JURNAL INFOTEL
VOL. 17, NO. 3, AUGUST 2025, PP. 472–489. DOI:10.20895/INFOTEL.V17I3.1344

RESEARCH ARTICLE

Semi-Supervised Sentiment Classification Using Self-Learning and Enhanced Co-Training

Agus Sasmito Aribowo[1,*], Siti Khomsah[2], and Shoffan Saifullah[3]

[1,3]Department of Informatics, Universitas Pembangunan Nasional "Veteran" Yogyakarta, Yogyakarta, 55283, Indonesia
[2]Data Science Study Program, Telkom University, Purwokerto, 53147, Central Java, Indonesia
[3]Faculty of Computer Science, AGH University of Krakow, Poland

*Corresponding email: sasmito.skom@upnyk.ac.id

Received: March 31, 2025; Revised: August 18, 2025; Accepted: August 26, 2025.

Abstract: Sentiment classification is often performed manually, but manual labeling is inefficient. Therefore, automated labeling with machine learning is essential. Building a computerized labeling model is challenging when labeled data is scarce, which reduces the accuracy of the model. This study proposes a semi-supervised learning (SSL) framework for sentiment analysis with limited labeled data. The framework integrates self-learning and enhanced co-training. The co-training model combines three machine learning methods: Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression (LR), with TF-IDF and FastText used for feature extraction. The model generates pseudo-labels, and the label with the highest confidence from SVM, RF, or LR is selected in the self-learning step.This framework is applied to English and Indonesian datasets, each run five times. The performance difference between the baseline model (without pseudo-labels) and SSL (with pseudo-labels) is small; the Wilcoxon signed rank test confirms this with a p-value <0.05. The results show that SSL produces pseudolabels with quality close to the original labels. Although the test performs well on four datasets, it has not surpassed the supervised baseline. SSL labeling has been shown to be more efficient than manual labeling, taking only 10-20 minutes to label thousands of samples. In conclusion, self-learning in SSL with co-training effectively labels unlabeled multilingual datasets with limited data, though it has not yet converged across all datasets.

Keywords: semi-supervised learning (SSL), sentiment analysis, self-learning, co-training

 ISSN: 2085-3688; e-ISSN: 2460-0997

1 Introduction

The growth of social media has led to an explosion of textual data, making sentiment analysis crucial for applications in customer feedback, brand management, and public opinion monitoring. However, traditional supervised learning methods are highly dependent on large, manually labeled datasets, which are costly and time-consuming [1]. To address these limitations, semi-supervised learning (SSL) has emerged as a promising approach, using limited labeled data to label unlabeled data. Unlabeled data are more readily available than labeled data with high validity. Therefore, semi-supervised learning approaches are expected to address traditional labeling issues [2]. SSL concept uses a small labeled dataset alongside a larger unlabeled dataset, reducing the inefficiency of manual annotation [3].

This research aims to evaluate the effectiveness of two SSL techniques - self-learning and enhanced co-training - in sentiment classification in English and Indonesian datasets. An ensemble-based co-training framework is proposed, in which three classifiers (SVM, Random Forest, and Logistic Regression) are combined with two vectorization methods (TF-IDF and FastText), forming a six-model ensemble designed to improve robustness through representation of various features (Figure 1). In this framework, self-learning is applied to iteratively label high-confidence predictions to expand the training set, while enhanced co-training is employed to utilize multiple classifiers trained in diverse feature subsets to annotate new instances [4].

To achieve the objectives of this research, the following steps are conducted: (1) assessing the performance of SSL in low-label scenarios, (2) comparing it with supervised baselines using accuracy, F1 Score, and computational efficiency, (3) evaluating its applicability to multilingual scenarios, and (4) identifying challenges for future work. It is hypothesized that the proposed SSL framework can address the problem of model performance with a limited amount of labeled data.

To provide context, this research is compared with previous studies that have employed various semi-supervised learning (SSL) approaches. An earlier study developed an ensemble SSL model using SVM and RF with TF-IDF and n-grams, showing that classifier choice significantly impacts performance: SVM excelled in binary tasks, while RF was more robust in multi-class settings [5]. Subsequent work demonstrated that the integration of hyperparameter adjustment within SSL iterations improves model performance, with random search often outperforming grid search [6]. Other studies applied hybrid supervised–unsupervised methods to English digital payment reviews, where RF achieved an F1 Score of 73.8% for sentiment and 58.8% for emotion analysis [7]. For Indonesian texts, ensemble models using unigrams, bigrams, and trigrams showed that SSL performance depends on data volume and feature–algorithm compatibility [8].

Previous research has demonstrated the potential of SSL in sentiment analysis [5–8], but its effectiveness in multilingual contexts remains uncertain. The present study contributes to this area by employing a co-training approach with limited labeled training data. The proposed SSL framework is applied to six datasets, including English datasets (IMDB, US Airlines) and Indonesian datasets (Hate Speech, Sentiment 1, Sentiment 2, Emotion). This study advances the development of SSL methods for sentiment classification by leveraging limited labeled data, with a particular emphasis on multilingual sentiment analysis.

2 Research Method

2.1 Dataset Preparation

This study utilizes datasets in English and Indonesian, sourced from public repositories such as US Airline, IMDB, Indonesian Hate Speech, Sentiment 1, Emotion, and Sentiment 2, all collected under ethical guidelines and proper licensing. All data is user-generated and publicly accessible; although the raw data contains a wealth of information such as a username, the user's personal identity is not used. The analysis focuses on the content (comments or reviews) without taking into account the profile of the commentor or reviewer.

Our proposed SSL model was applied to six diverse datasets to assess its generalization across languages, domains, and varying numbers of classes. As summarized in Table 1, the datasets consist of English corpora (IMDB [9], US Airline [10]) and Indonesian corpora (Hate Speech [11], Sentiment 1 [12], Emotion and Sentiment 2 [13]), with varying numbers of classes (binary up to six classes). This selection enables a comprehensive evaluation of the robustness of the model to handle linguistic variation, class imbalance, and multilingual sentiment tasks, ensuring its applicability to real-world scenarios.

Table 1: Datasets Used for SSL Modeling

No	Dataset	Language	Number Records	Number Class
1	IMDB [9]	English	50000	Positif, negative
2	US Airline [10]	English	14848	Positive, Negative, Neutral
3	Hate Speech [11]	Indonesian	13169	Hate Speech, Non-Hate Speech
4	Sentiment 1 [12]	Indonesian	10807	Positive, Negative, Neutral
5	Emotion [13]	Indonesian	7079	Joy, Love, Sad, Fear, Anger, Neutral
6	Sentiment 2 [13]	Indonesian	12759	Positive, Negative, Neutral

2.2 Data Preprocessing and Feature Extraction

In the preprocessing phase, all datasets go through preprocessing steps, including data cleaning, vectorization, and feature extraction [14]. The data cleaning process aims to eliminate noise and meaningless characters. This stage consists of several tasks, such as conversion to lowercase, removal of stop words, tokenization, removal of numbers, removal of all nonalphabetic characters and punctuation, and stemming. First, all text was converted to lowercase to ensure consistency. Stop words or words that have no meaning, such as conjunctions, adverbs, to be, and the like, are removed. The text is then tokenized by splitting sentences into individual words (tokens) for easier processing. Numerical characters were eliminated as they did not contribute to sentiment analysis. Non-alphabetic symbols and punctuation were removed to reduce unnecessary noise. The data is then transformed into its basic word form (stemming). e.g., running $\rightarrow run, therebyconsolidatingwordvariantsintoasinglerepresentation.$

TF-IDF and FastText are used for word embedding. TF-IDF is used to calculate the weight of each word in the corpus. The frequency of the term is the number of occurrences of the term t in document D divided by the number of terms in document D. IDF is how the terms are distributed in document D. TF-IDF has been extensively used in text

analysis of the Indonesian language, with applications spanning various domains. For example, it has been utilized in a Twitter sentiment analysis model for feature extraction [15], feature extraction of reviews of mobile banking application opinions [16]. Furthermore, TF-IDF was used in an experimental study of kernel functions in the SVM algorithm for sentiment analysis in Indonesian [17], and feature extraction was used to detect anxiety in social media [18]. In addition to the Indonesian context, TF-IDF has also been applied in research on semi-supervised learning for sentiment analysis classifiers [5], emotion detection in text [19], and sentiment analysis of financial news data [20].

FastText was developed based on the Word2Vec model and was introduced by Facebook AI Research in 2016 [21]. Compared to Word2Vec, FastText offers several advantages. One key advantage is its ability to handle words outside the dictionary (OOV) - words that have not been seen during training or are not included in the dictionary [22].

For example, FastText can process words with affixes, such as "me-" and "di-" in Indonesian, or suffixes like "-y", "-ly", and "-ism" in English. Words like "fun" and "funny", which are semantically related but differ in their affixes, can still be represented meaningfully. In Indonesian, the affixation increases the variation of the word, although the root meaning often remains the same, for example, "memakan" and "makan" both refer to the activity of eating. FastText represents words as a collection of character n-grams. For instance, if n = 3, the word "happy" is split into trigrams such as "hap", "app","ppy". The resulting word vector is computed as the average of the embeddings of these n-grams.

2.3 Proposed Design Semi-Supervised Learning

The proposed semi-supervised learning strategy for document annotation using both self-learning and co-training techniques. Self-learning iteratively labels unannotated data, stopping after three iterations or upon convergence, when no further data can be labeled. Meanwhile, co-training employs a multi-vectorization approach: V1 (using TF-IDF) and V2 (using FastText) combined with three classifiers (ML1 = SVM, ML2 = RF, and ML3 = LR), forming six models in an ensemble. This dual representation enhances the model's ability to capture both statistical and contextual features of text. Only models achieving the highest F1 scores contribute to classification decisions through weighted voting, while weaker models are excluded.

Figure 1 illustrates the SSL architecture for handling annotated and unannotated datasets. The process begins with dataset division, for example 20% labeled and 80% unlabeled records are separated, with 80% of the labeled data used for training and 20% for testing. Unknown keywords in the testing and unlabeled datasets are then cleaned using FastText word vectors, replacing unrecognized words with their closest vector match. In the co-training phase, word vectors are generated using two vectorization algorithms: one that determines the significance of each word within a document (TF-IDF) and another that captures contextual meaning by analyzing the structure of words at the character level (FastText), forming six classifier models (RF, SVM, and LR for each vectorization method). Two different vectorizers are combined, namely the statistical-based vectorizer (TF-IDF) and the semantic-based vectorizer (FastText), in order to capture a broader range of linguistic features.

In addition, three different machine learning models, each with distinct working mechanisms, are employed to extract insights according to their respective strengths. These models are optimized through hyperparameter tuning. The unlabeled dataset is then

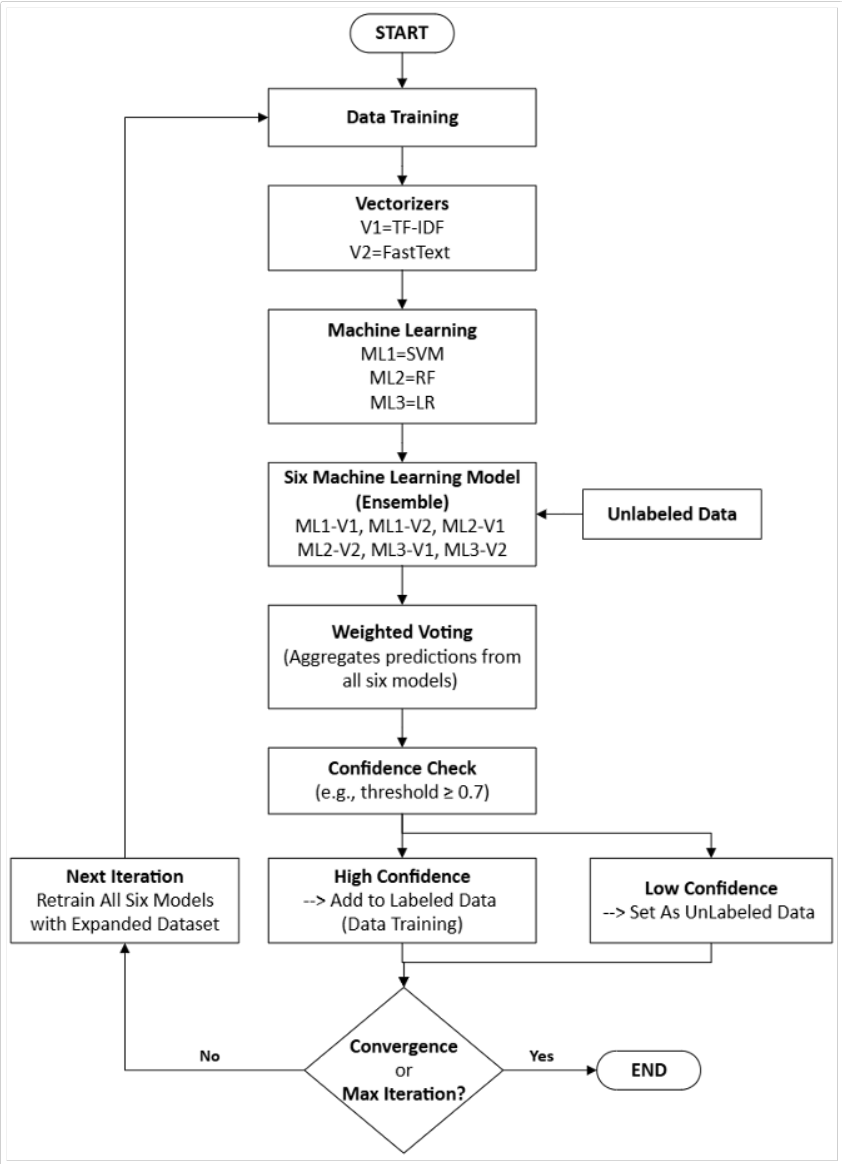

Figure 1: Proposed SSL Framework

pseudo-labeled using the six models, assigning probability scores to each comment. Six base models (SVM, RF, LR with TF-IDF, and FastText) are trained on the initial labeled subset. In each SSL iteration, the four models with the highest F1 Score on the validation set are selected for weighted voting to generate pseudo-labels. This selection is dynamic and not tied to the vectorization method, allowing the ensemble to adapt to performance changes across iterations. The self-learning process utilizes weighted voting to determine whether pseudo-labeled data should be incorporated into the training dataset. This cycle

is repeated up to three times or until convergence is achieved, ensuring efficient annotation of the unlabeled dataset while optimizing processing time.

To assess the performance improvements introduced by the semi-supervised learning framework, each experiment was evaluated by comparing the performance of the model under two conditions: baseline model (model without co-training) and SSL proposed. The baseline refers to the accuracy of the model when trained solely on the small-labeled subset without any incorporation of pseudo-labeled data, while SSL represents the performance after the semi-supervised iterative learning process, in which additional data were automatically labeled and added to the training set. The difference in accuracy, along with changes in precision, recall, and F1 Score, provides a clear measure of the impact of the SSL approach.

2.4 Performance Measure: Matrix Confusion

A confusion matrix is used for evaluating the performance of classification models, providing a detailed comparison between the predicted and actual labels. It consists of four key components Table 2 True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), which allow the calculation of important evaluation metrics such as accuracy, precision, recall, and F1 Score [23]. By analyzing these components, researchers can gain deeper insight into model behavior, optimize performance, and address specific shortcomings in classification tasks.

Table 2: Confusion Matrix

	Actual Positive	**Actual Negative**
Predicted Positive	True Positive (TP)	False Positive (FP)
Predicted Negative	False Negative (FN)	True Negative (TN)

To validate the performance of the model, this study uses four key evaluation metrics: Accuracy, precision, recall, and F1 score. Accuracy measures the overall correctness of the model. At the same time, precision assesses the accuracy of positive predictions and recall assesses the model's ability to identify positive instances. Accuracy, as defined in Eq.(1), is a useful metric for balanced datasets where false positive and false negative values are relatively equal. However, it may not be reliable when dealing with imbalanced data distributions.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

Unlike accuracy, the F1 Score ensures a better balance between precision and recall. The F1 Score combines both precision and recall to offer a balanced view of model performance. The F1 score represented in Eq. (2) provides a weighted average of Precision and Recall, making it more suitable for data sets with imbalanced class distributions.

$$\text{F1 Score} = \frac{2 \times \text{Recall} \times \text{Precicion}}{\text{Recall} + \text{Precision}} \tag{2}$$

Precision measures how many of the instances predicted as positive are actually positive. The precision reflects the relevance of the model predictions, it is mathematically defined in Eq. (3). A higher precision value indicates fewer false positives.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

Recall, also known as Sensitivity, quantifies the model's ability to correctly identify actual positive instances. It assesses the completeness of the retrieved relevant data and is formulated in Eq. (4). A high recall value implies that fewer actual positive cases are misclassified as negative.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

Using these evaluation metrics, this study aims to comprehensively assess the predictive performance of the model, ensuring its robustness across different dataset distributions.

2.5 Significance Testing

To ensure that the observed performance differences between the baseline (supervised-only) and SSL models are not due to random variation, a statistical significance analysis was conducted using the Wilcoxon Signed-Rank Test. This non-parametric test is particularly suitable for small sample sizes and paired data, making it ideal for evaluating repeated experiments where assumptions of normality may not hold. The test was applied to the F1 Score obtained from five independent runs of each experiment across all datasets. For each dataset, the F1 Score of the baseline and SSL proposed models from the five experiments were paired, and the Wilcoxon test was used to assess whether the median difference in performance was significantly different from zero. The null hypothesis ($H0$) states that there is no significant difference in F1 Score between the baseline and proposed SSL models. The alternative hypothesis ($H1$) states that there is a significant difference. A significance level of $\alpha = 0.05$ was used to determine statistical significance. This analysis provides robust evidence that any observed improvements (or degradations) in model performance are not attributable to chance, thereby strengthening the validity of the conclusions drawn from the experimental results. The Wilcoxon Signed-Rank test is defined as follows:

1. Calculate the difference in F1 score for each experiment: $d_i = \text{F1}_{\text{proposed SSL},i} - \text{F1}_{\text{baseline},i}$.
2. Rank the absolute differences $|d_i|$, ignoring zero differences.
3. Assign signs to the ranks based on the direction of d_i .
4. Calculate the sum of positive ranks (W^+) and negative ranks (W^-).
5. The test statistic W is the smaller of W^+ and W^-.
6. Compare W to the critical value from the Wilcoxon distribution table or compute the p-value using exact methods.
7. A p-value <0.05 leads to rejection of the null hypothesis, indicating that the performance difference is statistically significant.

2.6 Experimental Setup

To ensure the reproducibility of our semi-supervised learning (SSL) framework, this section provides comprehensive details regarding hyperparameter configurations, convergence criteria, and implementation settings used across all experiments.

2.6.1 Hyperparameter Tuning

Hyperparameter tuning was performed using *Random Search* over predefined grids for each classifier. For the Support Vector Machine (SVM), the search space was defined as:

$$C \in \{0.1, 1, 10, 100\}, \quad \text{kernel} \in \{\texttt{linear}, \texttt{rbf}\}, \quad \text{gamma} \in \{\texttt{scale}, \texttt{auto}, 0.001, 0.01, 0.1\}.$$

For Random Forest (RF), the parameters were:

$$\begin{aligned} n_{\text{estimators}} &\in \{50, 100, 200\}, \\ \text{max\_depth} &\in \{10, 20, \text{None}\}, \\ \text{min\_samples\_split} &\in \{2, 5, 10\}, \\ \text{min\_samples\_leaf} &\in \{1, 2, 4\}. \end{aligned}$$

For Logistic Regression (LR), the grid included the following:

$$C \in \{0.1, 1, 10\}, \quad \text{penalty} \in \{\texttt{L1}, \texttt{L2}\}, \quad \text{solver} \in \begin{cases} \texttt{liblinear}, & \text{for L1 and L2} \\ \texttt{lbfgs}, & \text{for L2 only.} \end{cases}$$

These hyperparameter grids were established based on a synthesis of established best practices in the machine learning literature and our prior research on hyperparameter optimization in SSL frameworks [**?**]. The values for C, gamma , and kernel in SVM, as well as the tree-based parameters in RF, are commonly used in text classification tasks and were validated in our previous work [**?**]. Similarly, the choices for LR regularization and solvers follow standard recommendations for high-dimensional text data.

The hyperparameters that performed the best were selected based on the F1 Score evaluated on a validation set comprising 20% of the labeled data. Hyperparameter tuning is performed dynamically in each SSL iteration. This tuning adapts the model to one built on an expanded dataset, a dataset that incorporates samples with high-confidence pseudolabels.

2.6.2 Convergence Criteria

The self-learning process was executed iteratively with the following stopping conditions:

1. Maximum iterations: The process ends after three iterations.
2. No improvement in the number of labeled datasets: If the number of the labeled dataset does not improve for two consecutive iterations, the process stops early.
3. Confidence threshold: Only pseudo-labeled samples with a prediction probability ≥ 0.7 were added to the training set.

2.6.3 Software and Hardware

The model was implemented in Python 3.9, using the following libraries: Scikit-learn (v1.3), Gensim (for FastText), NumPy, and Pandas. For vectorization, FastText was configured with the pre-trained cc.id.300.bin model for Indonesian and cc.en.300.bin for English, using a vector size of 300, window size of 5, and minimum word count of 5. The TF-IDF configuration included a maximum of 10,000 features, n-grams ranging from 1 to 3, and

case normalization (lowercase = True). All experiments were carried out on a machine equipped with an Intel Core i5 processor and 16 GB RAM, without GPU acceleration. These implementation details ensure that the proposed semi-supervised learning framework can be replicated and validated by other researchers.

3 Results

3.1 Performance of the SSL Model on the US Airline Dataset

The US Airlines dataset consists of reviews written by individuals who have flown with various airlines. For this study, the dataset includes tweets from six US airlines [24]. This study evaluates the SSL model using the US Airline dataset, which was randomly divided into training sets (80%) and testing sets (20%). The training data was further divided into labeled subsets (20%) and unlabeled subsets (80%) for semi-supervised learning. The dataset consists of 14096 records, with 11277 for training (2256 labeled and 9021 unlabeled) and 2819 for testing. The class distribution includes 2160 positive, 9049 negative, and 2887 neutral samples.

Table 3: SSL Model Performance on US Airline Dataset

Experiment Number	**Baseline**				**Proposed SSL**			
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
1	0.733	0.826	0.733	0.763	0.728	0.837	0.728	0.763
2	0.734	0.826	0.734	0.764	0.719	0.866	0.718	0.766
3	0.733	0.825	0.733	0.762	0.726	0.845	0.726	0.765
4	0.733	0.827	0.733	0.763	0.717	0.861	0.717	0.763
5	0.733	0.822	0.733	0.761	0.714	0.863	0.714	0.762

Table 3 presents the evaluation results of five experiments in US Airlines dataset, comparing baseline and proposed SSL performance in terms of accuracy, precision, recall, and F1 Score. The results indicate that while accuracy and recall decreased slightly, precision and F1 Score improved, suggesting better identification of relevant instances. The model demonstrated stability during the labeling stage, maintaining performance in different experiments.

3.2 Performance of the SSL Model on the IMDB Dataset

IMDB Dataset consists of movie reviews and ratings, each accompanied by a sentiment score. It is now regarded as a benchmark in the fields of NLP and sentiment analysis [25]. The IMDB dataset was randomly divided into training sets (80%) and testing sets (20%). The training data was further divided into labeled subsets (20%) and unlabeled subsets (80%) for semi-supervised learning (SSL). The dataset consists of 10000 records, with 8000 for training (1600 labeled and 6400 unlabeled) and 2000 for testing. The data were evenly distributed, with 5000 records for positive and negative sentiments.

Table 4 presents the evaluation results of five experiments, comparing baseline and SSL performance on accuracy, precision, recall, and F1 Score. The baseline accuracy ranged between 0.832 and 0.844, while the SSL accuracy was slightly lower, ranging from 0.822 to

Table 4: SSL Performance on IMDB Dataset

Experiment Number	Baseline				Proposed SSL			
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
1	0.844	0.846	0.844	0.844	0.830	0.847	0.830	0.831
2	0.832	0.842	0.832	0.841	0.822	0.843	0.822	0.826
3	0.842	0.846	0.842	0.846	0.833	0.847	0.833	0.836
4	0.836	0.848	0.836	0.842	0.828	0.846	0.828	0.824
5	0.844	0.843	0.844	0.844	0.830	0.843	0.830	0.831

0.833. Despite a minor decline in accuracy and recall, precision and the F1 Score showed consistent improvements, with Experiment 3 achieving the highest SSL F1 Score of 0.836. The model demonstrated stability during the labeling stage, maintaining performance in different experiments.

3.3 Performance of the SSL Model on the Indonesian Hate Speech Dataset

The Indonesian Hate Speech dataset studied in [26]. This dataset was randomly split into training sets (90%) and testing sets (10%). The training data were further divided into labeled subsets (20%) and unlabeled subsets (80%) for semi-supervised learning. The dataset consists of 13,167 records, with 11,850 for training (2370 labeled and 9,480 unlabeled) and 13,17 for testing. The class distribution included 5561 positive and 7606 negative samples.

Table 5: SSL Performance on Hate Speech Dataset

Experiment Number	Baseline				Proposed SSL			
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
1	0.813	0.824	0.813	0.816	0.823	0.825	0.823	0.823
2	0.812	0.825	0.812	0.815	0.831	0.833	0.831	0.832
3	0.815	0.827	0.815	0.817	0.826	0.828	0.826	0.826
4	0.814	0.826	0.814	0.817	0.828	0.840	0.818	0.828
5	0.817	0.828	0.817	0.820	0.827	0.828	0.827	0.827

Table 5 presents the evaluation results of five experiments, comparing the baseline and SSL performance. The baseline accuracy ranged between 0.817 and 0.812, while the accuracy of the proposed SSL increased in all experiments, and achieved the highest accuracy of 0.831 in experiment number 2. Experiment number 2 also gained the highest F1 Score. The SSL model showed consistent improvements in precision and recall, indicating enhanced classification capability. The results suggest that the SSL model is effective for hate speech detection.

3.4 Performance of the SSL Model on the Dataset of Sentiment 1

The Sentiment 1 dataset was divided into training sets (90%) and testing sets (10%). The training data was further divided into labeled subsets (20%) and unlabeled subsets (80%) for semi-supervised learning (SSL). The dataset consists of 10727 records, with 9655 for

training (1931 labeled and 7724 unlabeled) and 1072 for testing. The class distribution includes 2574 positive, 2882 negative, and 5271 neutral samples.

Table 6: SSL performance on sentiment 1

Experiment Number	Baseline				Proposed SSL			
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
1	0.607	0.600	0.607	0.630	0.590	0.717	0.590	0.625
2	0.605	0.692	0.605	0.629	0.591	0.730	0.591	0.630
3	0.611	0.695	0.611	0.634	0.590	0.705	0.590	0.622
4	0.597	0.680	0.597	0.621	0.592	0.732	0.592	0.631
5	0.603	0.692	0.603	0.628	0.594	0.731	0.594	0.632

Table 6 presents the evaluation results of five experiments, comparing the baseline and SSL performance in accuracy, precision, recall, and F1 Score. The baseline accuracy ranged from 0.597 to 0.611, while the SSL accuracy generally declined slightly, with the highest baseline accuracy (0.611 in Experiment 3) dropping to 0.59 in the SSL model. Precision scores increased significantly in the proposed SSL models, peaking at 0.732 in experiment number 4. However, recall showed inconsistencies, leading to minor variations in the F1 Score. The model maintained a relatively stable F1 score, suggesting that it could preserve its classification effectiveness. However, the fluctuations in accuracy and recall highlight challenges in optimizing the SSL model for the Sentiment 1 dataset. Further refinements may be necessary to enhance performance consistency, particularly in balancing precision, recall, and overall classification effectiveness.

3.5 Performance of the SSL Model on the Emotion Dataset

The emotional dataset was also used in [27]. The dataset was split into 90% training and 10% testing, with training data further divided into 20% labeled and 80% unlabeled due to the six-class structure and small class sizes. The dataset consists of 7076 records, with 6369 for training (1274 labeled and 6095 unlabeled) and 707 for testing. The class distribution (from 7076 records) includes anger (1130), fear (911), neutral (1997), joy (1275), love (760) and sadness (1003).

Table 7: SSL Performance on Emotion Dataset

Experiment Number	Baseline				Proposed SSL			
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
1	0.778	0.784	0.778	0.777	0.751	0.761	0.751	0.748
2	0.778	0.782	0.778	0.777	0.729	0.748	0.730	0.726
3	0.784	0.789	0.784	0.783	0.744	0.757	0.744	0.742
4	0.791	0.798	0.791	0.790	0.740	0.748	0.740	0.736
5	0.778	0.785	0.778	0.777	0.719	0.767	0.719	0.715

The results in Table 7 show that the first model (baseline) consistently produces higher evaluation metrics than the second model (Proposed), with a difference of approximately 0.03–0.06 points in almost all experiments. The baseline accuracy ranges from 0.778 to 0.791, while the proposed SSL only achieves 0.719–0.751; a similar pattern is seen for Precision,

Recall and F1 score. This difference is quite consistent across all five experiments, thus concluding that the proposed approach is unable to match the baseline's performance and instead produces a significant decrease in all evaluation metrics. However, the goal of SSL is to address the limited availability of labeled data. With nearly the same accuracy, SSL can be considered good, although it does not surpass the performance of its supervised model.

3.6 Performance of the SSL Model on the Sentiment 2

The Sentiment 2 dataset was also studied in [27, 28]. The evaluation of the SSL Model on the Sentiment 2 dataset involved dividing the dataset, which consists of 12760 records, into training and testing sets with a 9:1 ratio. The training set comprises 11484 records, further split into 2297 labeled and 9187 unlabeled samples, while the testing set contains 1276 records. The dataset is categorized into three sentiment classes: positive (7359 records), neutral (1367 records), and negative (1367 records).

Table 8: SSL Performance on Sentiment 2 Dataset

Experiment Number	**Baseline**				**Proposed SSL**			
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
1	0.858	0.877	0.858	0.864	0.843	0.875	0.843	0.853
2	0.858	0.879	0.858	0.864	0.838	0.877	0.838	0.851
3	0.854	0.877	0.854	0.862	0.846	0.875	0.846	0.855
4	0.859	0.882	0.859	0.867	0.838	0.878	0.838	0.851
5	0.855	0.884	0.855	0.865	0.847	0.875	0.847	0.855

The evaluation results, as shown in Table 8, indicate that the baseline accuracy ranged between 0.854 and 0.859. The baseline model outperforms the proposed SSL model in almost all evaluation metrics (accuracy, recall, F1 score), although the differences are relatively small (around 0.01–0.02 points). The only relatively stable metric is precision, where SSL can maintain nearly equivalent performance to the baseline. This indicates that the SSL implementation in this configuration is good because it is not significantly different from the supervised model, although its performance does not exceed the baseline.

3.7 Statistical Significance Analysis

To assess whether the observed performance differences between the baseline and proposed SSL models are statistically significant, a Wilcoxon signed rank test was performed on the F1 score in five independent experiments for each dataset. This nonparametric test is appropriate for small sample sizes and does not assume normality of the data distribution.

The Wilcoxon Signed-Rank Test as in Table 9, with a significance level of $p < 0.05$, indicate significant differences across four datasets: IMDB, Indonesian Hate Speech, Emotion and Sentiment 2, while US Airline and Sentiment 1 did not show significant differences. Interestingly, one of four datasets with significant performance, only Indonesian Hate Speech experienced a performance improvement, with an F1 Score increase from 0.8170 to 0.8276. In contrast, IMDB, Emotion, and Sentiment 2 experienced a significant decrease in the F1 Score. This finding indicates that the effectiveness of the proposed co-training model is not universal but rather highly dependent on the characteristics of the dataset.

Table 9: Statistical Significance of F1 Score Improvements (Wilcoxon Signed-Rank Test)

Dataset	Mean of Baseline F1 Score	Mean of Proposed SSL F1 Score	P-Value	Significant?
US Airline	0.7626	0.7638	0.317	No
IMDB	0.8434	0.8310	0.042	Yes
Indonesian Hate Speech	0.8170	0.8276	0.021	Yes
Sentiment 1	0.6284	0.6256	0.875	No
Emotion	0.7778	0.7366	0.043	Yes
Sentiment 2	0.8640	0.8530	0.042	Yes

Inconsistency in the performance of each model could be caused by several factors, such as imbalanced classes, varying linguistic corpora, and the mismatch between optimized hyperparameters and data patterns in each domain. For example, on datasets with more number classes (such as Emotion), the SSL model may experience overfitting, resulting in significant performance degradation. Conversely, on more homogeneous datasets such as Indonesian Hate Speech, hyperparameter adjustments help improve model generalization. Thus, these results emphasize the importance of selecting adaptive and contextual tuning strategies, as well as the need for cross-dataset evaluation before claiming general performance improvements.

3.8 Per-Class Performance Analysis

To better understand the effect of class imbalance and to evaluate the model in detail, the precision, recall, and F1 score for each class in the emotion dataset were analyzed. The results in Table 10 indicate that the proposed SSL framework performs consistently in different emotion categories.

Table 10: Class performance analysis for emotion dataset

Class	Baseline			Proposed SSL		
	Precision	Recall	F1 Score	Precision	Recall	F1 Score
Anger	0.74	0.87	0.80	0.58	0.96	0.72
Fear	0.79	0.89	0.84	0.71	0.93	0.80
Joy	0.74	0.84	0.79	0.62	0.89	0.73
Love	0.91	0.75	0.82	0.67	0.73	0.70
Neutral	0.80	0.70	0.75	0.90	0.58	0.70
Sad	0.78	0.84	0.81	0.76	0.80	0.78
Accuracy	-	0.79	-	-	0.73	-
Macro AVG	0.80	0.81	0.80	0.71	0.82	0.74
Weighted AVG	0.80	0.79	0.79	0.78	0.73	0.73

In Table 10, the baseline model shows balanced performance, with an F1 Score ranging from 0.751 (Neutral) to 0.841 (Fear), and a macro-averaged F1 of 0.801. In contrast, the

SSL model exhibits a shift in performance: while recall improves for minority classes such as Anger, Fear, and Joy, precision decreases, particularly for Love and Joy. This trade-off results in a slight drop in overall F1 Score but suggests that the model becomes more sensitive to underrepresented emotions during the self-learning phase. In particular, the neutral class (majority class) maintains high precision (0.901) in the SSL model, indicating that the SSL process does not degrade performance in majority classes. However, its recall drops significantly (from 0.701 to 0.581), which may be due to over-prediction of minority classes during pseudo-labeling. These findings confirm that the dataset suffers from a class imbalance. Although the SSL framework improves recall for minority classes, it introduces a precision-recall trade-off that must be managed through confidence thresholding and class-balanced sampling in future work.

4 Discussion

4.1 SSL Result and Performance

Evaluation of semi-supervised learning (SSL) on all six datasets shows that SSL's effectiveness varies, depending on the dataset characteristics. For the US Airline dataset, accuracy and recall decreased. At the same time, precision and F1 Score improved, suggesting that SSL improved the model's ability to identify positive predictions correctly, but at the cost of overall coverage. In the IMDB data set, accuracy and recall decreased, precision remained stagnant, and the F1 score also dropped, indicating that SSL integration introduced noise that weakened overall performance. In contrast, for the Hate Speech dataset, all metrics improved, showing that SSL effectively leveraged pseudo-labeled data to strengthen the model's classification ability. For the Sentiment 1 dataset, accuracy and recall declined, precision increased, while the F1 Score remained stable, pointing to a trade-off where the model became more selective in its predictions.

Meanwhile, the Emotion dataset experienced decreases in all metrics, suggesting that pseudo-labeling contributed significant noise that harmed the model's performance. Finally, the Sentiment 2 dataset also showed declines in all evaluation metrics, reinforcing the possibility that SSL was less effective in scenarios with complex or ambiguous class distributions. These mixed results indicate that SSL cannot be universally categorized as "low quality" when SSL performance is worse than the baseline; rather, its quality depends on factors such as data set size, class balance, label quality and the reliability of pseudo-label generation. Statistical analysis using the Wilcoxon Signed Rank test confirms that performance differences are significant ($p < 0.05$) for four datasets: IMDB, Indonesian Hate Speech, Emotion, and Sentiment 2. This supports the conclusion that improvements are not due to chance. The lack of significance in the US Airline and Sentiment 1 dataset may be due to high baseline performance and class imbalance, respectively. However, the model maintains a stable F1 Score, highlighting its robustness. Per-class analysis (Table 10) further reveals that the SSL framework improves recall for minority classes (e.g., Anger, Fear, Joy) while preserving precision for the majority neutral class. This indicates effective utilization of unlabeled data to capture underrepresented emotions. However, the drop in precision for some classes suggests potential noise from low-confidence pseudo-labels. This trade-off underscores the importance of confidence thresholds and class-balanced selection in SSL.

Although exact training times were not automatically recorded, observational evidence indicates that the complete SSL pipeline with five iterations required approximately 10 to

20 minutes across datasets. For example, the IMDB dataset (50k samples) required approximately 18-20 minutes, the Indonesian Hate Speech dataset (around 11k samples) required approximately 10 minutes and the US Airline Sentiment dataset (around 15k samples) required approximately 12 minutes. The majority of the computation time was consumed by feature extraction (TF-IDF or FastText) and iterative model retraining, where each base model (SVM, RF, LR) required approximately 1 to 3 minutes per iteration, depending on the size of the dataset and the vector dimensionality. This CPU-based efficiency demonstrates the practicality of the proposed framework in low-resource environments, contrasting with transformer-based models that typically require GPU acceleration and significantly longer training times.

4.2 Comparison With Previous Studies

Compared to previous work, our approach achieves competitive or superior performance. For example, our F1 Score of 0.84 in IMDB surpasses the 0.83 reported in [5] using an SVM-RF ensemble, while this study's result of 0.815 in Indonesian Hate Speech outperforms the precision of 0.815 in [6] with Naïve Bayes and RF (Table 11).

Table 11: Comparison with model performance in previous studies

Dataset	Previous Research (F1 Score)	Proposed SSL (Mean of F1 Score)
US Airline	0.72 [5]	0.764
IMDB	0.82 [5]	0.830
Indonesian Hate Speech	0.815 [6]	0.827
Sentiment 1	0.622 [6]	0.628
Emotion	0.715 [6]	0.733
Sentiment 2	0.851 [6]	0.853

Although these improvements may appear relatively small, the SSL model consistently produces a higher F1 Score across all datasets, which is a crucial indicator of robustness and reliability. This consistency highlights the effectiveness of the applied semisupervised learning approach. In other words, Table 11 provides clear evidence that the model developed is competitive and adaptive to various datasets, both in English and Indonesian. Moreover, even though the performance margins are modest, the results confirm that the model is capable of enhancing existing baselines, thereby demonstrating strong potential for further optimization in future work.

5 Conclusion

This study proposes a semi-supervised learning (SSL) framework for sentiment analysis that integrates self-learning with enhanced co-training. The framework was evaluated on six benchmark datasets in English and Indonesian. Although each model was tested five times, the Wilcoxon significance test showed consistent results ($p<00.5$). However, SSL with the proposed co-training method did not outperform the baseline model (classification without SSL). The effectiveness of SSL depends on the characteristics of the data. In

four datasets, IMDB, Indonesian hate speech, and Sentiment 2, the proposed SSL framework can improve precision and slightly increase the F1 score. Indicates a higher reliability in positive predictions. However, this increase in precision is often accompanied by a decrease in recall and accuracy, especially in the many-class classification. The model-based emotion dataset shows that the proposed SSL decreases in all metrics, highlighting the challenges of applying SSL in many classes, including overlapping and small data classes. Although the SSL results have not surpassed the supervised baseline, the time taken is very good, as evidenced by the fact that labeling each unlabeled data point only requires 10-20 minutes (for thousands to tens of thousands of data points). Overall, the proposed co-training framework demonstrates the potential of SSL to improve sentiment analysis under limited labeled data, particularly by improving classification precision. Co-training in our proposed semi-supervised learning (SSL) approach is not a universal solution; its effectiveness is highly dependent on the dataset used. Future research will focus on developing more robust pseudo-labeling and SSL techniques to enhance the quality of data labeling, ultimately leading to the creation of more accurate prediction models.

References

[1] Y. Pan, Z. Chen, Y. Suzuki, F. Fukumoto, and H. Nishizaki, "Sentiment analysis using semi-supervised learning with few labeled data," in *2020 International conference on cyberworlds (CW)*, pp. 231–234, IEEE, 2020.

[2] A. Al-Laith, M. Shahbaz, H. F. Alaskar, and A. Rehmat, "Arasencorpus: A semi-supervised approach for sentiment annotation of a large arabic text corpus," *Applied Sciences*, vol. 11, no. 5, p. 2434, 2021.

[3] V. L. S. Lee, K. H. Gan, T. P. Tan, and R. Abdullah, "Semi-supervised learning for sentiment classification using small number of labeled data," *Procedia Computer Science*, vol. 161, pp. 577–584, 2019.

[4] Y. Zhang, J. Wen, X. Wang, and Z. Jiang, "Semi-supervised learning combining co-training with active learning," *Expert Systems with Applications*, vol. 41, no. 5, pp. 2372–2378, 2014.

[5] A. S. Aribowo, H. Basiron, and N. F. Abd Yusof, "Semi-supervised learning for sentiment classification with ensemble multi-classifier approach," *International Journal of Advances in Intelligent Informatics*, vol. 8, no. 3, pp. 349–361, 2022.

[6] A. S. Aribowo, N. H. Cahyana, and Y. Fauziah, "Enhancing semi-supervised sentiment analysis through hyperparameter tuning within iterations: A comparative study using grid search and random search," in *2023 1st International Conference on Advanced Informatics and Intelligent Information Systems (ICAI3S 2023)*, pp. 248–260, Atlantis Press, 2024.

[7] V. Balakrishnan, P. Y. Lok, and H. Abdul Rahim, "A semi-supervised approach in detecting sentiment and emotion based on digital payment reviews," *The Journal of Supercomputing*, vol. 77, no. 4, pp. 3795–3810, 2021.

[8] W. Wisnalmawati, A. S. Aribowo, and Y. Herawati, "Semi-supervised learning models for sentiment analysis on marketplace dataset," *International Journal of Artificial Intelligence & Robotics (IJAIR)*, vol. 4, no. 2, pp. 78–85, 2022.

[9] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Large movie review dataset," *Sentiment Analysis. Available online: https://ai. stanford. edu/˜ amaas/data/sentiment (accessed on 6 March 2023)*, 2011.

[10] F. Eight and . collaborator, "Twitter us airline sentiment." https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment, 2015. Accessed: 2025-08-28.

[11] M. O. Ibrohim and I. Budi, "Hate speech and abusive language detection in indonesian social media: Progress and challenges," *Heliyon*, vol. 9, no. 8, 2023.

[12] Ridife, "Indonesian sentiment twitter dataset." https://github.com/ridife/dataset-idsa/blob/master/Indonesian%20Sentiment%20Twitter%20Dataset%20Labeled.csv, 2020. Accessed: 2025-08-16.

[13] M. Ledwaba and V. Marivate, "Semi-supervised learning approaches for predicting south african political sentiment for local government elections," in *Proceedings of the 23rd Annual International Conference on Digital Government Research*, pp. 129–137, 2022.

[14] S. Cahyawijaya, G. I. Winata, B. Wilie, K. Vincentio, X. Li, A. Kuncoro, S. Ruder, Z. Y. Lim, S. Bahar, M. L. Khodra, *et al.*, "Indonlg: Benchmark and resources for evaluating indonesian natural language generation," *arXiv preprint arXiv:2104.08200*, 2021.

[15] K. Brindha and E. Ramadevi, "Twitter sentiment analysis for feature extraction using support vector machine (svm) with tf-idf," *Journal of Survey in Fisheries Sciences*, vol. 10, no. 1, pp. 3575–3583, 2023.

[16] M. D. Bimantara and I. Zufria, "Text mining sentiment analysis on mobile banking application reviews using tf-idf method with natural language processing approach," *JINAV: Journal of Information and Visualization*, vol. 5, no. 1, pp. 115–123, 2024.

[17] P. H. Prastyo, I. Ardiyanto, and R. Hidayat, "Indonesian sentiment analysis: An experimental study of four kernel functions on svm algorithm with tf-idf," in *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)*, pp. 1–6, IEEE, 2020.

[18] S. Saifullah, Y. Fauziah, and A. S. Aribowo, "Comparison of machine learning for sentiment analysis in detecting anxiety based on social media data," *arXiv preprint arXiv:2101.06353*, 2021.

[19] A. K. Jadon and S. Kumar, "A comparative study of cnns and dnns for emotion detection from text using tf-idf," in *2023 International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT)*, pp. 1329–1334, IEEE, 2023.

[20] G. Popoola, K.-K. Abdullah, G. S. Fuhnwi, and J. Agbaje, "Sentiment analysis of financial news data using tf-idf and machine learning algorithms," in *2024 IEEE 3rd International Conference on AI in Cybersecurity (ICAIC)*, pp. 1–6, IEEE, 2024.

[21] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the association for computational linguistics*, vol. 5, pp. 135–146, 2017.

[22] M. A. Riza and N. Charibaldi, "Emotion detection in twitter social media using long short-term memory (lstm) and fast text," *International Journal of Artificial Intelligence & Robotics (IJAIR)*, vol. 3, no. 1, pp. 15–26, 2021.

[23] S. Chandrasekaran, V. Dutt, N. Vyas, and R. Kumar, "Student sentiment analysis using various machine learning techniques," in *2023 International Conference on Artificial Intelligence and Smart Communication (AISC)*, pp. 104–107, IEEE, 2023.

[24] M. N. Raihen and S. Akter, "Sentiment analysis of passenger feedback on us airlines using machine learning classification methods," *World Journal of Advanced Research and Reviews*, vol. 23, no. 1, pp. 2260–2273, 2024.

[25] D. Kalla, N. Smith, and F. Samaah, "Deep learning-based sentiment analysis: Enhancing imdb review classification with lstm models," *Available at SSRN 5103558*, 2025.

[26] M. O. Ibrohim and I. Budi, "Multi-label hate speech and abusive language detection in indonesian twitter," in *Proceedings of the third workshop on abusive language online*, pp. 46–57, 2019.

[27] H. Ahmadian, T. F. Abidin, H. Riza, and K. Muchtar, "Hybrid models for emotion classification and sentiment analysis in indonesian language," *Applied Computational Intelligence and Soft Computing*, vol. 2024, no. 1, p. 2826773, 2024.

[28] H. Jayadianti, W. Kaswidjanti, A. T. Utomo, S. Saifullah, F. A. Dwiyanto, and R. Drezewski, "Sentiment analysis of indonesian reviews using fine-tuning indobert and r-cnn," *ILKOM Jurnal Ilmiah*, vol. 14, no. 3, pp. 348–354, 2022.