# Iterative pseudo-labelling with SoftMax probability in text classification

**Jiyu Wang**

School of Computing, National University of Singapore, Singapore, 138600

E0925524@u.nus.edu

**Abstract**. Semi-supervised learning is one of the potential research fields in text classification. In this paper, semi-supervised pseudo-label training experiments are conducted using the BERT model that has been pre-trained as a baseline. Only 20% of the original dataset is used for the new training set after segmenting the training set. The raw corpus used for pseudo-label training consists of the remaining 80% of data after labels are removed, while the original test set is still utilized. The results indicate that the key to the semi-supervised pseudo-labelling method is the performance of the original model and reasonable data filtering techniques. Even though the SoftMax value used for data filtering is not precisely equivalent to model prediction accuracy, experimental results show it can somewhat reduce the error propagation problem of the model. This is consistent with earlier research. However, using SoftMax as the threshold for data screening can't bring enough benefits to the model training and make it surpass the training performance of the original data set. As a result, future studies will focus on improving the accuracy of pseudo-labelling with a more suitable data selection method to better the model's performance.

**Keywords:** Semi-Supervised Learning, Text Classification, Softmax Probability, Deep Learning.

## 1. Introduction

Text classification, a significant subfield of natural language processing, aims to label text. When utilizing fully labeled datasets, many deep learning models have achieved acceptable classification results [1]. On the other hand, when presented with limited labeled datasets, i.e., in the absence of sufficient training data, the models perform unsatisfactorily. Despite this, a solution to this problem needs to be found as quickly as possible because massive, labeled datasets are uncommon in many applications that take place in the industry. However, labeling datasets is a time-consuming process, whether done manually or mechanically, and it takes a significant amount of subject expertise to ensure accurate labeling. Consequently, there is an urgent need to investigate semi-supervised text classification using a limited amount of labeled training data in the context of the deep learning paradigm. It is essential to optimize the effective utilization of structural and feature information of unlabeled data for semi-supervised text classification to be successful.

Consistency regularization and pseudo-labeling are the two primary approaches most frequently utilized in semi-supervised learning. To achieve consistent regularization, the samples must be combined with various variations while maintaining the ability to produce consistent results within specific restrictions. On the other hand, pseudo-labeling involves filtering the label values of a limited number of samples predicted by the model as the actual labels of the samples and then training the model based on those values. To make this selection from among them, the majority of the time, the values of SoftMax are computed and used as the foundation for label classification.

However, the possibility of using the computed values of SoftMax as the foundation for label classification is still up for debate, as is the question of whether or not it is even practicable and whether it can improve the performance of the model. Because of this, the reason for this research will be based on a text sentiment analysis task to verify the two points presented above. The model's performance will serve as the primary focus of this investigation, examining how a semi-supervised learning strategy involving pseudo-labeling can improve accuracy. First, to evaluate the model's results using simply pseudo-labeling, and second, to examine the results using SoftMax thresholding as a technique to filter the pseudo-label data of the model. Both of these checks will be performed in order. According to the findings of the previous research, it is clear that introducing SoftMax into the data filtering process lowers the rate at which errors are propagated by the model; however, the performance of the initial model has a significant impact on how well semi-supervised learning works. Therefore, it is possible to conclude that the criterion for the semi-supervised learning method of pseudo-labeling is in the initial model performance and an appropriate data selection technique.

The following outline describes how this paper is structured. Section 2. is some previous works that are relevant to this topic. In Section 3. , some preliminary information and the algorithm for iterative pseudo-labeling utilizing SoftMax probabilities are presented. In Section 4. , the effectiveness of the method is evaluated through the use of several experiments. Section 5. contains a discussion of the conclusion and ideas for the future.

## 2. Related work

Research on the topic of semi-supervised learning is extensive. Because these are the most prominent semi-supervised learning approach types at the moment, the majority of the attention in this research has been concentrated on consistency regularization, pseudo-labeling-based methodologies and sentiment analysis.

**Consistency Regularization.** Different perturbations and constraints can be added to the same sample to give a consistent output. The prior assumed constraints are specified by adding a consistency regularization term to the final loss function. At this point, the data-augmented consistency regularization approach is the foundation for the mainstream consistency regularization method. In text classification, the main ways of data augmentation are back translation, lexical substitution, and adding perturbations. With its tremendous impact and low learning curve, back translation has become one of the most popular data augmentation techniques. Xie et al. translated the original samples from English to French and back to English to achieve data augmentation [2]. After obtaining sufficient data, they minimized the consistency loss between the original samples and augmented samples based on the original samples. On the other hand, the procedure of lexical substitution is analogous to that used in image data augmentation, like random clipping and scaling. Wei & Zou referred to image processing to propose the Easy Data Augmentation (EDA) method, including synonym replacement, random insertion, random exchange, and random deletion for better utilizing the limited dataset [3]. A different approach to data augmentation was provided by Miyato et al., who suggested decreasing the adversarial loss during back-propagation by adding adversarial perturbation to the word embedding [4].

**Pseudo-labeling.** Unlike consistency regularization techniques, pseudo-labeling techniques are more intuitive and adaptive. One key distinction is that consistency regularization methods typically rely on consistency constraints to perform a wide variety of transformations and augmentations to the underlying data. On the other hand, Pseudo-labeling approaches focus on highly reliable pseudo-labels to supplement the training dataset with additional labels. For the purpose of semi-supervised training of neural networks, Lee developed a simple and powerful formulation [5]. In this method, the network is trained in a supervised way with both labeled and unlabeled data simultaneously. With the goal of choosing a high-quality subset of the pseudo-labeled documents in each iteration during weakly supervised learning process, Mekala et al. examined the different pseudo-label selection techniques based on learning order [6]. Using the straightforward mathematical function called SoftMax, Chung was able to effectively and rapidly label the unlabeled data, which improved the functionality of the supervised model [7].

**Sentiment analysis.** Text sentiment analysis, also used for opinion mining or trend analysis, is a technique for determining the emotional tone of a text. Internet-based mediums, such as online discussion boards, blogs, and forums, as well as social service networks like public review sites, are responsible for the generation of insightful commentary and data pertaining to individuals, events, and goods. These comments represent a wide range of human emotions and proclivities, including joy, rage, grief, happiness, criticism, praise, and a host of other feelings and dispositions. Considering this, prospective users can read through these personal opinions to gain insight into how the general public views an event or product. Methods for analyzing the sentiment of text can be classified into two categories: 1) Sentence level sentiment analysis. Work currently being done to determine sentence sentiment in textual content typically involves the construction of sentiment databases containing various sentiment symbols, abbreviations, words, modifiers, etc. In certain experiments, a number of emotions, including anger, hostility, fear, guilt, curiosity, happiness, and sadness, will be classified. Sentences will then be labeled with one of the emotion categories and the intensity values associated with that category in order to achieve classification of the emotions contained within sentences. 2) In order to establish whether a whole piece (for instance, an online review) expresses a generally positive or negative viewpoint, it is necessary to do document-level sentiment analysis [8]. 3) The primary components of aspect-based sentiment analysis, often known as ABSA, are the extraction and categorization of aspect words and aspect sentiments. The objective of the ABSA task is to determine the emotional orientation of a certain facet of a particular target. Typically, it is broken down into two subtasks: aspect type sentiment analysis and aspect term sentiment analysis [9]. This research focuses on the sentence-level affective analysis task and verifies the effectiveness of semi-supervised learning strategies based on a binary evaluation dataset.

## 3. Methods

**Pseudo-labeling.** Let $D_L = \{(x_i, y_i)\}_{i=1}^{N_L}$ be a labeled dataset with $N_L$ samples, where $x_i$ is the input and $y_i \in \{0, 1\}$ is the corresponding label with the binary class. Consider dataset $D_U = \{x_i\}_{i=1}^{N_U}$ to be an unlabeled set of input samples (of size $N_U$) that does not include any corresponding labels. Pseudo labels $\widetilde{y_i}$ are generated for the unlabeled samples by the specific semi-supervised learning model, i.e., BERT in this research. Thus, the dataset used for training in this pseudo-labeling research is $D = \{(x_i, \widetilde{y_i})\}_{i=1}^{N_L+N_U}$, with $\widetilde{y_i} = y_i$ for the $N_L$ labeled samples. The key to decide how to generate $\widetilde{y_i}$ for the $N_U$ unlabeled samples is the softmax predictions produced by the model. To calculate the softmax value, given an unlabeled sample $\{x_i\}_{i=1}^{N_U}$, the probability that it may belong to the corresponding label $t \in \{0, 1\}$ is estimated by the formulation $\mathrm{P}(\widetilde{y_i} = t|x_i)$. In particular, the softmax predictions of the model are saved at the end of each iteration, which are used to adjust the pseudo-label $\widetilde{y_i}$ of the $N_U$ unlabeled samples.

**Text classifier.** The BERT model is the most cutting-edge example of the pre-trained contextual representations, which are constructed on a multi-layer bidirectional Transformer encoder architecture [10]. The self-attention mechanism strengthens the encoder design of the transformer by better representing the input data through the use of a multi-head attention system to focus on different areas of the text. A pre-trained BERT model was created after 800 million words were collected from BooksCorpus and 25 million words were collected from the English Wikipedia. Important steps in the BERT architecture include both pre-training and fine tuning. As part of its pre-training phase, BERT performs two types of self-supervised tasks on data from an unlabeled multilingual general domain. During the fine tuning step, trained parameters are used to initialize BERT. These parameters, along with the entire of the model, are refined using labeled data from downstream tasks especially the text classification. To apply pre-trained language models for certain downstream tasks, one needs just to make some minor adjustments to the model and modify it utilizing efficient and extensible approaches.

## 4. Experimental results and analysis

Within the scope of this research, a semi-supervised pseudo-labelling experiment is conducted based on a binary dataset of tourism evaluation (a total of 7765 data samples). Initially, the training set in the original dataset will be partitioned into two parts: only 20% of the data from the original training set will be used to create a new training dataset. In contrast, the remaining 80% will be de-annotated and used as a raw corpus for pseudo-labelling. The test set from the original dataset is still being utilized to verify the model.

First, a baseline model is developed by training it using the new training dataset. After that, the dataset will then be segmented into ten equal groups or deciles (raw corpus 01, raw corpus 02, ..., raw corpus 10). Following partitioning, the original baseline model will be used to fit and predict the raw corpus 01, with the predicted label serving as the pseudo-label for raw corpus 01. The original baseline model is retrained to make predictions on the test set once the raw corpus 01 with pseudo-label has been added to the training set. As long as the preceding steps are taken after each new data set is collected, the entire dataset can be labeled, and the prediction model can be retrained. Once all of those things have been done, the experimental results are shown in **Table 1**.

**Table 1.** Pseudo-labels model.

| Accuracy | Baseline Model | Iteration | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Pseudo-labels | 0.7231 | 0.6322 | 0.6133 | 0.6156 | 0.5998 | 0.5832 | 0.5864 | 0.6044 | 0.6111 | 0.5723 | 0.6045 |

The experimental results demonstrate that, even though the experimental setup described above can be considered a data augmentation, the performance on the model test set after 10 rounds of iterative training not only does not improve but also exhibits a relatively significant decreasing trend. The hypothesized explanation may be connected to the performance of the original baseline model. When the performance of the original baseline model is poor (0.7231), error propagation will have a significant effect after ten iterations. Therefore, it can be inferred that the performance of the original baseline model is the first crucial part of semi-supervised pseudo-labeling. In most circumstances, however, it is difficult to overcome the performance limitations of the initial model by adding data with pseudo labels iteratively to improve model performance. Therefore, establishing an appropriate approach to filter the pseudo-labeled data throughout the iterative process and enhancing data quality are required to enhance the performance of the supervised learning model version. In this research, the pseudo labels created by each iteration are filtered to illustrate inference further. In particular, it is assumed that the size of the SoftMax value denotes the degree of model confidence for the prediction outcomes, and only those prediction results with SoftMax values greater than 0.90 are maintained as the new pseudo-label data. **Table 2** displays the outcomes of the experiment.

**Table 2.** Pseudo-labels with SoftMax.

| Accuracy | Baseline Model | Iteration | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Pseudo-labels with SoftMax | 0.7231 | 0.6571 | 0.6832 | 0.6998 | 0.6475 | 0.6832 | 0.6662 | 0.6360 | 0.6512 | 0.6681 | 0.6332 |

The experimental results in **Table 2** demonstrate that when the SoftMax threshold is chosen for pseudo-labeled data filtering, the outcomes of 10 iterations of pseudo-label selection are superior to those of the experimental setting in Table 1. Nevertheless, the outcomes are still poorer than the initial base model's performance. This outcome is regarded as being in line with expectations. The SoftMax value can somewhat alleviate the error propagation issue of the model from the experimental data, although not an exact parallel to the confidence of the model prediction results. In reality, some earlier publications have successfully used SoftMax values as thresholds for pseudo-labeled data selection [9]. However, the experimental results presented in **Figure 1** suggest that the key to semi-supervised pseudo-labeling rests in the performance of the original model and an appropriate data-selecting strategy. The following study will concentrate on this as well.

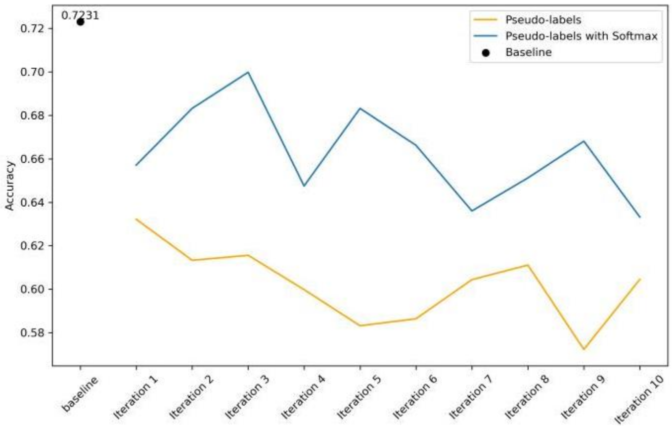

**Figure 1.** Comparison with two models.

## 5. Conclusion

Semi-supervised learning is one of the most fundamental machine learning tasks. It can save time and money because it requires less labelled data during training. Theoretically, semi-supervised learning relies on the continuity and consistency of the distribution of labelled and unlabelled data, which can be exploited through methods such as clustering, graph propagation, data augmentation, and generalized learning. As a result, methods of machine learning can take use of this for efficient structured learning, which can then be used to improve the representation of the model and, consequently, the accuracy of predictions. The progress made by semi-supervised machine learning is promising, but it has yet to catch up to the achievements of supervised learning methods like ResNet and BERT.

In this research, semi-supervised pseudo-label training experiments utilize the BERT pre-trained model as a baseline. After segmenting the training set of the original dataset, only 20% of the dataset is used for the new training set. The raw corpus utilized for pseudo-label training comprises the remaining 80% of the data after the labels have been eliminated. The test set included in the original dataset is still used when testing the model. The experiments' findings indicate that the original model's performance in conjunction with an appropriate method of data selection is necessary for successful semi-supervised pseudo-label training. Even though the SoftMax value used as the basis for data filtering is not precisely equivalent to the reliability of the model prediction results, the experimental results show that it can still alleviate the error propagation problem of the model to some extent. This is consistent with some previous works, but the achieved effect still fails to reach the expectation, making it difficult to break the performance bottleneck of the original model. As a consequence, further research will concentrate on enhancing the precision of pseudo-labeling by selecting a threshold that is more applicable to real-world scenarios to boost the model's performance.

## References

[1] Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. Information, 10(4), 150.

[2] Xie, Q., Dai, Z., Hovy, E., Luong, T., & Le, Q. (2020). Unsupervised data augmentation for con sistency training. Advances in Neural Information Processing Systems, 33, 6256-6268.

[3] Wei, J., & Zou, K. (2019). Eda: Easy data augmentation techniques for boosting performance o n text classification tasks. arXiv preprint arXiv:1901.11196.

[4] Miyato, T., Dai, A. M., & Goodfellow, I. (2016). Adversarial training methods for semi-supervi sed text classification. arXiv preprint arXiv:1605.07725.

[5] Lee, D. H. (2013, June). Pseudo-label: The simple and efficient semi-supervised learning metho d for deep neural networks. In Workshop on challenges in representation learning, ICML (V ol. 3, No. 2, p. 896).

[6] Mekala, D., Dong, C., & Shang, J. (2022). LOPS: Learning Order Inspired Pseudo-Label Selecti on for Weakly Supervised Text Classification. arXiv preprint arXiv:2205.12528.

[7] Chung, H., & Lee, J. (2022). Iterative Semi-Supervised Learning Using Softmax Probability. Co mputers, Materials and Continua, 72(3), 5607-5628.

[8] Zhang, Y., Wang, J., & Zhang, X. (2021). Conciseness is better: recurrent attention lstm model f or document-level sentiment analysis. Neurocomputing, 462, 101-112.

[9] D'Aniello, G., Gaeta, M., & La Rocca, I. (2022). KnowMIS-ABSA: an overview and a referenc e model for applications of sentiment analysis and aspect-based sentiment analysis. Artificial Intelligence Review, 1-32.

[10] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectio nal transformers for language understanding. arXiv preprint arXiv:1810.04805.