CERIAS

The Center for Education and Research in Information Assurance and Security

Text Augmentation: Improving Classification Accuracy at the Expense of Calibration?

Geetanjali Bihani, Julia Taylor Rayz

Overview

Motivation
Text data augmentation (TDA) has shown to improve model generalization in neural NLP pipelines by increasing data diversity. But does it come at the cost of calibration?

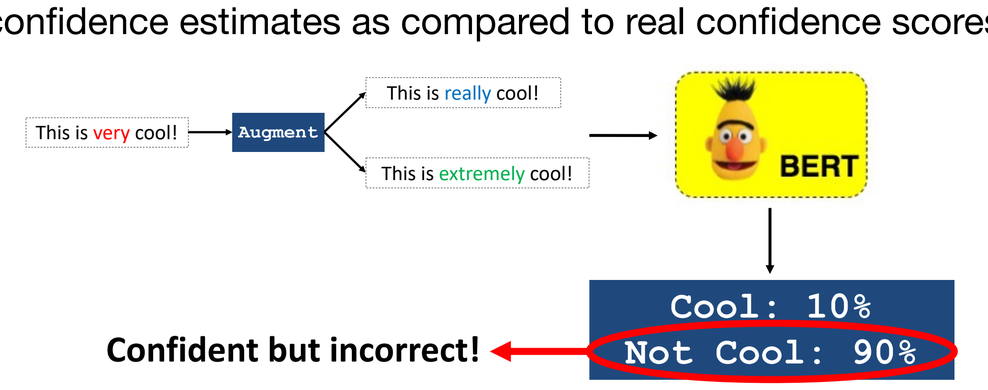
Question
Does Text Data Augmentation improve model reliability in terms of calibration, on NLP classification tasks?

Approach
Study how *overconfidence*, *under confidence,* and *Expected Calibration Error (ECE)* in model decisions varies across different levels of text augmentation for NLP classification

Text Augmentation and (Mis)-Calibration

Text Augmentation: Increases data diversity by modifying existing data or creating synthetic data

Confidence Calibration: How reliable are a model's predicted confidence estimates as compared to real confidence scores?



- When mis-calibrated models are employed in downstream NLP applications, can result in high-risk scenarios due to overconfident but incorrect predictions

Methodology

- **Model:** BERT (base) finetuned on using one epoch training to observe the regularizing effects of text augmentation in isolation
- **Text Augmentation Techniques:** Token based (wordnet based synonym replacement, deletion, swapping)
- **Calibration Evaluation Metrics**:
  - **Expected Calibration Error (ECE)**

$$\mathrm{ECE}_p = \mathbb{E}[|\hat{z} - \mathbb{E}[1_{\hat{y}=y}|\hat{z}]|^p]^{\frac{1}{p}}$$

  - **Overconfidence:** Expectation of confidence over incorrect predictions
  - **Underconfidence:** Expectation of 1-confidence over correct predictions

Results

With more augmentation comes more overconfidence

- As we add more augmented samples, the models become increasingly overconfident
- Underconfident instances reduce with an overall boost in predicted confidence
- Expected Calibration Error (ECE) also increases as more augmented samples are added, regardless of augmentation type

	Overconfidence			*Underconfidence*			ECE		
+ % DA	wordnet	delete	swap	wordnet	delete	swap	wordnet	delete	swap
+ 0%	0.479	0.477	0.489	0.237	0.237	0.226	0.023	0.020	0.018
+ 50%	0.523	0.540	0.535	0.187	0.179	0.184	0.022	0.028	0.023
+ 100%	0.549	0.566	0.566	0.161	0.150	0.155	0.042	0.049	0.044

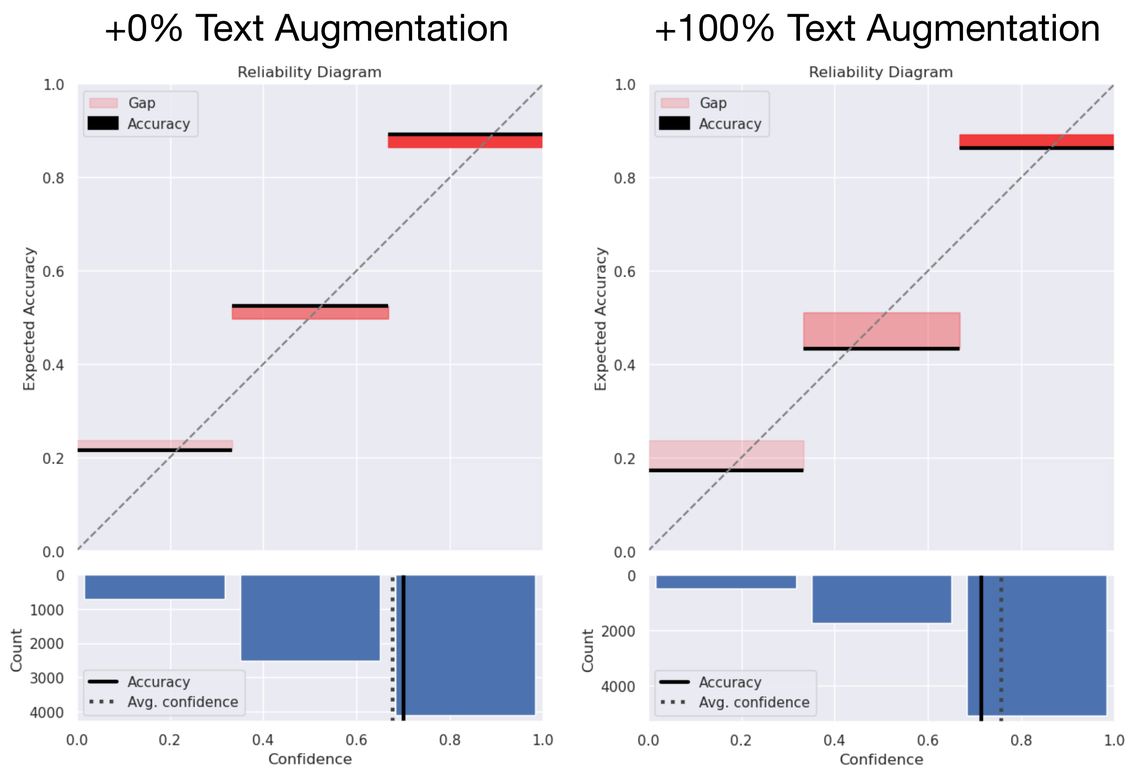


Figure: The reliability diagrams are presented on the top half (Red), while confidence histograms are in the bottom (Blue). Black lines in reliability diagrams show prediction accuracy.

- As we add more augmented samples, the gap between model confidence and accuracy of instances in each bin increases
- While being underconfident in initial predictions, the model becomes increasingly overconfident post the addition of augmented samples
- Average confidence increases by a much larger margin of difference with the application of text augmentation

Future Work

- Current calibration error metrics are skewed towards high probability bins, work ongoing on estimating calibration error using soft binning of confidence