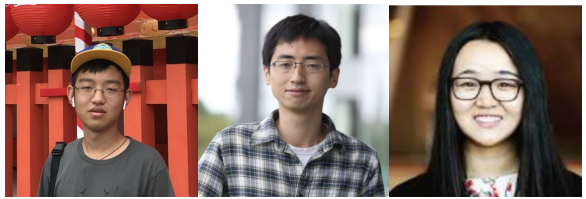
MixText: Linguistically-Informed Interpolation of Hidden Space for Semi-Supervised Text Classification

Jiaao Chen, Zichao Yang, Diyi Yang



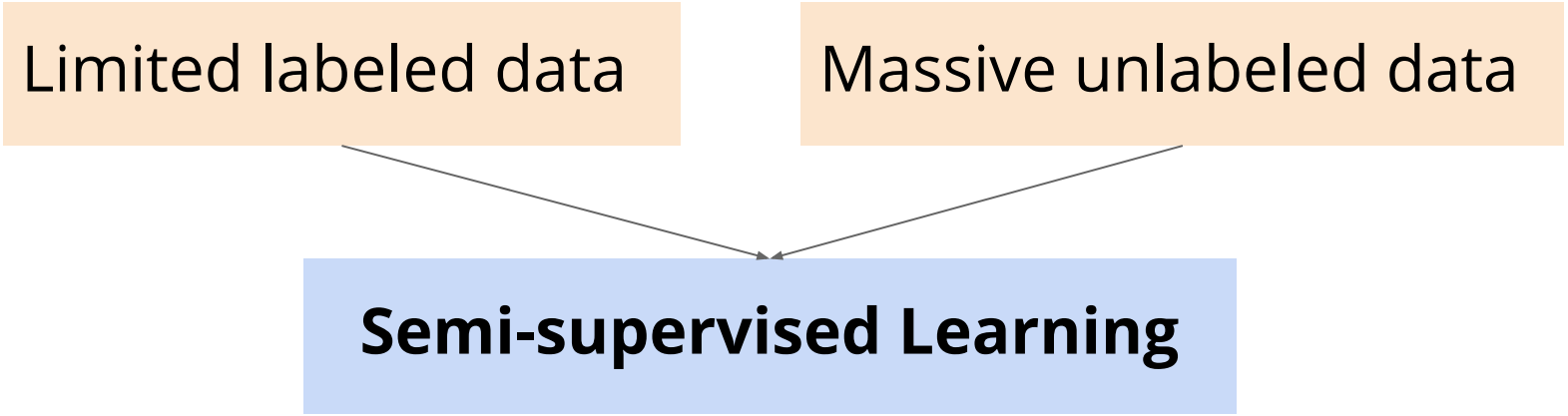
Georgia Tech College of Computing
School of Interactive Computing

In real-world applications, annotated data is often limited.

How to utilize limited labeled data for learning (e.g., for text classification)?

In real-world applications, annotated data is often limited.

How to utilize limited labeled data for learning (e.g., for text classification)?



Prior Work on Semi-Supervised Text Classification

- Variational autoencoders (Chen et al., 2018; Yang et al., 2017; Gururangan et al., 2019)
- Confident predictions on unlabeled data for self-training (Lee, 2013; Grandvalet and Bengio, 2004; Meng et al., 2018)
- Consistency training on unlabeled data (Miyato et al., 2019, 2017; Xie et al., 2019)
- Pre-training on unlabeled data, then fine-tuning on labeled data (Devlin et al., 2019)

Why Is It Not Enough?

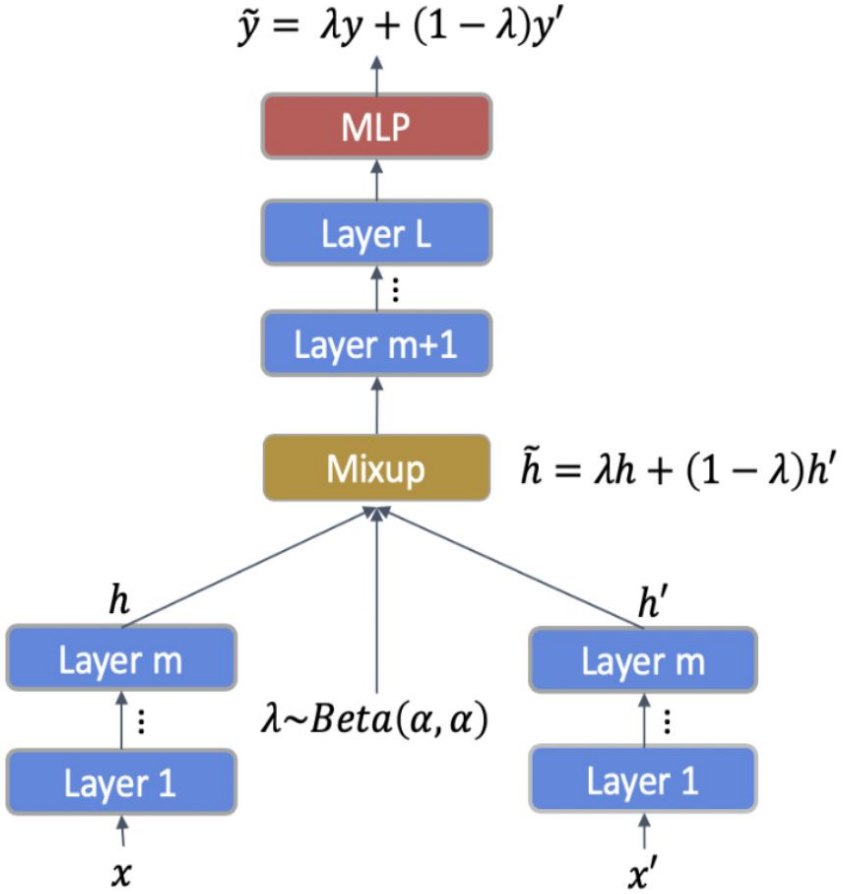
- Labeled and unlabeled data are treated **separately**
- Models may easily **overfit** on labeled data while still underfit on the unlabeled data

TextMix (TMix)

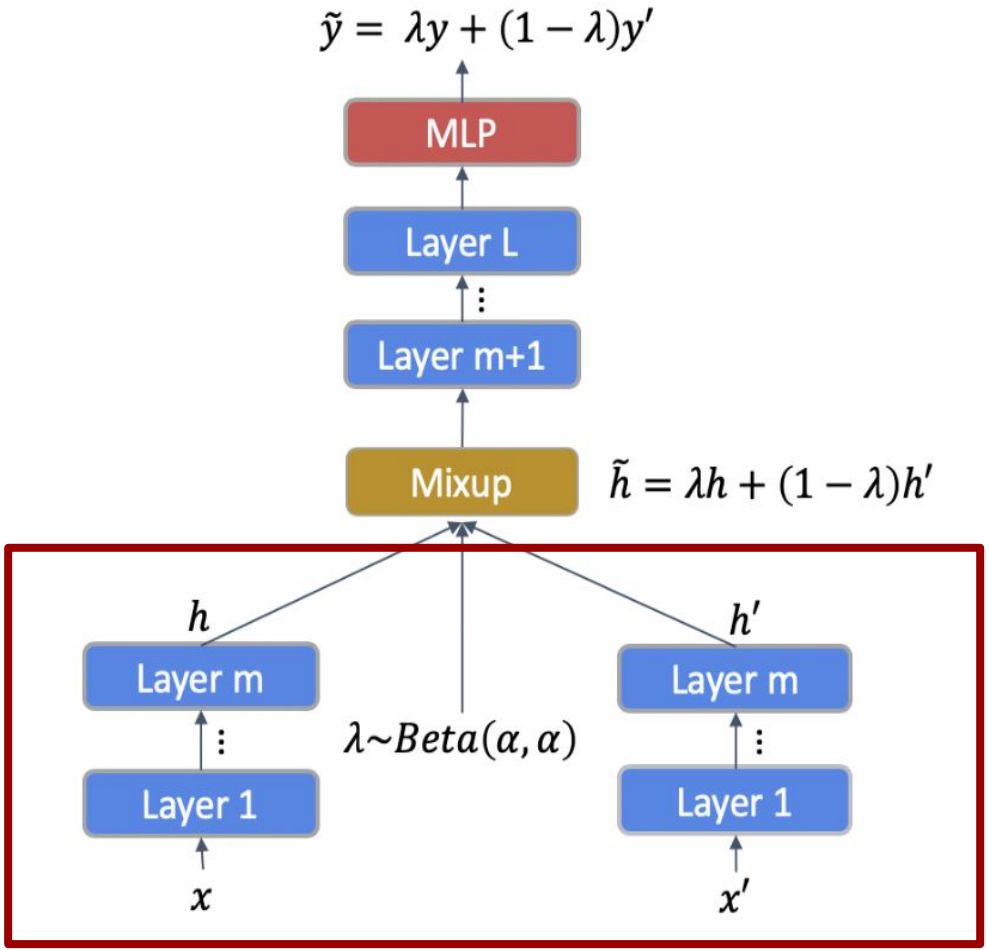
- performs linear interpolations (e.g., Mixup (Zhang et al., 2017)) in textual hidden space between different training sentences

- allows information to share across different sentences and creates infinite augmented training samples

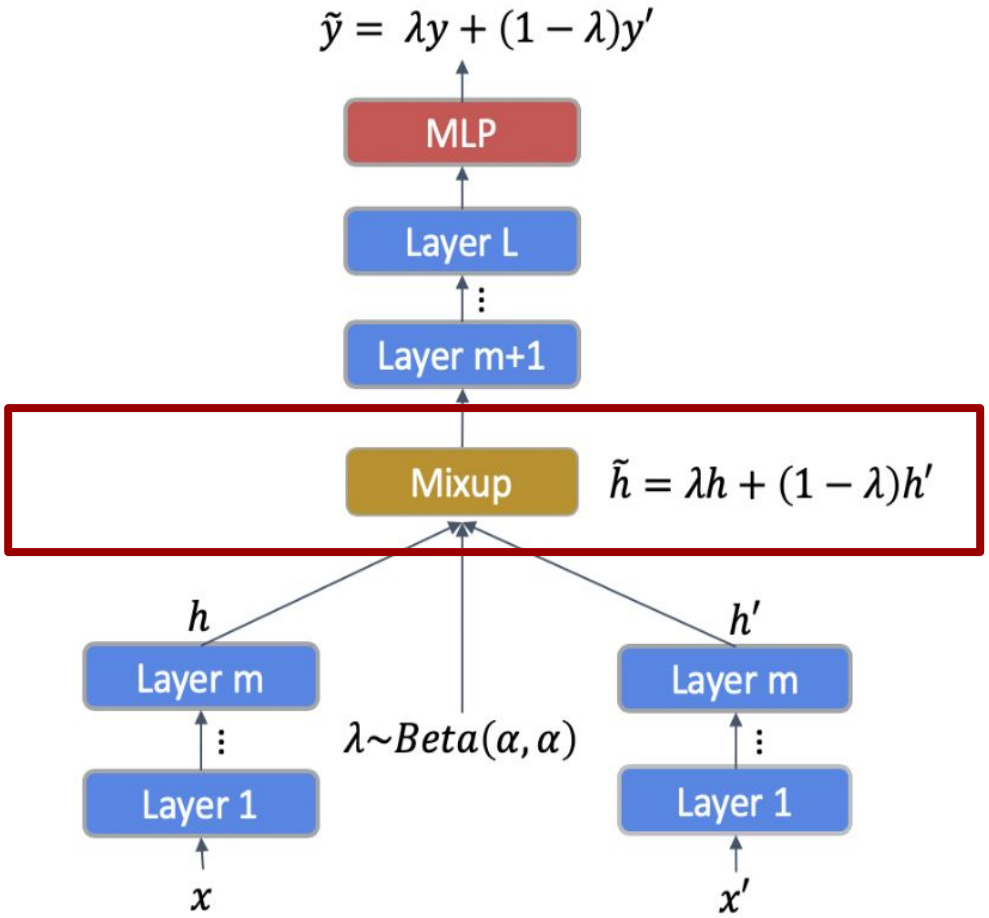
TMix



x: sentence 1
x′: sentence 2
y: label 1
y′: label 2

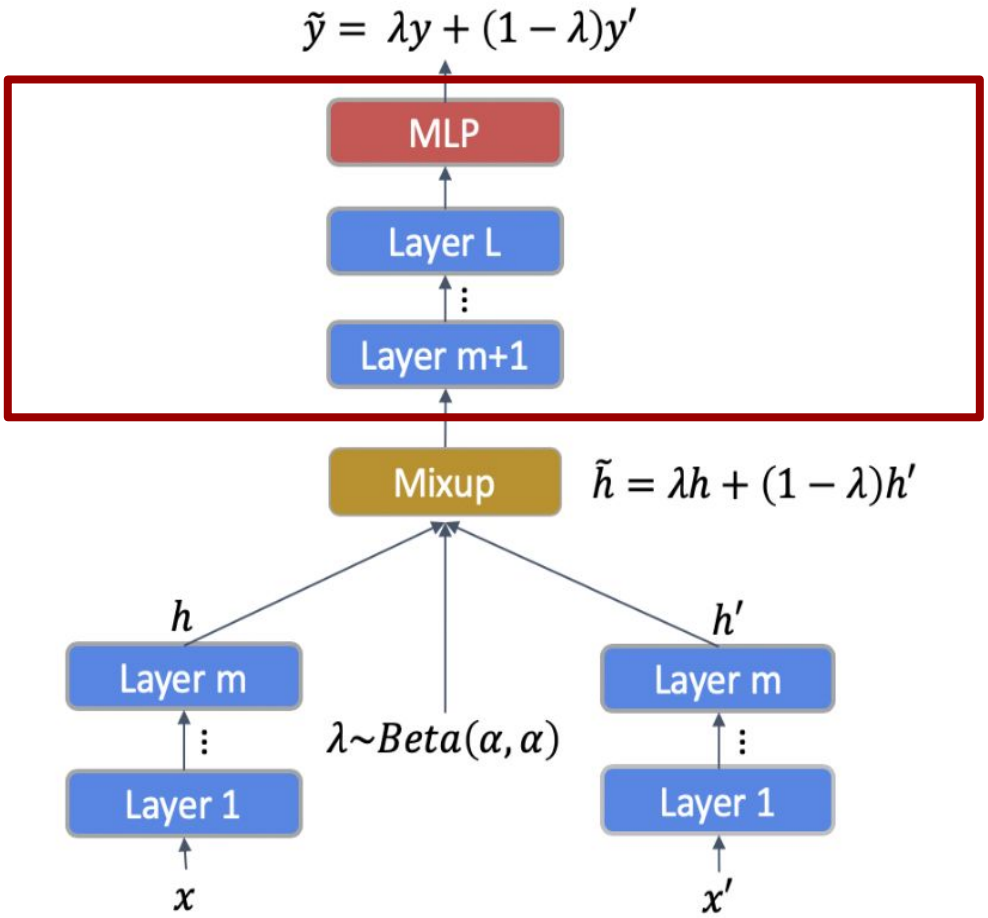


Encode separately



Linear interpolation

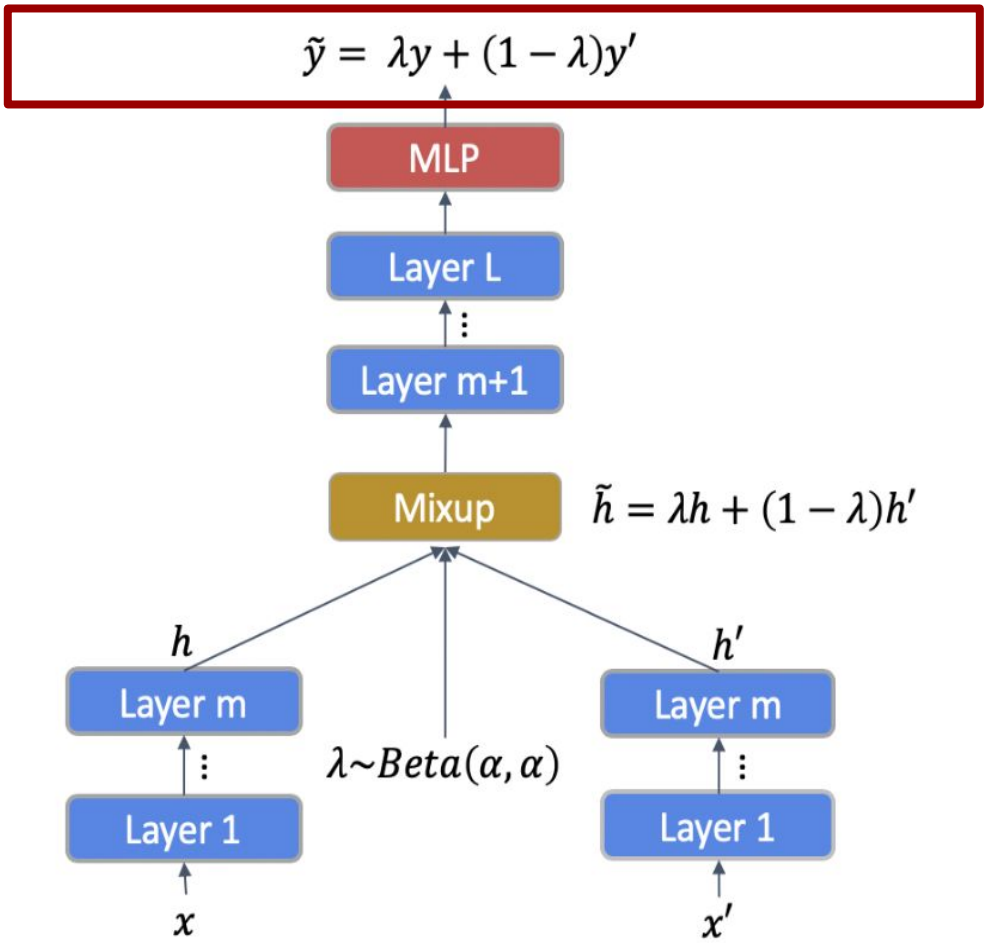
Encode separately



Forward-passing

Linear interpolation

Encode separately



Interpolate labels

Forward-passing

Linear interpolation

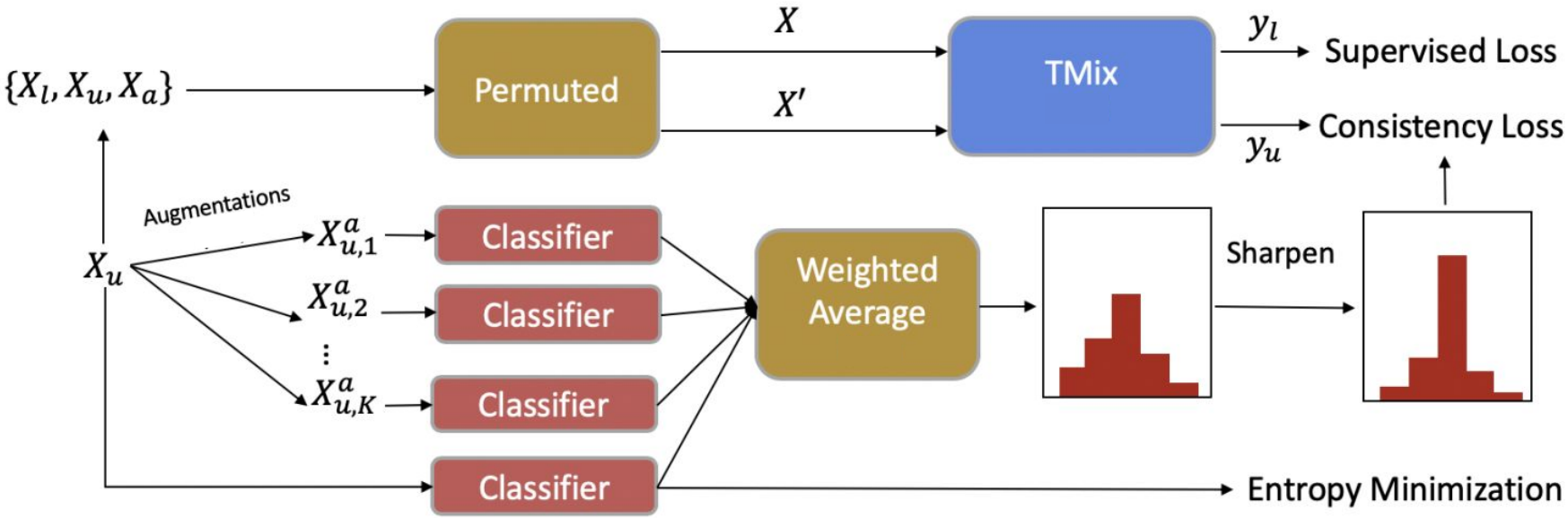
Encode separately

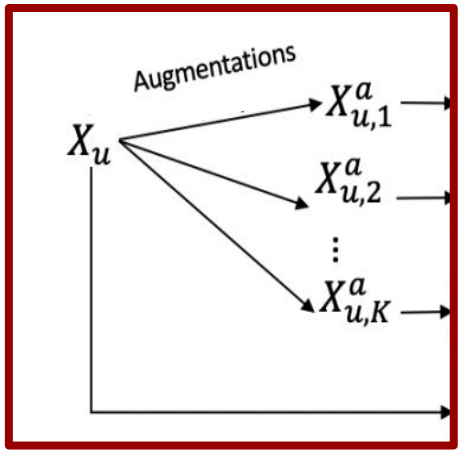
TMix: Which layers to mix?

Multi-layer encoders (e.g., BERT) capture different types of information in different layers (Jawahar et al., 2019)

- Surface, e.g., sentence length (3, 4)
- Syntactic, e.g., word order (6, 7)
- Semantic, e.g., tense, subject (7, 9, 12)

MixText = *TMix + Consistency Training* for Semi-supervised Text Classification

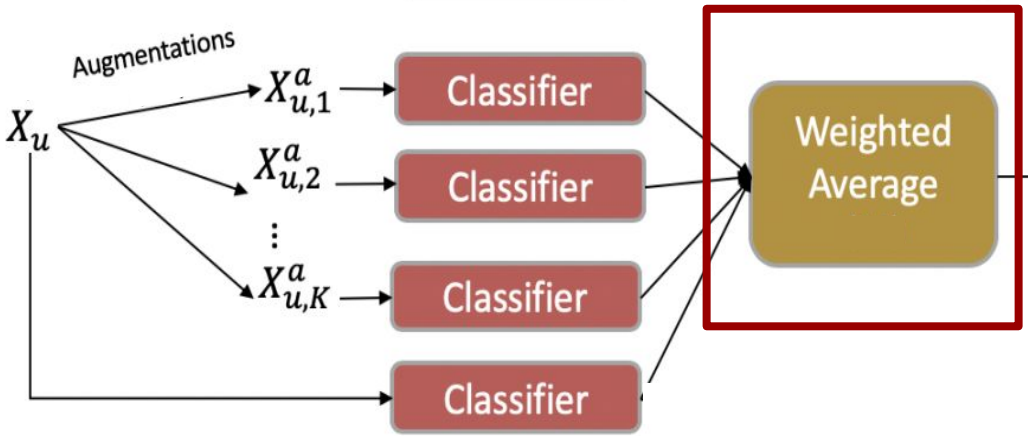




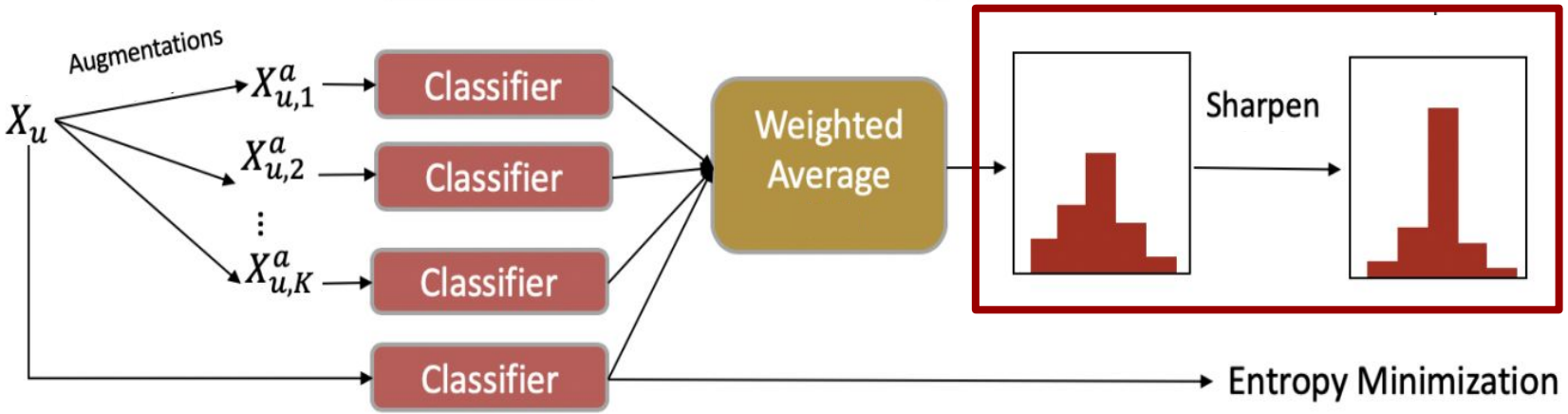
Back-translations

German & Russian as intermediate language

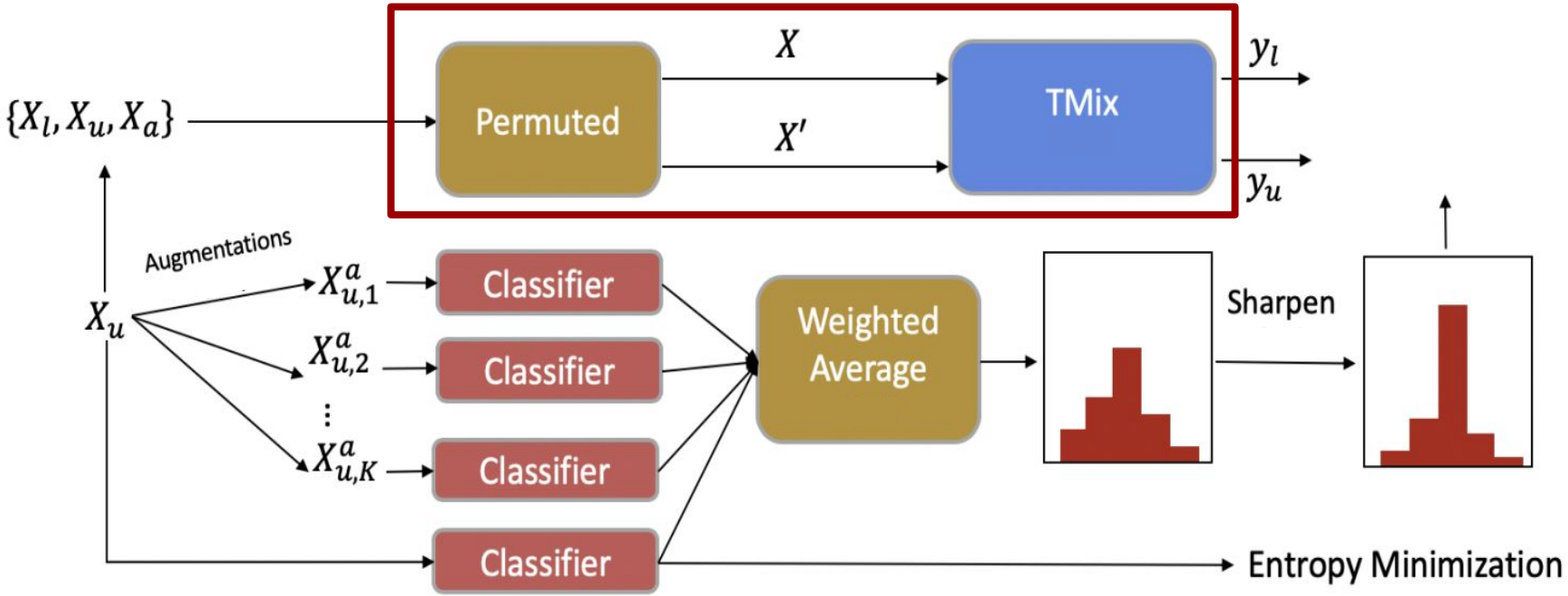
$$\mathbf{y}_i^u = \frac{1}{w_{ori} + \sum_k w_k}(w_{ori} p(\mathbf{x}_i^u) + \sum_{k=1}^{K} w_k p(\mathbf{x}_{i,k}^a)))$$



$$\text{Sharpen}(\mathbf{y}_i^u, T) = \frac{(\mathbf{y}_i^u)^{\frac{1}{T}}}{||(\mathbf{y}_i^u)^{\frac{1}{T}}||_1}$$

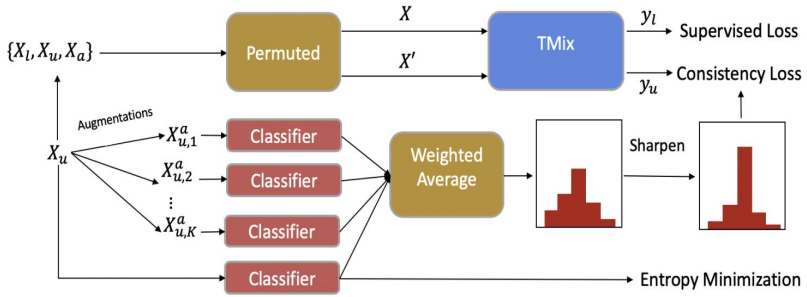


Interpolate labeled/unlabeled text



MixText

Training Objective



TMix Loss: $L_{\text{TMix}} = \mathbb{E}_{\mathbf{x},\mathbf{x}' \in \mathbf{X}} \text{KL}(\text{mix}(\mathbf{y}, \mathbf{y}') || p(\text{TMix}(\mathbf{x}, \mathbf{x}'))$

- *Supervised Loss / Consistency Loss*

Entropy Minimization: $L_{\text{margin}} = \mathbb{E}_{\mathbf{x} \in \mathbf{X}_u} \max(0, \gamma - ||\mathbf{y}^u||_2^2)$

Overall Loss: $L_{\text{MixText}} = L_{\text{TMix}} + \gamma_m L_{\text{margin}}$

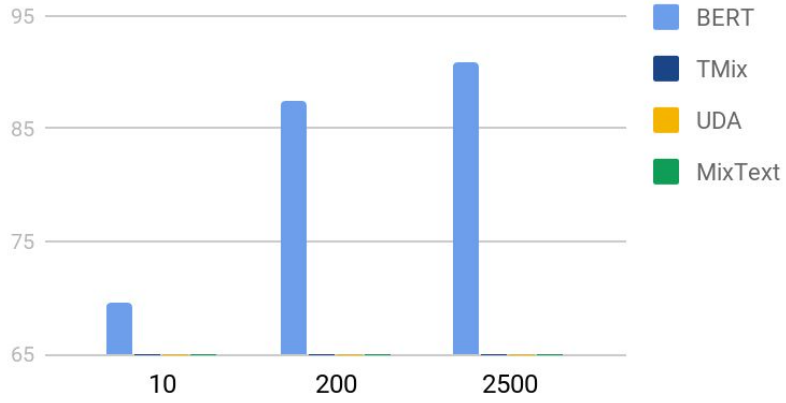
Dataset and Baselines

Dataset	Label Type	Classes	Unlabeled	Dev	Test
AG News	News Topic	4	5000	2000	1900
DBpedia	Wikipeida Topic	14	5000	2000	5000
Yahoo! Answer	QA Topic	10	5000	5000	6000
IMDB	Review Sentiment	2	5000	2000	12500

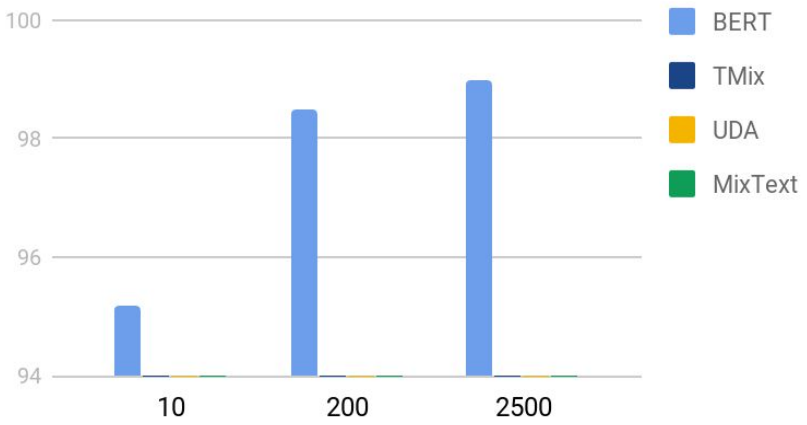
Baselines:

- BERT (Devlin et al., 2019)
- UDA (Xie et al., 2019)

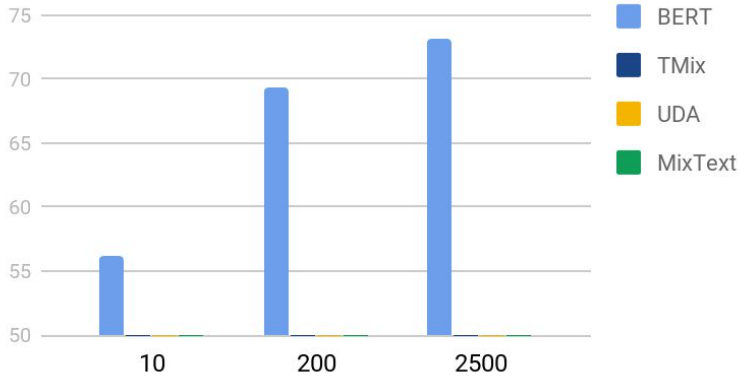
AG News



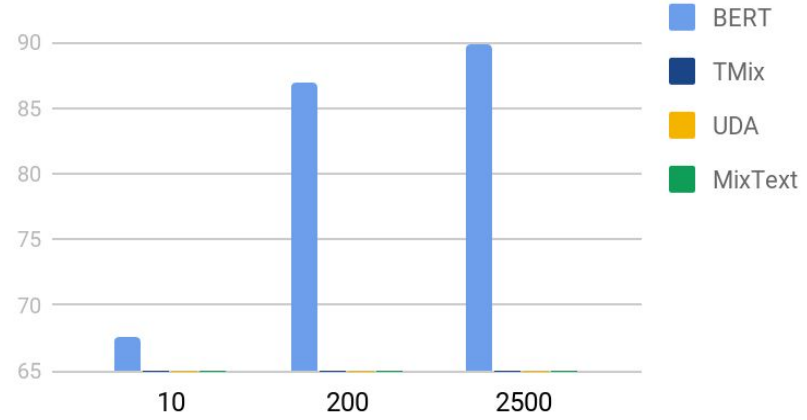
DBPedia



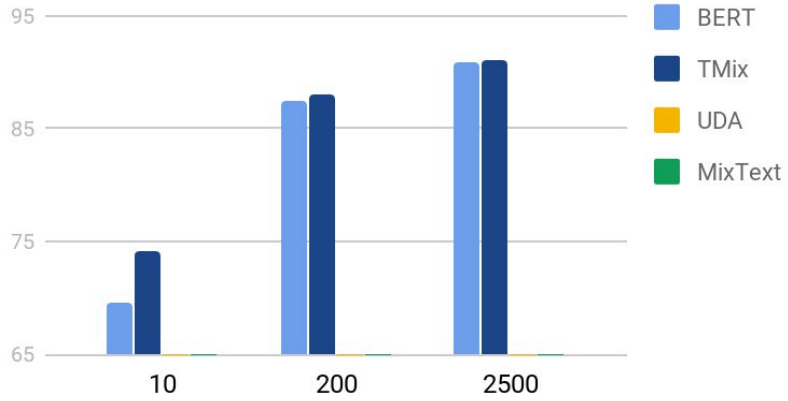
Yahoo! Answer



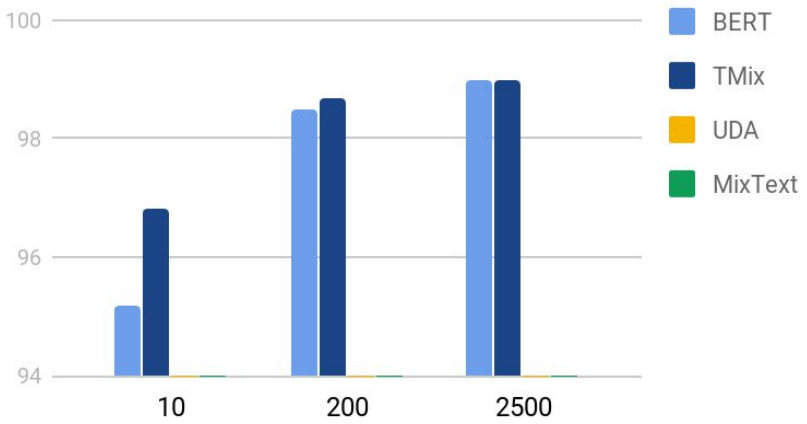
IMDB



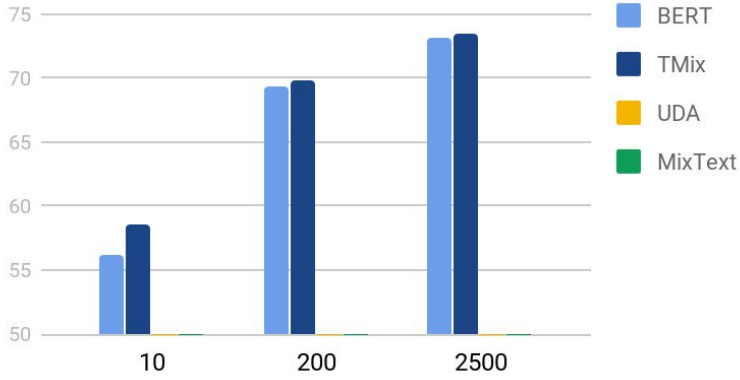
AG News



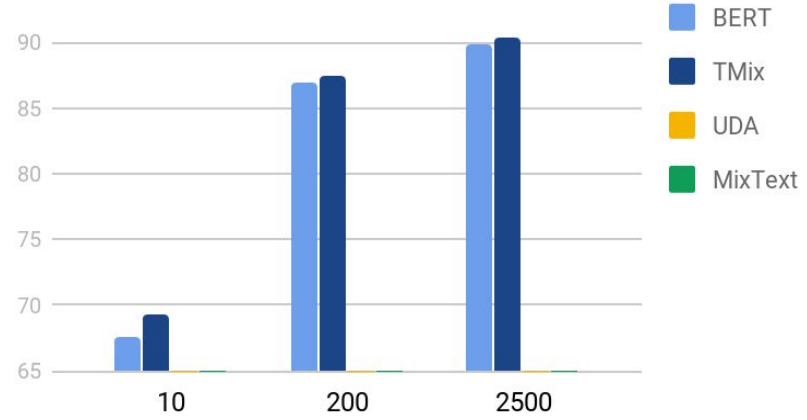
DBPedia



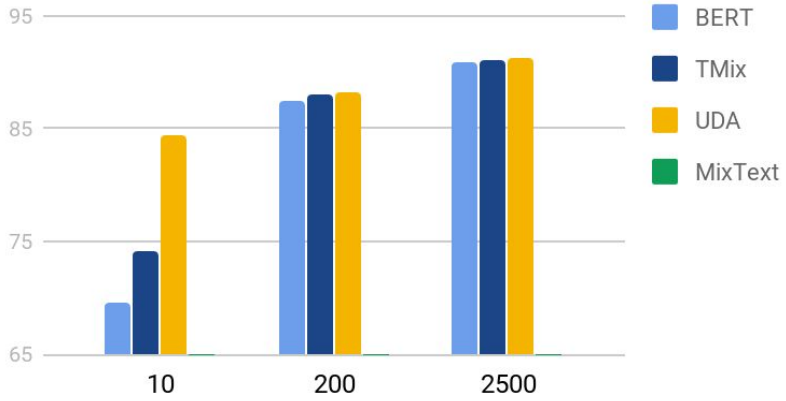
Yahoo! Answer



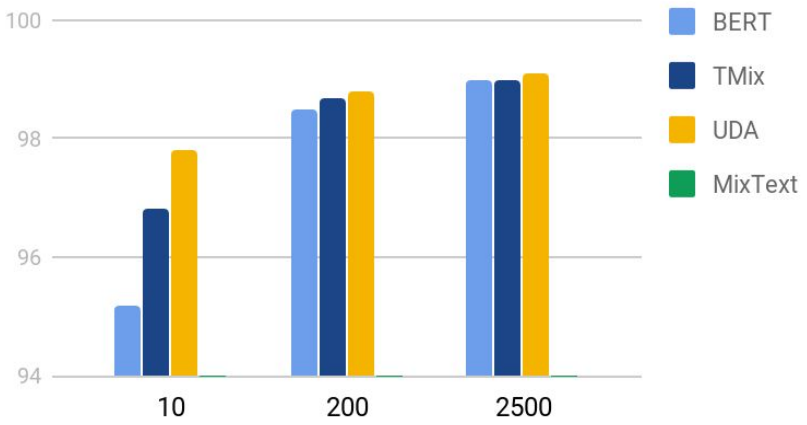
IMDB



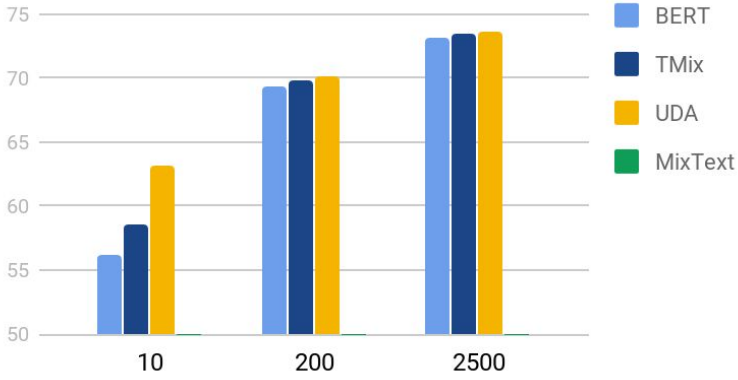
AG News



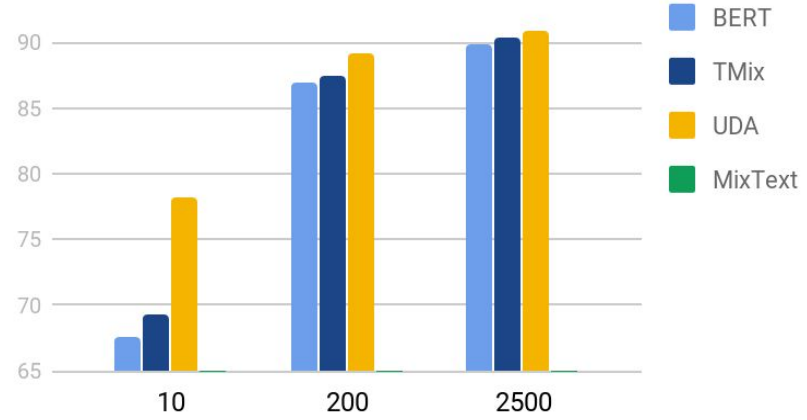
DBPedia



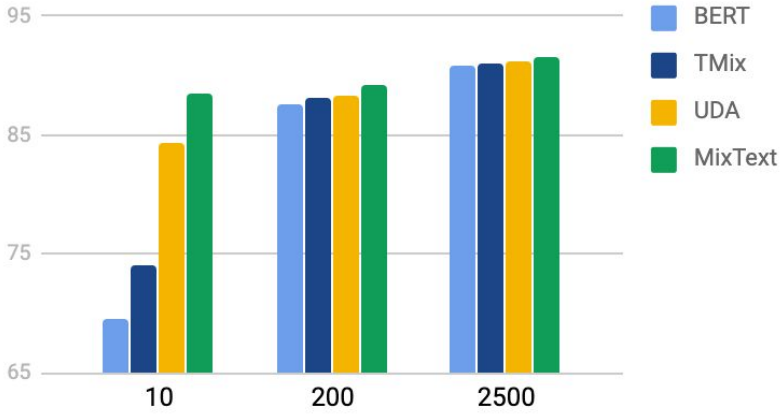
Yahoo! Answer



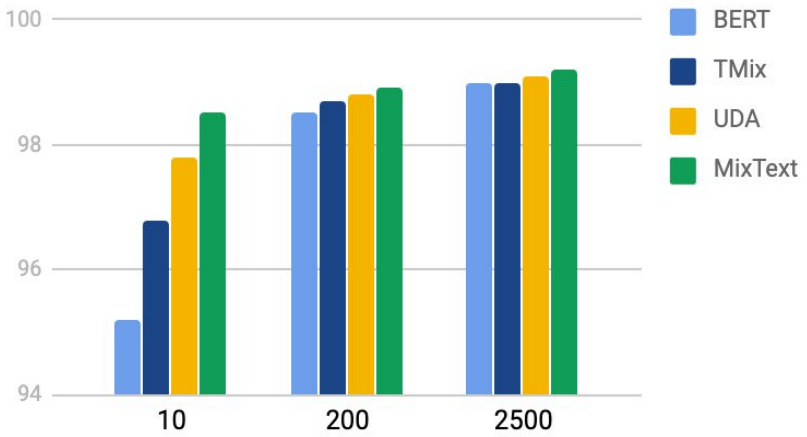
IMDB



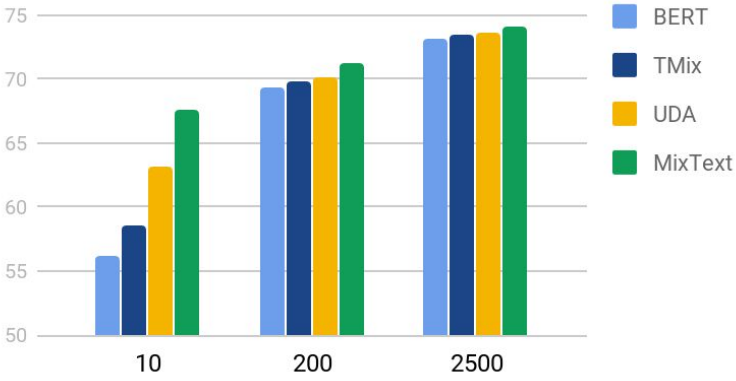
AG News



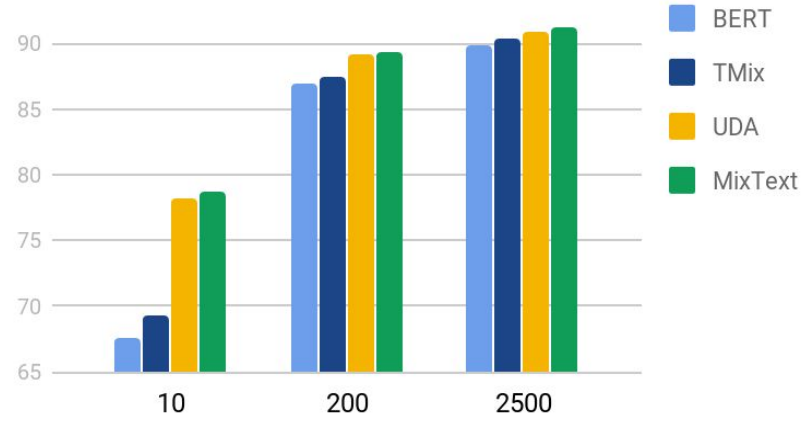
DBPedia



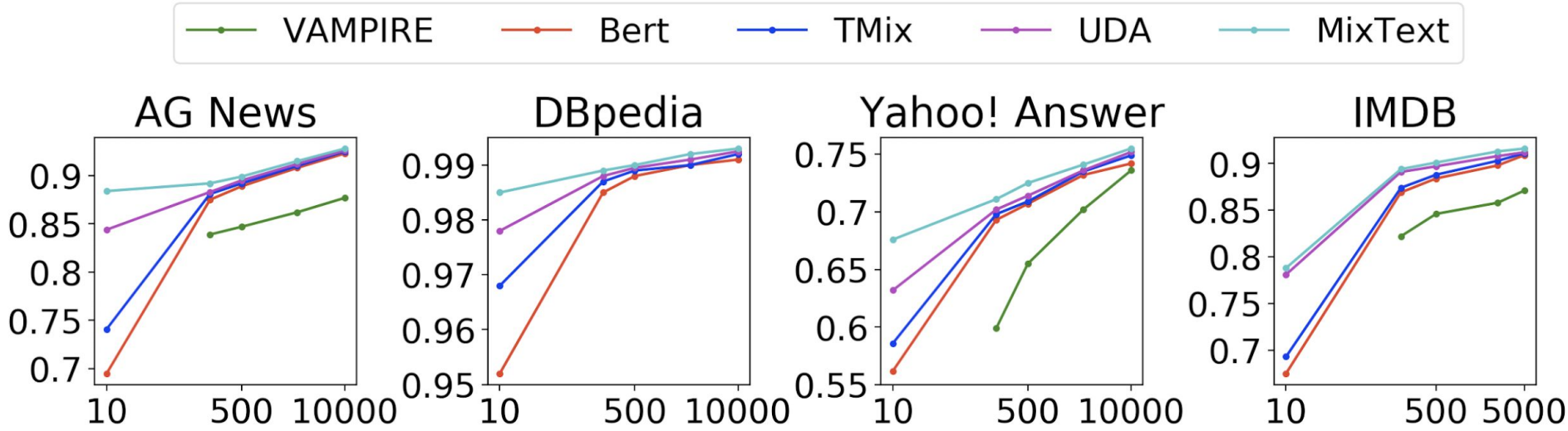
Yahoo! Answer



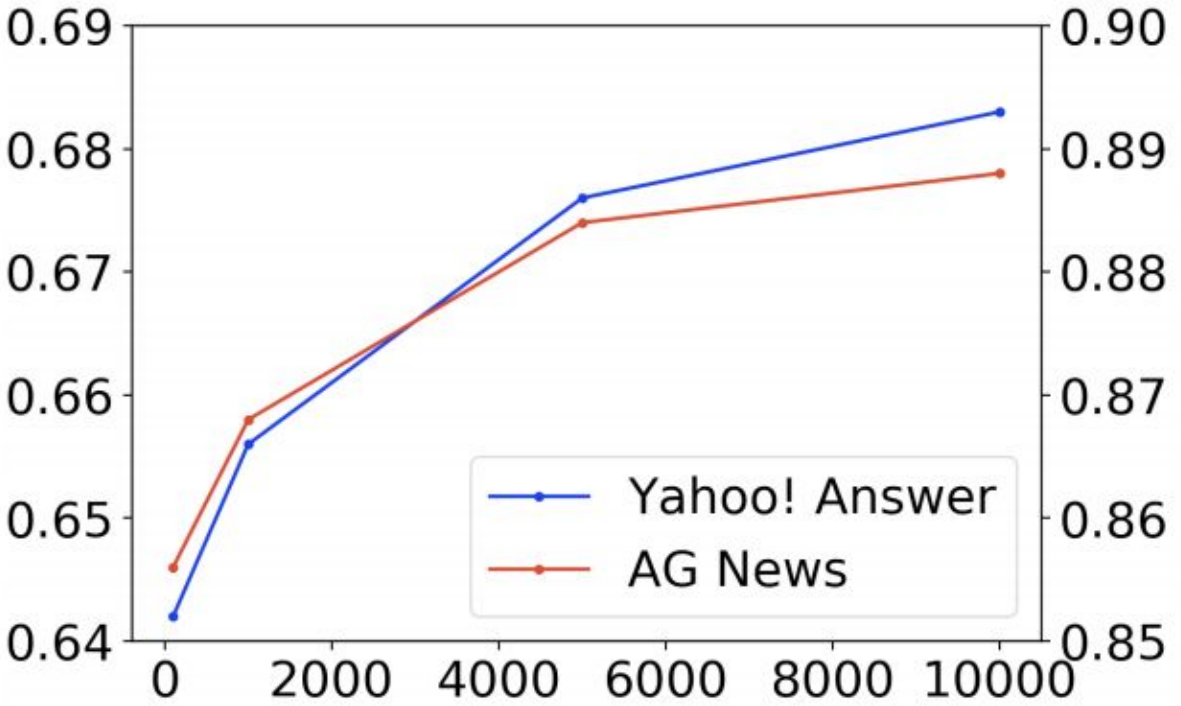
IMDB



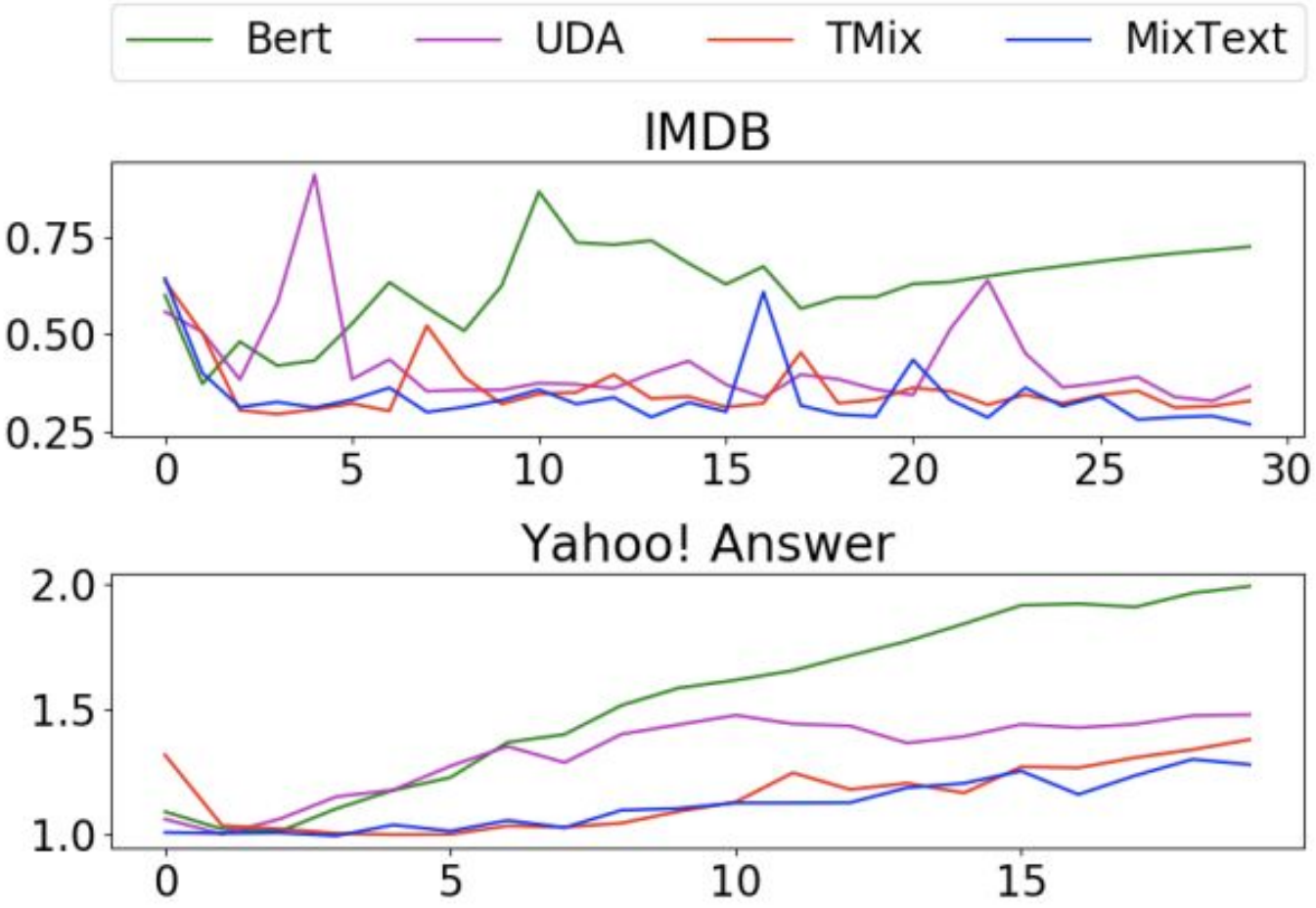
Varying the Number of Labeled Data



Varying the Number of Unlabeled Data



Avoid Overfitting



Ablation #1: Use different layer set in **TMix**

Mixup Layers Set	Accuracy(%)
$\emptyset$	69.5
{0,1,2}	69.3
{3,4}	70.4
{6,7,9}	71.9
{7,9,12}	**74.1**
{6,7,9,12}	72.2
{3,4,6,7,9,12}	71.6

Performance on *AG News*

(Here, 10 labeled data per class, consistent for other settings on different datasets)

Ablation #2: Remove parts from **MixText**

Model	Accuracy(%)
MixText	**67.6**
- weighted average	67.1
- TMix	63.5
- unlabeled data	58.6
- all	56.2

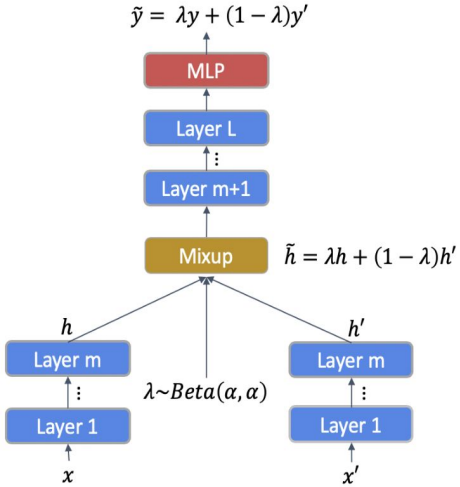
Performance on *Yahoo! Answer* (10 labeled data per class)

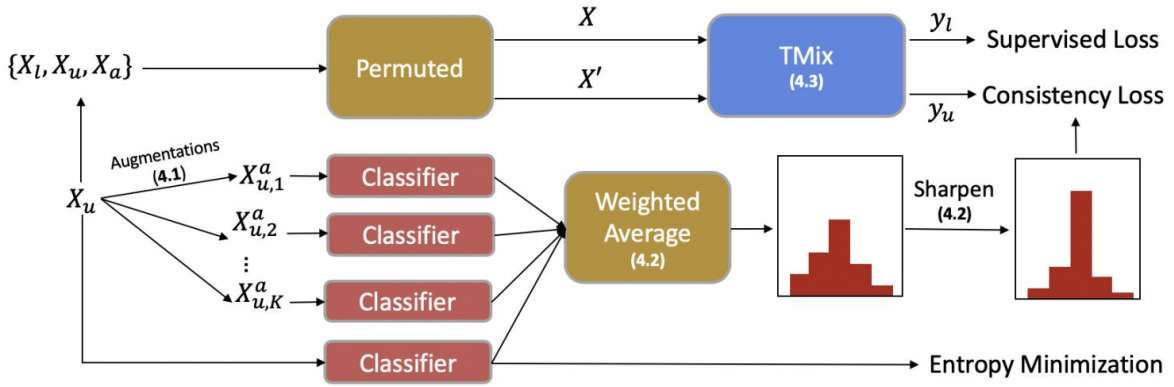
Conclusion

- ***TMix*** performs interpolations in hidden space to create infinite augmented training data
- ***Consistency training*** with interpolated data avoids overfitting in semi-supervised models
- ***MixText*** (= TMix + Consistency training) works for text classification with limited training data

MixText : Linguistically-Informed Interpolation of Hidden Space for Semi-Supervised Text Classification

Jiaao Chen, Zichao Yang, Diyi Yang





Email: jiaaochen@gatech.edu

Github: https://github.com/GT-SALT/MixText