Co-training pseudo-labeling for text classification with support vector machine and long short-term memory

Sri Handayani[1,2], **Rizal Isnanto**[3], **Budi Warsito**[4]
[1]Doctoral Program of Information Systems, Diponegoro University, Semarang, Indonesia
[2]Department of Information and Communication Technology, Semarang University, Semarang, Indonesia
[3]Department of Computer Engineering, Diponegoro University, Semarang, Indonesia
[4]Department of Statistics, Faculty of Science and Mathematics, Diponegoro University, Semarang, Indonesia

Article Info

Article history:

Received May 4, 2024
Revised Feb 3, 2025
Accepted Mar 15, 2025

Keywords:

Co-training
Long short-term memory
Pseudo-labeling
Semi-supervised learning
Support vector machine

ABSTRACT

The scarcity of labeled data may hamper training text-processing models. In response to this issue, a novel and intriguing strategy that combines the co-training method and pseudo-labeling design is applied to enhance the model's performance. This method, a component of an efficient semi-supervised learning paradigm for processing and comprehending text, is a fresh perspective in the field. The model, which combines a support vector machine (SVM) for classification and long short-term memory (LSTM) for text sequence interpretation, is a unique approach. By introducing samples that may be marginalized in the labeled data, the co-training approach could help solve the class imbalance problem by using a small amount of labeled data and the rest unlabeled. This study assesses the model's performance using a student dataset from higher education institutions to establish a threshold for each model's degree of confidence and ascertain how much the model can be generalized depending on the threshold. The SVM threshold was calculated as >=0.88, and the LSTM threshold was calculated as >=0.5 using a mixture of confidence metrics.

Corresponding Author:

Sri Handayani
Department Technology of Information and Communication, Semarang University
Soekarno Hatta, West Tlogosari, Semarang City, Central Java, Indonesia
Email: sri@usm.ac.id

1. INTRODUCTION

Text clustering in the era of big data and machine learning is one of the most critical tasks in text mining, which aims to group texts according to their classification. Text clustering has been beneficial in many applications, such as recommendation systems [1], [2], data mining management [3], and position detection [4]. With the presence of the digital era, most people enjoy sharing and finding various content on social media and using text is very influential in carrying information. Text clustering can identify trending topics, detect sentiment patterns, or recommend personalized content, offering a wealth of valuable insights on social media. Text clustering is a challenging task. Conversely, the diversity of text categories often presents a significant challenge that cannot be ignored-data data imbalance. This situation arises when the classes in a dataset are not evenly distributed, leading to one class having significantly more samples than the others. In the context of classification, this imbalance can pose a serious problem. Machine learning algorithms typically prioritize the majority class and may overlook the less representative minority class.

One study highlights the challenges of data measurement in text classification. It proposes a promising solution-the use of text generation with mode GPT-2 and long short-term memory (LSTM) to balance the dataset. This approach shows potential in improving the performance of the classification model

[5]. Text clustering methods have also been tried to deal with highly imbalanced data, which tend to lead to poor solutions when minor data clusters or clusters with small proportions of cases disappear [6], [7]. They may also obtain easy solutions in which all text instances are in the same cluster [8]. Furthermore, contrastive language learning methods compare the first and second words to help correct errors and improve word mastery. However, these methods still need better solutions, especially in highly imbalanced text collections [7]. It's important to note that imbalanced data, if processed directly, can lead to data proportions that reduce the performance of the classification algorithm [9].

The authors propose a solution to reducing data and improving model performance on unlabeled data using pseudo-labeling. This technique, which involves training a model on labeled data to produce pseudo labels for unlabeled data, offers several benefits. Pseudo-labeling in semi-supervised learning has been developed to classify data sets with partial labels [10]. It has also been developed in semi-supervised learning for pattern recognition [11], because it can significantly increase learning effectiveness through the distribution of information from unlabeled data and information from labeled data [12]. Pseudo-labeling will also be used for the classification process [13], [14], from unlabeled test data so that the semi-supervised learning algorithm can be used properly. Thus, the classification process that will be carried out can facilitate the search for information based on certain categories needed. Semi-supervised learning also functions by training the model on labeled and unlabeled data. Labeled data provides information about the relationship between input and output variables, while unlabeled data captures the underlying data structure [15].

Several studies in higher education have been conducted with a collaborative spirit, aiming to identify and improve student acceptance rates. One such collaborative effort involves the application of predictive models with the aid of machine learning. This approach can help Universities identify students who need help early, thereby enabling interventions that can improve student graduation rates [16]. Another area of interest is a hybrid machine learning approach that effectively combines unsupervised and supervised learning methods, leading to improve accuracy in predicting student academic performance [17]. Additionally, research on ensemble models, which involve the combination of three distinct algorithms (decision tree, logistic regression, and neural networks), has yielded an effective tool for improving student retention rates. These tools, developed through collaboration, have the potential to help higher education institutions reduce the risk of dropout and the associated financial losses [18]. There is also the application of the Feature selection and construction for radial basis function (FSC4RBF) using algorithm. This algorithm uses the grammatical evolution method for feature selection and feature construction in the RBF network, which aims to improve the generalization ability in predicting student academic performance, especially in predicting the duration of study and final grades based on previous learning data, to effectively improve the prediction ability in the college environment [19]. However, no research has applied the co-training method in college management. Through this article, the authors try to apply the pseudo-labeling technique using the co-training method, in which the student data obtained is text data from a questionnaire distributed at a college to determine the potential of students at the college. While processing the questionnaire results, the data will be labeled based on potential and non-potential criteria. A small portion of the student data will be used as training data, which then helps form a wrapped dataset to classify labels on unlabeled data (testing data) using the co-training semi-supervised learning concept.

Pseudo labeling in semi-supervised learning can degrade the labeling quality because it does not represent the classes in the labeled dataset. The co-training method can utilize semantic labels to overcome this problem [20]. The labeling framework with the co-training method can solve the multi-view weak label learning problem using pseudo label vectors [21]. Comprehensive experimental studies have been conducted to demonstrate the effectiveness of the semi-supervised learning and co-training method for multi-label [22]. Over the past decade, the co-training method has gained popularity in the research and industry communities and has been successfully applied in several real-world applications [23]. The co-training method has also been used for transfer learning (TL) [24]. Co-training is an emerging topic in machine learning. It aims to extract knowledge gained in a source task or domain and use it to facilitate target predictive learning functions in a different task or domain. Co-training can improve the generalization ability of a model by utilizing information from unlabeled data. Co-training enables machines to think from multiple perspectives like humans by dividing data into multiple views, designing learners scientifically, and estimating confidence labels accurately. This method can also improve classification accuracy and model convergence. In addition, co-training's versatility in various research tasks, such as error classification and person identification based on audio-visuals, opens up a world of potential applications and research opportunities [25].

The co-training process generally involves three main steps: view acquisition, learner differentiation, and label confidence estimation [25]. View acquisition aims to obtain two independent and sufficient data views, each of which can be used by a different learning model. If independent views are not naturally available, they can be constructed using pre-trained models [26]. Once the views are acquired, the next step is learner differentiation, in which two different learning models are drilled separately for each view. This approach ensures that each model learns a unique representation of the data, which increases the models' ability to complete and correct each other's errors [27]. In the label confidence estimation stage, each model makes predictions on unlabeled data and assesses the model's confidence in those predictions. Predictions with high confidence are then used as pseudo-labels to train the other model, allowing both models to reinforce each other and improve overall prediction accuracy [20]. This emphasis on high-confidence predictions provides a reassuring framework for the models to learn and improve.

This article reviews the co-training method of confidence label estimation, in which the label confidence estimation is an important step in the co-training method with the aim that the data testing is not mislabeled, which can reduce the ability of data testing. According to the label confidence estimation method, the label confidence estimation can be divided into implicit and explicit estimations. Most algorithms in implicit estimation use the degree of difference between the results of training data and testing data to reflect the confidence of the pseudo label. Still, if there is unlabeled data, if it is mislabeled, it will reduce performance during iteration, so the accuracy of implicit estimation is lower than explicit estimation [28]. The algorithm in explicit estimation uses exact numbers to display the confidence label. Basically, this algorithm uses learning output in probabilistic form, the difference in model accuracy before and after using pseudo-labeled samples, or the similarity between the pseudo label of the current unlabeled data and the surrounding labeled data [25].

Co-training uses implicit estimation of the support vector machine (SVM) and LSTM algorithms. The reason for choosing SVM to be drilled together with LSTM is that SVM is commonly used for text classification [29]. SVM has a more mature and mathematically clearer concept than other classification techniques. SVM can also solve classification and regression problems with linear and non-linear. SVM is also used because it has advantages in handling classification problems, especially when the data has high dimensions or an imbalance between classes. SVM effectively separates classes with maximum margin, making it suitable for classification with complex feature data, such as medical images or text in a visual context. In the case of image classification, SVM is often relied on because of its ability to find the optimal hyperplane that separates data based on relevant features, which ultimately helps reduce classification errors [30]. The LSTM algorithm is a predictive model to process sequential data in virtual learning environments (VLEs). LSTM is applied to evaluate and predict student learning outcomes based on data collected from student interactions with online learning platforms. This model can analyze sequences of student behavioral data, such as login patterns, quiz interactions, and course activities, to identify trends that may indicate success or difficult in learning [31], [32]. LSTM is also chosen because of its ability to handle time series data, which is very important for modeling learning data continuously generated in online learning environments. This algorithm allows the system to remember important information in the long term, which makes it ideal for predicting student performance based on student activity history [5]. In another article, the LSTM algorithm is used as part of a combined CNN-LSTM model to detect insults in text. LSTM plays a role in processing the sequence of words from the text to understand the temporal or sequential context, which is important in detecting the meaning of insults based on the context of the sentence. Combining CNN for initial feature extraction and LSTM to understand the context sequence makes this model more effective in accurately detecting insults in text [26], [32]. Research on algorithms in machine learning that exists so far is comparing the SVM with LSTM algorithm and the deep believe network (DBN) algorithm using the fuzzy logical relationships (FLR) model [14].

Recent research on pseudo-labeling in semi-supervised learning using the SVM and LSTM co-training methods has shown that pseudo-labeling not only improves model performance but is also effective in situations in which labeled data is very limited. The similar pseudo label exploitation for semi-supervised classification (SIMPLE) algorithm, which effectively utilizes pseudo-labels to enhance accuracy in semi-supervised classification, is a real-world proof of concept [13]. The co-training method, which combines several algorithms to train models simultaneously, is gaining traction, particularly in addressing data imbalance problems. The combination of the co-training method with modern algorithms in the medical field demonstrates that using pseudo-labels with a high level of confidence (high-confidence pseudo labels) can significantly enhance the accuracy and stability of the model in medical image segmentation [33]. In a textual context, co-training has proven effective in overcoming the problem of under-labeled data by combining SVM and LSTM-based models to optimize data sequence classification. In the visual context, label clustering and co-training improve the quality of pseudo-labels and overall model performance by grouping labels based on visual similarity to reduce misclassification of similar classes. This approach, using an embedding-based label representation, creates groups of visually similar labels and uses these groups in the pseudo-labeling process [30]. The future of semi-supervised learning looks promising with recent techniques in co-training, such as combining implicit and explicit confidence estimates, to improve labeling accuracy. Entropy-based and margin-based confidence algorithms are currently widely used to ensure the accuracy of labels on test data. For example, the right step is to develop consistency-based semi-supervised active learning techniques that minimize labeling costs while improving model accuracy through accurate confidence estimates [34].

This paper offers a unique contribution by combining two different models, SVM and LSTM, in a co-training framework. SVM is used to handle static feature-based classification, while LSTM is used to manage and understand text data sequences. This integration provides advantages because it allows more comprehensive modeling of sequential text data compared to traditional co-training methods that rely only on one model type. Furthermore, both models will be given a specific trust threshold: SVM with a threshold value of 0.88 and LSTM with 0.5. This meticulous parameterization of the models, a feature that is not always explained in detail in previous studies, instills trust in the application of semi-supervised models. In addition, another contribution of this article is that while previous studies have focused on semi-supervised learning applications in domains such as image recognition or medical analysis, this article focuses on its application to educational data, specifically to assess the condition of students in higher education institutions. This is an under-explored area, and this study shows how semi-supervised techniques can be applied for more effective educational evaluation purposes.

In co-training methods, semi-supervised learning, label confidence refers to the extent to which the model has confidence or certainty about the label given to data. These methods, which combine models trained on two feature sets, are particularly effective in improving performance. Here are some general concepts that can be used to measure label confidence in the context of co-training,

- Entropy: a prevalent approach involving entropy utilization to assess the model's uncertainty regarding labels. Entropy quantifies the model's probability distribution to resemble a uniform discrete distribution. Higher entropy values indicate lower confidence levels in the assigned label [34]. Using two models will produce joint entropy or M_j representing the entropy associated with the joint occurrence of classes i and j. This entropy value measures the uncertainty or irregularity level associated with the two classes' joint probability distribution [35]. The higher the entropy value, the greater the uncertainty in the occurrence of the two classes simultaneously. The joint entropy can be calculated using the formula,

$$M_j = -\Sigma_{j=1}^{s_j}\left(P_{ij}\right)\log_2\left(P_{ij}\right)\ j= 1,\ldots,\mathrm{n} \qquad (1)$$

in which,
M_j is this representing the joint entropy for the j-th category.
S_j is total number of categories or classes in the j-th variable being considered.
P_{ij} is the probability of joint occurrence of two classes, namely class i and class j, in which i represents the first class and j represents the second class, and $\log_2\left(P_{ij}\right)$ This is the logarithm base 2 of the joint probability P_{ij} .
n is the total number of classes.

The minus sign (-) in the joint entropy formula is because entropy measures uncertainty in a probability distribution. The minus sign ensures that the entropy value is always positive. The logarithm of a number between 0 and 1 produces a negative value. Therefore, the minus sign ensures that entropy is positive because uncertainty (entropy) cannot be negative. In the context of the Co-training method, entropy is used to evaluate the model's confidence in pseudo-labels.

- Margin-based confidence: label confidence can also be measured based on the margin between the probability of the primary label and the second most likely label. The larger the margin, the higher the model's confidence in the selected label [36].
- Vote-based confidence: in co-training, data can be labeled by both models. If both models give the same label, then the confidence in the label is higher than if they give different labels [20].
- Confidence measures combination: sometimes, combining multiple confidence measures can provide a more comprehensive picture of the model's confidence in a given label [33].

2. METHOD

Co-training is a semi-supervised learning approach including two models: SVM and LSTM. They mutually oversee each other and refine themselves by leveraging insights from unlabeled data [37]. The co-training algorithm presented is as Algorithm 1.

```
Algorithm 1: Co-training Algorithm
    LD ← Labeled data
    UD ← Unlabeled data
    RD ← Random data UD → UD’
```

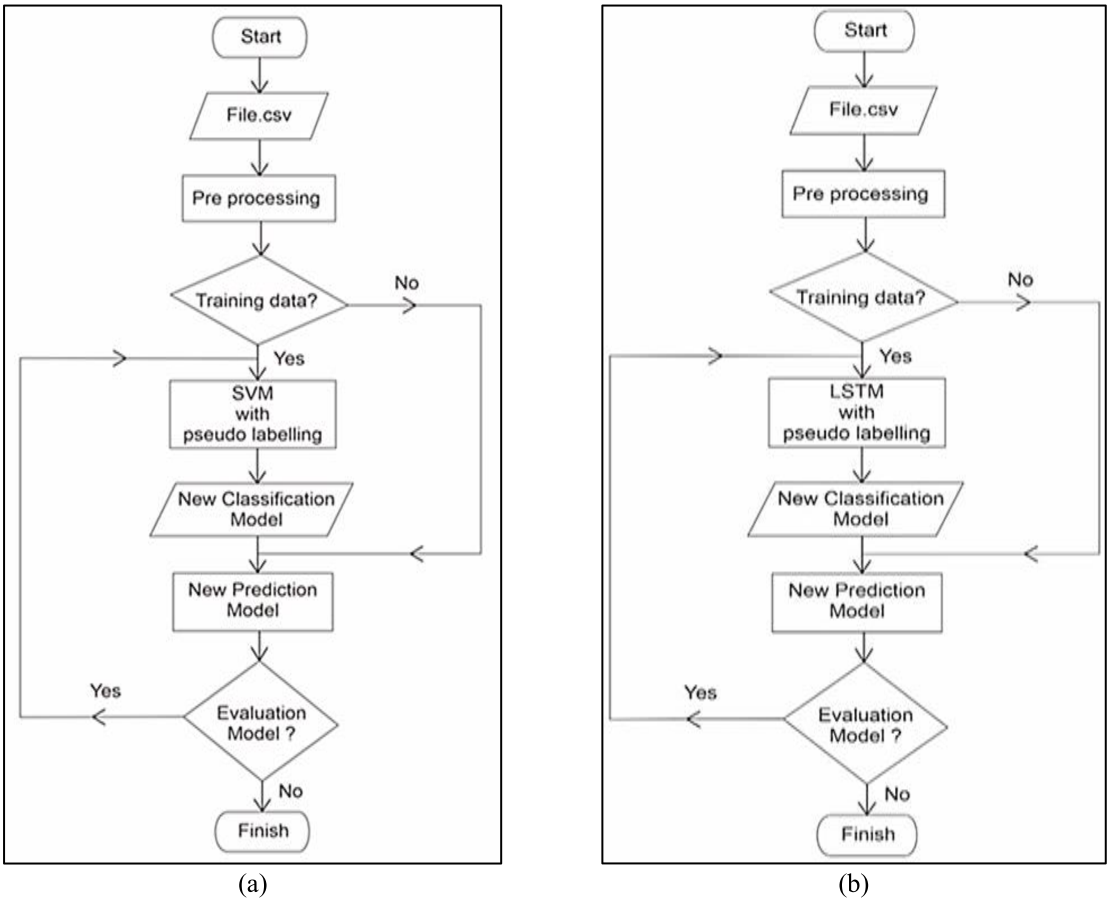
```
4     Failedlter ← 0
5     While (Failedlter < n)
6             h1 (SVM) ← Training ← First LD
7             h2 (LSTM) ← Training ← Second LD
8             UD'1 ← UD' Classification with h1
9             UD'2 ← UD' Classification with h2
10            UD'1 ← Label best classification example
11            UD'2 ← Label best classification example
12            LD ← Append UD'1 and UD'2
13            if No example added
14                    Faildlter++
15            else
16                    Faildlter ← 0
17            end
18    End
```

In the co-training process, the unlabeled data (UD) is labeled when both classifiers (LSTM and SVM) independently agree to assign a label. The first classifier (h1 or SVM) assigns a label with a confidence level greater than a certain threshold value X, indicating a high level of certainty. Similarly, the second classifier (h2 or LSTM) assigns a label with a confidence level greater than a threshold value of Y. The co-training process is halted when both classifiers fail to agree on the same label during an iteration. The co-training algorithm in Algorithm 1 tests on 100 datasets with RD =5 set as the fixed value. These datasets include 90 unlabeled testing data points and 10 labeled training data points. This proportion of labeled and unlabeled data is used to analyze the performance of co-training. Nigam and Ghani [38] have argued that although the greatest performance improvement occurs when the labeled data is very small, adding further labeled data provides a smaller improvement.

The co-training algorithm flow in Algorithm 1 begins with the initialization process of LD and UD. At this initialization stage, the confidence score threshold for SVM and LSTM is determined, and the labeled data set to train each model and the unlabeled data set (RD) are determined. In line 6, iterations will be carried out on the training process of the SVM classifier. Line 8 produces predictions with SVM (UD'1), and in line 10, each prediction made by the SVM model is accompanied by a confidence score. This score indicates the level of confidence of the model in the classification decision made by the SVM, in which the confidence score will be the threshold for unlabeled data when given a pseudo-label so that the unlabeled data gets a higher confidence score than the SVM threshold during the training set. In line 7, iterations will be carried out on the training process of the LSTM classifier to produce predictions with LSTM (UD'2) and confidence scores on unlabeled data, then add pseudo-labels with higher confidence than the LSTM threshold to the training set. For the SVM and LSTM algorithms that are being tested, the co-training approach aims to determine the confidence level threshold that is used during the co-training procedure. For the SVM and LSTM algorithms being tested, the co-training approach aims to determine the confidence level threshold used during the co-training procedure.

The phases in the co-training process flow are detailed in Figures 1(a) and 1(b). Figure 1(a) shows that the technique begins by loading a dataset from a file (e.g., File.csv). The dataset is then pre-processed to clean and prepare the data for training. Once preprocessing is complete, the workflow checks whether training data is available. If no training data is present, the process terminates. If training data is available, an SVM classifier is applied with pseudo-labeling to generate labels for the unlabeled data. These pseudo-labels are then used to create a new classification model. The updated classification model is utilized to generate predictions for the data. The workflow then evaluates the new prediction model to determine its performance. If the model's performance meets the desired criteria, the process may iterate again to refine the predictions further. The process concludes if the performance is unsatisfactory or further iterations are not required. This iterative cycle aims to improve the classification model by integrating pseudo-labeling with SVM.

Similarly, the co-training methodology depicted in Figure 1(b) starts by loading a dataset from a file (e.g., File.csv). Once the dataset is loaded, it undergoes preprocessing to clean and prepare the data for further use, such as handling missing values, normalizing features, or text tokenization. After preprocessing, the system checks if training data is available. If no training data is present, the process ends. However, if training data is available, the next step involves using an LSTM model with pseudo-labeling. The LSTM model is trained on the existing labeled data, and pseudo-labels are generated for the unlabeled data with a confidence threshold to ensure quality. These pseudo-labeled samples are then used to update the training dataset, which allows for creating a new classification model. This updated model is used to generate new predictions. The newly trained model's performance is evaluated based on predefined criteria. If the evaluation indicates the model's performance is satisfactory, the process can be repeated to refine the model with additional pseudo-labeling iterations. The process concludes if the evaluation fails to meet the required standards or if further iterations are unnecessary. This iterative workflow enhances the classification model's accuracy by integrating pseudo-labeling and LSTM for sequential data.



(a) (b)

Figure 1. Flowchart co-training process for (a) SVM and (b) LSTM

3. RESULTS AND DISCUSSION

The implementation of co-training to produce an information framework focuses on leveraging labeled and unlabeled data to enhance the accuracy and reliability of predictive models. This approach integrates complementary algorithms, such as SVM and LSTM, to generate pseudo-labels iteratively, enabling the development of a robust framework that supports data classification, decision-making, and performance evaluation [33]. The information system framework provides an outline of a simple information system. It shows the main parts that are the same in all information storage and retrieval institutions, such as libraries, archives, documentation and information centers, regardless of the level of mechanism or type of information managed by institutions. By looking at the information system framework, it is expected to look the information system framework what components information units can be understood and what process should occur [37]. The authors will use labeled input data (training data), which makes up around 5% of the entire dataset, to test the suggested information system framework. This selection of the percentage of training data is consistent with study literature practices, where the percentage of training data usually falls between 1% and 50% of the total dataset, depending on the particular problem domain and resources available, in addition to testing data (unlabeled data) [38].

To determine the threshold value of the label confidence level of the SVM classifier and the LSTM classifier in the co-training method, the steps are to train the first and second classification models using the data labeled as desired in the pseudo-labeling process. For the first classification model training, the results of the labeled data classification are used to classify unlabeled data so that the label confidence level of each prediction will be obtained. Likewise, for the second classification model training, the results of the labeled data classification will be used to classify unlabeled data to obtain the label confidence level of each prediction in the second classification.

The diagram in Figure 2 illustrates the system framework for co-training using SVM and LSTM with a pseudo-labeling approach to process and classify student data. The system begins with two datasets: training data, which consists of a small set of supervised data with labeled samples, and testing data, which contains many unsupervised, unlabeled data. The LSTM model first processes the training data and generates pseudo-labels for the testing data based on its learning from the labeled samples. Simultaneously, the SVM model is applied to the testing data to generate its pseudo-labels. These pseudo-labels from both models are then compared in three key stages.

- Significance of comparing pseudo-labeling results: the results of pseudo-labeling, a crucial step in model evaluation performed by both LSTM and SVM, are compared with the test data. This comparison is vital in evaluating how well the models perform on the unlabeled dataset.
- Confidence prediction comparison: the predictions made by the LSTM and SVM models are compared in terms of their confidence levels. This step is invaluable in identifying which model produces more reliable pseudo-labels.
- Comparison with training data: the pseudo-labeling results for the testing data are compared against the original labeled training data to evaluate consistency and alignment with the initial dataset.

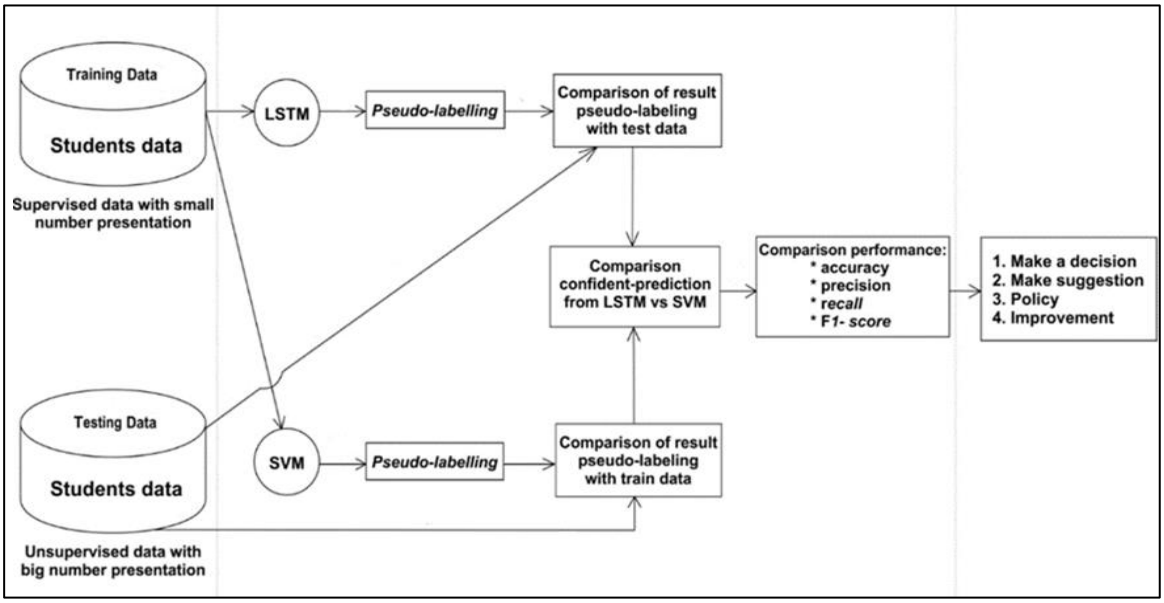

Training Data
Students data
Supervised data with small number presentation
LSTM
Pseudo-labelling
Comparison of result pseudo-labeling with test data
Comparison confident-prediction from LSTM vs SVM
Comparison performance:
* accuracy
* precision
* recall
* F1- score
1. Make a decision
2. Make suggestion
3. Policy
4. Improvement
Testing Data
Students data
Unsupervised data with big number presentation
SVM
Pseudo-labelling
Comparison of result pseudo-labeling with train data

Figure 2. Co-training method information system framework

The system's performance is then assessed using accuracy, precision, recall, and F1-score metrics. Based on this evaluation, the system provides actionable outputs, including decisions, suggestions, policies, and recommendations for further improvements. This iterative framework ensures robust classification of student data by leveraging both models' strengths and refining predictions through pseudo-labeling.

By leveraging the robust capabilities of Google Colab, the trial implementation of co-training between SVM and LSTM yielded a threshold label value for SVM of 0.88 and a threshold label value for LSTM of 0.5, as depicted in Figure 3(a). It demonstrates that the level of label confidence for SVM is >=0.88. Figure 3(b) demonstrates that the level of label confidence for LSTM is >=0.5.

Label confidence estimation is an important step in a semi-supervised algorithm [25]. Label confidence estimation aims to prevent unlabeled data from being mislabeled, which can lower the label threshold. According to the label confidence estimation method, it can be divided into explicit estimation and implicit estimation. Implicit estimation is highly dependent on the training data (labeled data); unlabeled data may be mislabeled, and it will degrade the learner's performance during iteration, so the accuracy of implicit estimation is lower than explicit estimation. However, the cost required is smaller than the explicit estimation.

Most algorithms in implicit estimation use the degree of difference in the results to reflect the current pseudo-label confidence. Algorithms in explicit estimation use exact numbers to represent the label confidence. Typically, these algorithms use learning outputs in probabilistic form, the difference in model accuracy before and after using pseudo-labeled data, or the similarity between the pseudo-labels of the current unlabeled data and the surrounding labeled data.

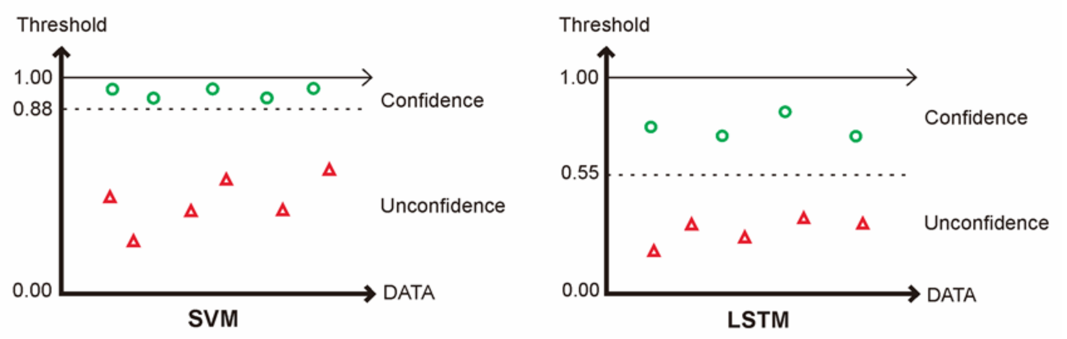


Figure 3. The relation of data and threshold for (a) SVM and (b) LSTM

Explicit estimation is applied in this SVM and LSTM Co-training trial. As the first classifier, SVM is used for prediction, and LSTM, as the second classifier, is used for understanding the sequence of text data in this co-training. In the first classification (SVM), when the label given to the testing data has a label confidence level (threshold) below 0.88, the training data will be iterated again so that the label that will be given to the testing data has a label confidence level (threshold) >=0.88. Likewise, in the second classification (LSTM), when the label given to the testing data has a label confidence level (threshold) below 0.5, the training data will be iterated again so that the label that will be given to the testing data has a label confidence level (threshold) >=0.5. The concept of measuring the label confidence level (threshold) in this article combines the label confidence measures from classification 1 and 2 to provide a more comprehensive view to which the model is confident with the label given.

Some studies align with the method the authors applied, using two different models (SVM and LSTM) in the co-training framework. A collaborative approach in which each model produces predictions on unlabeled data and highly confident predictions are used as pseudo-labels to train the other model. For example, research that combines two deep learning models trained on two different representations of text data for sentiment classification. Each model produces predictions on unlabeled data, and highly confident predictions are used as pseudo-labels to train the other model. This approach improves the accuracy of sentiment classification on massive open online courses (MOOC) or calls for large-scale online courses open to the public, usually provided by universities or online learning platforms such as Coursera, edX, or future learn [39]. Furthermore, studies introducing Meta co-training have the principle that two models trained independently on different data representations can correct each other's errors and improve the overall accuracy, or two pre-trained models can be combined on different representations of the same data. Each model produces predictions on unlabeled data, and predictions with a high confidence level are used as pseudo-labels to train the other model [26]. The approach taken by the authors uses two different models in the co-training framework, or the approach using two learning models for the MOOC forum and the Meta co-training technique. These models showed improved performance on semi-supervised classification tasks, which shows that this strategy effectively improves model performance on various machine-learning tasks.

4. CONCLUSION

In this study, Google Colab was used to facilitate the application of the SVM and LSTM co-training techniques in the context of semi-supervised learning. The first datasets used to calculate label confidence levels from SVM and LSTM were obtained from student responses to questionnaires at three Central Java universities, which were collected via Google Form submissions. The determined threshold labels for SVM and LSTM are ≥0.88 and ≥0.5, respectively. Classifier 1 (SVM) and classifier 2 (LSTM) thresholds are crucial in defining the following iterations, especially when applying pseudo labels to testing data. In particular, iterations happen when the pseudo-label for the test data is less than the corresponding threshold values of 0.5 for LSTM and 0.88 for SVM.

ACKNOWLEDGMENTS

The author would like to thank Tri Endah Retno Kusumaningrum, S.Sos., M.I.Kom., for her valuable assistance during her visit to Tidar University in Magelang for data collection.

FUNDING INFORMATION

The research described in this article was funded by Semarang University.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Sri Handayani	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓		✓
Rizal Isnanto		✓		✓		✓				✓		✓	✓	
Budi Warsito	✓	✓		✓	✓	✓				✓		✓	✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis

I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting

Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no known financial conflicts of interest or personal relationships that could have influenced the work reported in this paper.

INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

DATA AVAILABILITY

The data supporting the findings of this study are available from Universitas Semarang. Access to these data is restricted and can only be obtained with permission from Universitas Semarang. Data requests can be made to the author at sri@usm.ac.id, with permission from Universitas Semarang.

REFERENCES

[1] W. Liu, X. Zheng, M. Hu, and C. Chen, “Collaborative filtering with attribution alignment for review-based non-overlapped cross domain recommendation,” *WWW 2022-Proceedings of the ACM Web Conference 2022*, pp. 1181–1190, 2022, doi: 10.1145/3485447.3512166.

[2] W. Liu, X. Zheng, J. Su, M. Hu, Y. Tan, and C. Chen, “Exploiting variational domain-invariant user embedding for partially overlapped cross domain recommendation,” *SIGIR 2022-Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 312–321, 2022, doi: 10.1145/3477495.3531975.

[3] S. Stieglitz, M. Mirbabaie, B. Ross, and C. Neuberger, “Social media analytics-challenges in topic discovery, data collection, and data preparation,” *International Journal of Information Management*, vol. 39, pp. 156–168, 2018, doi: 10.1016/j.ijinfomgt.2017.12.002.

[4] J. Li *et al.*, “Unsupervised belief representation learning with information-theoretic variational graph auto-encoders,” *SIGIR 2022-Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1728–1738, 2022, doi: 10.1145/3477495.3532072.

[5] S. Shaikh, S. M. Daudpota, A. S. Imran, and Z. Kastrati, “Towards improved classification accuracy on highly imbalanced text dataset using deep neural language models,” *Applied Sciences*, vol. 11, no. 2, pp. 1–20, 2021, doi: 10.3390/app11020869.

[6] A. Hadifar, L. Sterckx, T. Demeester, and C. Develder, “A self-training approach for short text clustering,” *ACL 2019-4th Workshop on Representation Learning for NLP, RepL4NLP 2019-Proceedings of the Workshop*, pp. 194–199, 2019, doi: 10.18653/v1/w19-4322.

[7] D. Zhang *et al.*, “Supporting clustering with contrastive learning,” *NAACL-HLT 2021-2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, pp. 5419–5430, 2021, doi: 10.18653/v1/2021.naacl-main.427.

[8] X. Ji, A. Vedaldi, and J. Henriques, “Invariant information clustering for unsupervised image classification and segmentation,”

2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019, pp. 9864-9873, doi: 10.1109/ICCV.2019.00996.

[9] T. E. Tallo and A. Musdholifah, "The implementation of genetic algorithm in smote (Synthetic minority oversampling technique) for handling imbalanced dataset problem," *2018 4th International Conference on Science and Technology (ICST)*, Yogyakarta, Indonesia, 2018, pp. 1-4, doi: 10.1109/ICSTC.2018.8528591.

[10] C. Gong, D. Tao, S. J. Maybank, W. Liu, G. Kang, and J. Yang, "Multi-modal curriculum learning for semi-supervised image classification," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3249–3260, 2016, doi: 10.1109/TIP.2016.2563981.

[11] J. Enguehard, P. O'Halloran, and A. Gholipour, "Semi-supervised learning with deep embedded clustering for image classification and segmentation," *IEEE Access*, vol. 7, pp. 11093–11104, 2019, doi: 10.1109/ACCESS.2019.2891970.

[12] M. K. Xie, J. H. Xiao, H. Z. Liu, G. Niu, M. Sugiyama, and S. J. Huang, "Class-distribution-aware pseudo-labeling for semi-supervised multi-label learning," *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[13] Z. Hu, Z. Yang, X. Hu, and R. Nevatia, "Simple: similar pseudo label exploitation for semi-supervised classification," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 15094–15103, 2021, doi: 10.1109/CVPR46437.2021.01485.

[14] S. Panigrahi and H. S. Behera, "A study on leading machine learning techniques for high order fuzzy time series forecasting," *Engineering Applications of Artificial Intelligence*, vol. 87, 2020, doi: 10.1016/j.engappai.2019.103245.

[15] S. Emtiyaz and M. R. Keyvanpour, "Customers behavior modeling by semi-supervised learning in customer relationship management," *Advances in Information Sciences and Service Sciences*, vol. 3, no. 9, pp. 229–236, 2011, doi: 10.4156/AISS.vol3.issue9.31.

[16] Z. Azizah, T. Ohyama, X. Zhao, Y. Ohkawa, and T. Mitsuishi, "Predicting at-risk students in the early stage of a blended learning course via machine learning using limited data," *Computers and Education: Artificial Intelligence*, vol. 7, 2024, doi: 10.1016/j.caeai.2024.100261.

[17] G. Al-Tameemi, J. Xue, I. H. Ali, and S. Ajit, "A hybrid machine learning approach for predicting student performance using multi-class educational datasets," *Procedia Computer Science*, vol. 238, pp. 888–895, 2024, doi: 10.1016/j.procs.2024.06.108.

[18] A. M. Rabelo and L. E. Zárate, "A model for predicting dropout of higher education students," *Data Science and Management*, 2024, doi: 10.1016/j.dsm.2024.07.001.

[19] V. Christou *et al.*, "Performance and early drop prediction for higher education students using machine learning," *Expert Systems with Applications*, vol. 225, 2023, doi: 10.1016/j.eswa.2023.120079.

[20] I. Nassar, S. Herath, E. Abbasnejad, W. Buntine, and G. Haffari, "All labels are not created equal: enhancing semi-supervision via label grouping and co-training," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 7237–7246, 2021, doi: 10.1109/CVPR46437.2021.00716.

[21] X. Xu, W. Li, D. Xu, and I. W. Tsang, "Co-labeling for multi-view weakly labeled learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 6, pp. 1113–1125, 2016, doi: 10.1109/TPAMI.2015.2476813.

[22] W. Zhan and M. L. Zhang, "Inductive semi-supervised multi-label learning with co-training," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. Part F129685, pp. 1305–1314, 2017, doi: 10.1145/3097983.3098141.

[23] Z. H. Zhou, "A brief introduction to weakly supervised learning," *National Science Review*, vol. 5, no. 1, pp. 44–53, 2018, doi: 10.1093/nsr/nwx106.

[24] F. Zhuang *et al.*, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2021, doi: 10.1109/JPROC.2020.3004555.

[25] V. Piccialli and M. Sciandrone, "Nonlinear optimization and support vector machines," *Annals of Operations Research*, vol. 314, no. 1, pp. 15–47, 2022, doi: 10.1007/s10479-022-04655-x.

[26] J. C. Rothenberger and D. I. Diochnos, "Meta co-training: two views are better than one," *arXiv-Computer Science*, pp. 1-17, 2023, doi: 10.1145/3474085.3475622.

[27] C. Tan, J. Xia, L. Wu, and S. Z. Li, "Co-learning: learning from noisy labels with self-supervision," in *MM 2021-Proceedings of the 29th ACM International Conference on Multimedia*, pp. 1405–1413, 2021, doi: 10.1145/3474085.3475622.

[28] D. T. Nguyen, C. K. Mummadi, T. P. N. Ngo, T. H. P. Nguyen, L. Beggel, and T. Brox, "Self: learning to filter noisy labels with self-ensembling," *8th International Conference on Learning Representations, ICLR 2020*, 2020.

[29] M. Goudjil, M. Koudil, M. Bedda, and N. Ghoggali, "A novel active learning method using svm for text classification," *International Journal of Automation and Computing*, vol. 15, no. 3, pp. 290–298, 2018, doi: 10.1007/s11633-015-0912-z.

[30] M. H. S. Enas, H. A. Saleh, and E. A. Khalel, "Classification of mammograms based on features extraction techniques using support vector machine," *Computer Science and Information Technologies*, vol. 2, no. 3, pp. 121–131, 2021, doi: 10.11591/csit.v2i3.pp121-131.

[31] E. Ismanto and N. Effendi, "An LSTM-based prediction model for gradient-descending optimization in virtual learning environments," *Computer Science and Information Technologies*, vol. 4, no. 3, pp. 199–207, 2024, doi: 10.11591/csit.v4i3.pp199-207.

[32] M. M. Ben Ismail, "Insult detection using a partitional CNN-LSTM model," *International Journal of Data Analysis Techniques and Strategies*, vol. 14, no. 4, pp. 336–349, 2023, doi: 10.1504/IJDATS.2022.129175.

[33] Z. Shen, P. Cao, H. Yang, X. Liu, J. Yang, and O. R. Zaiane, "Co-training with high-confidence pseudo labels for semi-supervised medical image segmentation," *IJCAI International Joint Conference on Artificial Intelligence*, vol. 2023, pp. 4199–4207, 2023, doi: 10.24963/ijcai.2023/467.

[34] M. Gao, Z. Zhang, G. Yu, S. Arık, L. S. Davis, and T. Pfister, "Consistency-based semi-supervised active learning: towards minimizing labeling cost," *Computer Vision – ECCV 2020*, pp. 510–526, 2020, doi: 10.1007/978-3-030-58607-2_30.

[35] W. Chen, H. R. Pourghasemi, and S. A. Naghibi, "A comparative study of landslide susceptibility maps produced using support vector machine with different kernel functions and entropy data mining models in China," *Bulletin of Engineering Geology and the Environment*, vol. 77, no. 2, pp. 647–664, 2018, doi: 10.1007/s10064-017-1010-y.

[36] G. Gigerenzer, U. Hoffrage, and H. Kleinbolting, "Probabilistic mental models: a brunswikian theory of confidence," *Psychological Review*, vol. 98, no. 4, pp. 506–528, 1991, doi: 10.1037/0033-295X.98.4.506.

[37] J. M. Griffiths and D. W. King, "US information retrieval system evolution and evaluation (1945-1975)," *IEEE Annals of the History of Computing*, vol. 24, no. 3, pp. 35–55, 2002, doi: 10.1109/MAHC.2002.1024761.

[38] K. Nigam and R. Ghani, "Analyzing the effectiveness and applicability of co-training," *International Conference on Information and Knowledge Management, Proceedings*, vol. 2000-Janua, pp. 86–93, 2000, doi: 10.1145/354756.354805.

[39] J. Chen, J. Feng, X. Sun, and Y. Liu, "Co-training semi-supervised deep learning for sentiment classification of MOOC forum posts," *Symmetry*, vol. 12, no. 1, 2020, doi: 10.3390/SYM12010008.

BIOGRAPHIES OF AUTHORS

Sri Handayani has been a permanent lecturer and studier at Semarang University since 2007, with competencies in networking, computer hardware, internet of thing and business intelligence. The author completed D3 Telecommunication Engineering from Polytechnic ITB in 1996, completed strata-1 telecommunication engineering from Sekolah Tinggi Telkom Bandung in 2001. The author continued his strata-2 studies at the Master of Information Technology at Gadjah Mada University and graduated in 2005. In 2024, the author is in the process of completing further studies in the doctoral program at the Doctor of Information Systems, Postgraduate School, Diponegoro University. She can be contacted at email: sri@usm.ac.id or srihandayanimansyur@students.undip.ac.id.

Rizal Isnanto obtained his bachelor and master degrees in electrical engineering at Gadjah Mada University, Yogyakarta - Indonesia, starting from 1994 to 2002. While the doctoral degree was received at the Department of Electrical Engineering and Information Technology, Gadjah Mada University in 2013. He received an honorary degree as a professor on June 1, 2023 as a professor in the field of image processing. He has held several positions, including Secretary of the Computer Systems Study Program 2 May-31 October 2016, Head of the Department of Computer Engineering, Faculty of Engineering, Diponegoro University (2016-2020), Head of the S1-Computer Engineering Study Program, Faculty of Engineering, Diponegoro University (2020-2022). He currently serves as Secretary of the Information Systems Doctoral Study Program at the Postgraduate School of Diponegoro University (2022-present). The study conducted to date is related to the fields of information systems, biomedical and biometric image processing and pattern recognition. He has received an award from the President of the Republic of Indonesia in the form of a 10-year Satyalancana Karya Satya Honor Certificate (2015) and a Charter of Appreciation as a Volunteer "Movement of one million volunteers for election supervision 2014" from the Central Bawaslu. He can be contacted at email: rizal@ce.undip.ac.id.

Budi Warsito completed elementary to high school in Sukoharjo in 1993. His undergraduate degree was obtained at the Statistics Study Program, Faculty of Mathematics and Natural Sciences, Gadjah Mada University in 1998. Then the S2 and S3 degrees were obtained from the Mathematics Study Program, Faculty of Mathematics and Natural Sciences, Gadjah Mada University in 2004 and 2016, respectively. Since 1999, he has been teaching at the Department of Statistics, Faculty of Science and Mathematics, Diponegoro University Semarang. His study interests are neural network modeling and machine learning with applications in various fields such as environment, finance, and health. Several book titles have been written, namely Kapita selecta statistica neural network; Spatial regression: application with R; and Modification of spatial regression model. He also has 2 patents and dozens of copyrights, especially in the field of computer programming. He has held several positions, including Secretary of the UNDIP FSM Statistics S1 Study Program (2008-2011), Secretary of the SPs UNDIP Information Systems Master Study Program (2018-2020), Chair of the SPs UNDIP Information Systems Master Study Program (2020-2021), Chair of the SPs UNDIP Environmental Science Doctoral Study Program (2021-present), Chair of the Computational Statistics and Data Science KBK Department of Statistics FSM UNDIP (2017-present). He can be contacted at email: budiwarsito@lecturer.undip.ac.id.