# Early Stopping Based on Unlabeled Samples in Text Classification


HongSeok Choi, Dongha Choi, and Hyunju Lee

Gwangju Institute of Science and Technology (GIST), South Korea
{hongking9,hyunjulee}@gist.ac.kr, dongha528@gm.gist.ac.kr

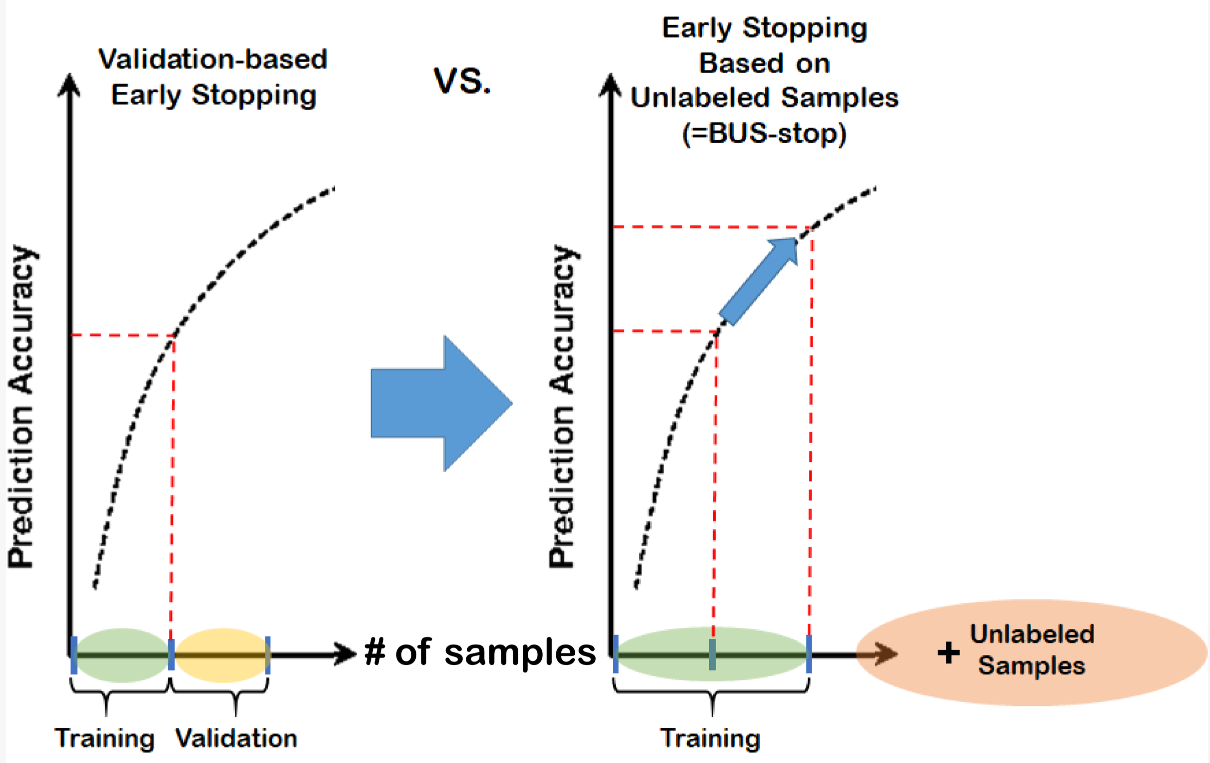


ACL 2022 22ND – 27TH MAY | 60TH MEETING | DUBLIN


## Abstract


Early stopping, which is widely used to prevent overfitting, is generally based on a separate validation set. However, in low resource settings, validation-based stopping can be risky because a small validation set may not be sufficiently representative, and the reduction in the number of samples by validation split may result in insufficient samples for training. In this study, we propose an early stopping method that uses unlabeled samples. The proposed method is based on confidence and class distribution similarities. To further improve the performance, we present a calibration method to better estimate the class distribution of the unlabeled samples. The proposed method is advantageous because it does not require a separate validation set and provides a better stopping point by using a large unlabeled set. Extensive experiments are conducted on five text classification datasets and several stop-methods are compared. Our results show that the proposed model even performs better than using an additional validation set as well as the existing stop-methods, in both balanced and imbalanced data settings. Our code is available at https://github.com/DMCB-GIST/BUS-stop.


## Motivation




1. Validation-based early stopping reduces the number of training samples for a validation set, and thus decreases the prediction accuracy.
2. In low resource settings, the small labeled set is not representative enough to be used as a stop-criterion.
3. In low resource settings, the prediction accuracy highly fluctuates during training.

## Advantages

Therefore, we propose an early **stop**ping method that is **b**ased on **u**nlabeled **s**amples, **BUS-stop**.

1. BUS-stop can *improve the performance* by using all the labeled samples for training.
2. BUS-stop can provide more *reliable stopping point* by using large unlabeled samples.
3. BUS-stop method is *related to performance metric* such as accuracy and loss.

## Method

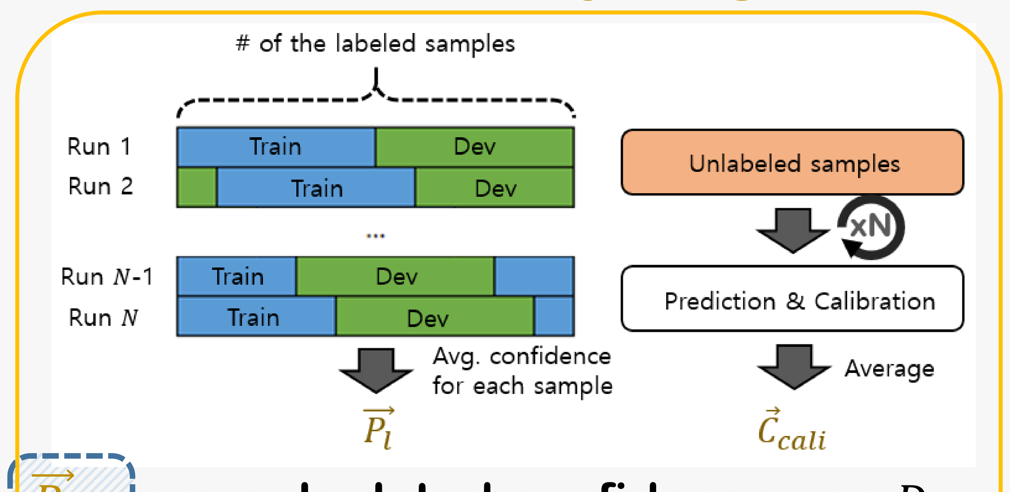
**● BUS-stop consists of two similarity measures:**

**(i) Conf-sim, $S_{conf}$**
**(ii) Class-sim, $S_{class}$**

$$S_{conf} = Euclidean(\vec{P_l}, \ddot{P}_u^{e})$$
$$S_{class} = Cosine(\vec{C}_{cali}, \widehat{C}_u^{e})$$

*** Conf-sim is the confidence similarity & Class-sim is the class distribution similarity.**

### Preliminary stage




$\vec{P_l}$ = **pre-calculated confidences on $D_l$**
$\vec{C}_{cali}$ = **pre-estimated class distribution for $D_u$**

Input: $\vec{P_l}$ & $\vec{C}_{cali}$

### Main stage

**For epoch $e$ in $\{1,2,3,\ldots\}$:**
**Train the model $M$ one epoch on $D_l$.**
**Feed $D_u$ into $M$ and obtains the confidences, $\ddot{P}_u^{e}$, and the output class distribution, $\widehat{C}_u^{e}$.**
$S_{conf} = Euclidean(\vec{P_l}, \ddot{P}_u^{e})$
$S_{class} = Cosine(\vec{C}_{cali}, \widehat{C}_u^{e})$
**BUS-stop($S_{conf}$, $S_{class}$) → stop?**
**if stop==yes → $M_{best} = M$**
**else stop==no → continue training**

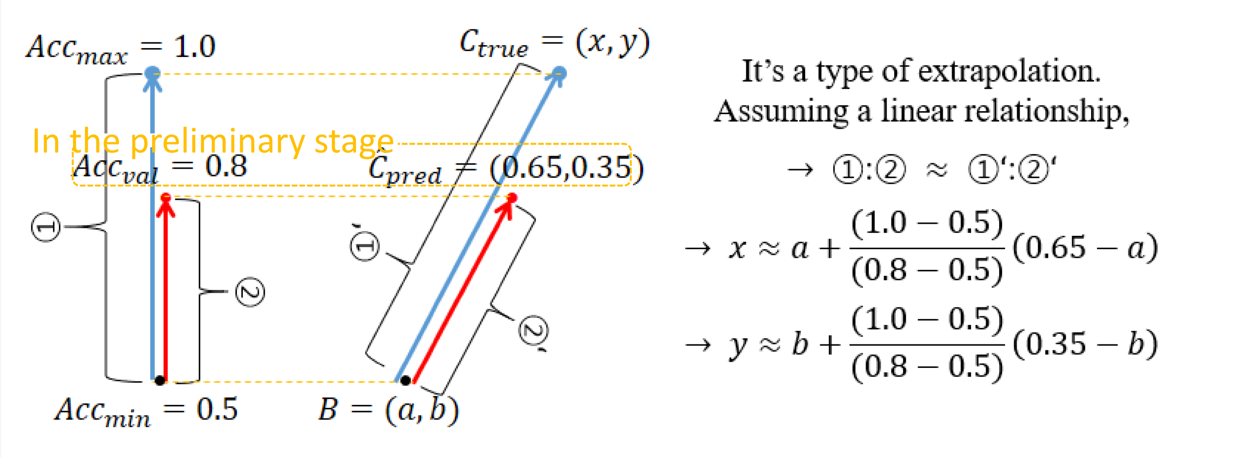
※ For $\forall j \in \{1,\ldots,n_{class}\}$,
the confidence, true class distribution, and output class distribution are defined as follows:

$$\text{Sample } x_i \rightarrow p_i = \max_j(p_{ij}), \quad C[j] = \frac{1}{n_{data}}\sum_{i=1}^{n_{data}} \mathbb{1}(y_i = j), \quad \hat{C}[j] = \frac{1}{n_{data}}\sum_{i=1}^{n_{data}} p_{ij}.$$

**● Calibration method to better estimate the true class distribution.**

$$C_{true} \approx \vec{C}_{cali} = B + \frac{(1 - Acc_{min})}{(Acc_{val} - Acc_{min})}(\hat{C}_{pred} - B)$$

**Ex) in binary classification**

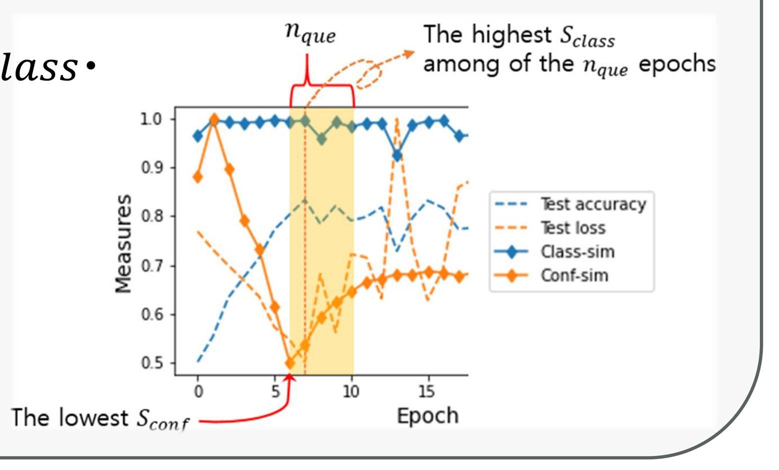



For example, if $(a, b) = (0.5, 0.5)$

$$\hat{C}_{pred} = (0.65, 0.35) \xrightarrow{\text{cali}} \vec{C}_{cali} = (0.75, 0.25)$$

**● BUS-stop is a combined method of $S_{conf}$ and $S_{class}$.**

- The combined stop-criterion is to save the model with the highest $S_{class}$ among of the epochs from the lowest $S_{conf}$ to the subsequent $(n_{que} - 1)$-th epoch.
- To implement this, we use a fixed-size ($n_{que}$) queue.




## Experimental Results

**● In various imbalanced settings**

SST-2

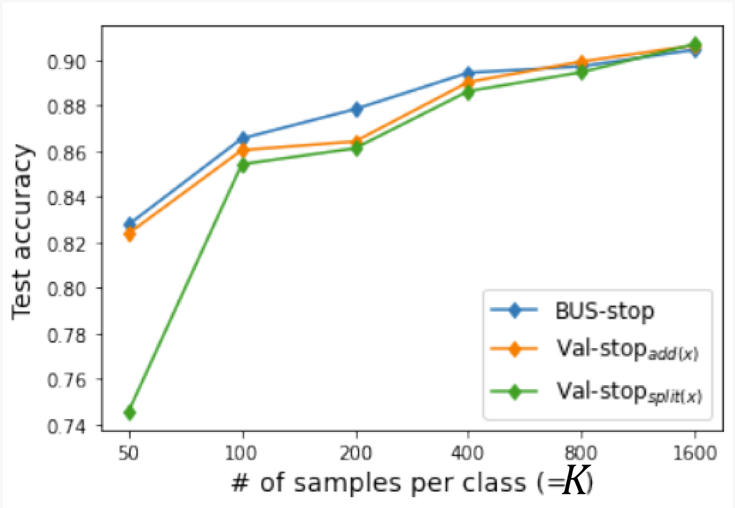
| Train | Test | 2:8 | 4:6 | 6:4 | 8:2 |
|---|---|---|---|---|---|
| 2:8 | EB | **0.845** | **0.732** | 0.643 | 0.511 |
| | BUS-stop (ours) | 0.828 | 0.719 | **0.669** | 0.521 |
| | Val-stop$_{add(25)}$ | 0.679 | 0.660 | 0.621 | **0.634** |
| 4:6 | EB | 0.860 | 0.820 | 0.790 | 0.728 |
| | BUS-stop (ours) | **0.864** | **0.825** | **0.815** | **0.808** |
| | Val-stop$_{add(25)}$ | 0.820 | 0.808 | 0.801 | 0.794 |
| 6:4 | EB | 0.790 | 0.816 | 0.825 | 0.845 |
| | BUS-stop (ours) | **0.845** | **0.826** | **0.833** | **0.864** |
| | Val-stop$_{add(25)}$ | 0.826 | 0.824 | 0.823 | 0.824 |
| 8:2 | EB | 0.611 | 0.696 | 0.774 | **0.870** |
| | BUS-stop (ours) | **0.682** | **0.714** | **0.793** | 0.865 |
| | Val-stop$_{add(25)}$ | 0.667 | 0.707 | 0.733 | 0.782 |

Avg.: EB=0.760, BUS-stop=**0.779**, Val-stop$_{add(25)}$=0.750

※ Accuracy

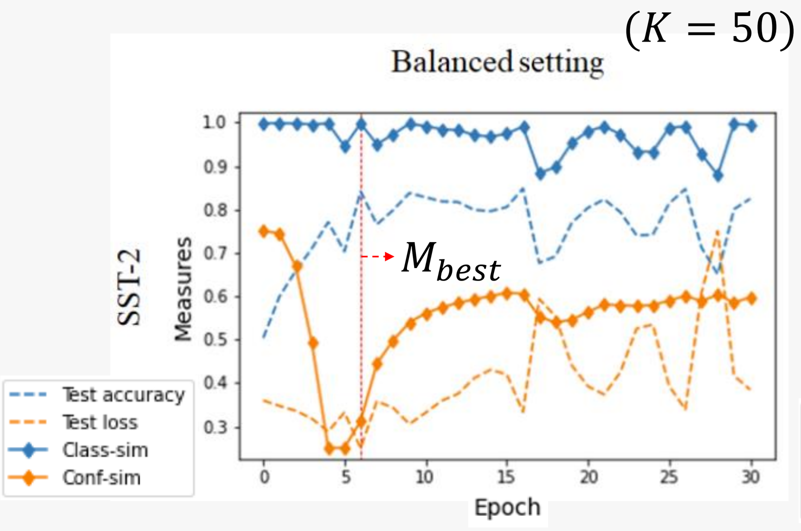
**● As the training size increases...**

IMDB




**● Balanced Classification**

($K = 50$)

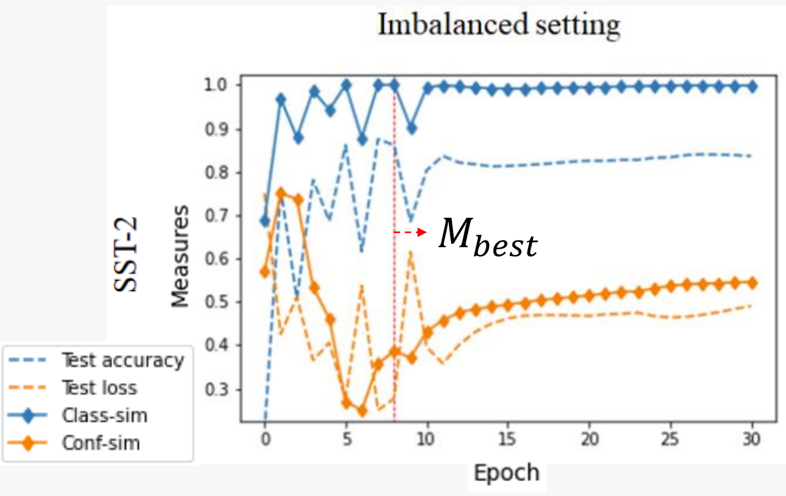



| Method | SST-2 | IMDB | ... | Avg. |
|---|---|---|---|---|
| Val-stop$_{split(25)}$ | 0.775 | 0.746 | ... | 0.826 |
| EB | 0.826 | 0.833 | ... | 0.869 |
| LID | 0.794 | 0.761 | ... | 0.84 |
| PE-stop-epoch | 0.816 | 0.826 | ... | 0.865 |
| Conf-sim (ours) | 0.807 | 0.793 | ... | 0.854 |
| Class-sim (ours) | 0.795 | 0.789 | ... | 0.844 |
| BUS-stop (ours) | 0.831 | 0.828 | ... | **0.872** |
| *Val-stop$_{add(25)}$ | 0.819 | 0.824 | ... | 0.868 |

※ Accuracy

**● Imbalanced Classification**

($K = 50$ & 2:8=neg:pos in $D_{test}$)




| Method | SST-2 | IMDB | Elec | Avg. |
|---|---|---|---|---|
| Val-stop$_{split(25)}$ | 0.788 | 0.732 | 0.783 | 0.768 |
| EB | 0.846 | 0.81 | 0.839 | 0.832 |
| LID | 0.75 | 0.712 | 0.78 | 0.747 |
| PE-stop-epoch | 0.843 | 0.821 | 0.843 | 0.836 |
| Conf-sim (ours) | 0.816 | 0.813 | 0.835 | 0.821 |
| Class-sim (ours) | 0.862 | 0.844 | 0.873 | 0.86 |
| BUS-stop (ours) | 0.86 | 0.849 | 0.876 | **0.861** |
| *Val-stop$_{add(25)}$ | 0.823 | 0.82 | 0.837 | 0.827 |

※ Accuracy

※ Val-stop$_{add(25)}$ is the validation-based stopping that uses *additional* 25 labeled samples per class, which is an unfair advantage.

## Datasets

**● Data**

| Data | Class | Test | Len |
|---|---|---|---|
| SST-2 | 2 | 1.8K | 19 |
| IMDB | 2 | 25K | 231 |
| Elec | 2 | 25K | 107 |
| AG-news | 4 | 7.6K | 38 |
| DBpedia | 14 | 70K | 49 |

- In the low resource settings, the number of training samples per class (= $K$) was set to 50.
- BERT-base was adopted as our text encoder.

## Conclusion

**Conclusion**: We conducted extensive experiments on five text classification datasets. *BUS-stop*, the proposed early stopping method, *achieved the best performance* among the existing stop-criteria, and the performance was *particularly better in imbalanced data* settings. In addition, *the proposed calibration method better estimates the true class distribution* and improves the BUS-stop performance.

**Limitation**: The running time increases with the number of unlabeled samples, and the preliminary stage requires additional running time.