entropy, as high entropy examples are often difficult while low entropy ones are which the model can classify confidently.
For our experiments, we use the following combinations:

1. Ambiguous (high entropy) + Ambiguous (low entropy)
2. Easy (high entropy) + Easy (low entropy)
3. Easy (random) + Ambiguous (random)