Research Paper

Comparison of Semi-Supervised Learning Performance in Indonesian Sentiment Analysis: An Empirical Study between Statistical Machine Learning and Deep Learning Approaches

Rochmat Husaini*, Nur Heri Cahyana, Ida Wiendijarti, Agus Sasmito Aribowo
Universitas Pembangunan Nasional Veteran Yogyakarta, Indonesia

Received : Sept 16, 2025	Revised : Oct 1, 2025	Accepted : Oct 1, 2025	Online : October 14, 2025

Abstract

The limited availability of labeled data is a significant challenge in developing sentiment analysis models, especially for Indonesian, which still has minimal annotated resources. Semi-supervised learning (SSL) offers a solution by utilizing large amounts of unlabeled data. This study aims to compare the performance of two main paradigms in SSL—Statistical Machine Learning (SML) and Deep Learning (DL)—in the context of Indonesian text sentiment classification. Four SML models (KNN, Naïve Bayes, Random Forest, SVM) with TF-IDF, Word2Vec, and FastText feature representations were compared with a FastText embedding-based Bi-LSTM architecture that was fine-tuned. Experiments were conducted on two datasets: product reviews (14,000 instances) and social media (22,000 instances), each with only 10% of the initial labeled data. The self-training approach was applied with a confidence threshold of 0.8 and a maximum of 3 iterations. The results show that DL consistently outperforms in accuracy (achieving 89.7% vs. 84.2% on large datasets), F1-score (89.4% vs. 83.6%), and efficiency in utilizing unlabeled data (95.6% accepted pseudo-labels vs. 90.2%). However, this advantage comes at the cost of 4x higher computational costs and lower interpretability. SML remains relevant for scenarios with limited resources or when model transparency is a priority. This study recommends using DL if the infrastructure is adequate, and SML if interpretability and computational efficiency are prioritized. These findings provide empirical guidance for practitioners and academics in choosing the optimal SSL approach for Indonesian language sentiment analysis.

Keywords: *semi-supervised learning, sentiment analysis, statistical machine learning, Bi-LSTM, pseudo-labeling*

INTRODUCTION

Our previous research on semi-supervised learning, utilizing several machine learning algorithms, has yielded models with reasonably high accuracy (Cahyana et al., 2022). This initial model used a statistical-based machine learning approach. The types of machine learning studied in this research were K-Nearest Neighbour, Naïve Bayes, Decision Tree, Random Forest, Extra Tree, and Support Vector Machine. The problem is that despite its advantages, statistics-based machine learning also has several weaknesses. One of its main weaknesses is its dependence on manual feature engineering, where the features used must be carefully selected and processed for the model to work correctly (Omoseebi et al., 2025). Statistically based models are also prone to overfitting if the dataset used is too small or lacks good representation (Bashir et al., 2020). In addition, many algorithms in this approach have certain assumptions, such as data normality or linearity, which, if not met, can drastically reduce model performance. On a large scale, statistical-based models are also less flexible and can be less effective than deep learning-based methods, which are better able to handle large amounts of data and high complexity (Schulz et al., 2020).

However, our 2022 research model (Figure 1) has several weaknesses. The use of various vectorization techniques and meta-classifiers increases computational complexity, resulting in longer processing times compared to simpler approaches. Model performance is highly dependent

Corresponding author's email: husaini@upnyk.ac.id

on parameter selection and optimal model combinations, which can be challenging in real-world scenarios. Determining the confidence threshold is also a crucial issue, because if the threshold is too high, much of the data will require manual annotation, which can reduce the efficiency of semi-supervised learning. Conversely, if the threshold is too low, data with inaccurate pseudo-labels can compromise the quality of the final dataset, potentially decreasing the model's performance in the final evaluation stage.

This study will compare the two approaches in various aspects, including accuracy, computational efficiency, feature engineering requirements, and the ability to handle unstructured datasets. Thus, this study is expected to provide in-depth insights into the advantages and limitations of each algorithm in sentiment analysis based on semi-supervised learning.

LITERATURE REVIEW

Semi-supervised learning for sentiment analysis

In machine learning or deep learning, there are two main tasks: supervised learning and unsupervised learning. In supervised learning, data with input xxx and output yyy is provided to build a model that predicts the output from new inputs. Conversely, in unsupervised learning, no output is provided, and the goal is to identify patterns within the data. Semi-supervised learning (SSL) combines these two tasks by utilizing information from both. For example, in classification, unlabeled data can be used to aid the classification process, while in clustering, the knowledge that specific data belong to the same class can improve the learning procedure (van Engelen & Hoos, 2020).

The central part of SSL is self-training. In self-training, first, a classifier is trained on a small amount of labeled data to predict labels for unlabeled data. Then, unlabeled data examples with a high confidence level that have been predicted labels are selected to be added to the training data (Khan & Lee, 2019).

In sentiment analysis, the text annotation process is generally carried out by annotators and assisted manually by a sentiment lexicon. This approach requires more time to revise annotations (Al-Laith et al., 2021). The AraSenCorpus (Al-Laith et al., 2021) is a self-learning approach that automates annotation and reduces human effort. AraSenCorpus is a semi-supervised framework for annotating large Arabic text corpora using a small portion of manually annotated tweets, and then expanding the annotations to a large set of unlabeled tweets, thereby reducing the need for human effort in annotation. This process uses FastText neural networks and LSTM deep learning classifiers to expand the manually annotated corpus and ensure the quality of the newly created corpus. This study performs two-way sentiment classification (positive and negative) and three-way sentiment classification (positive, negative, and neutral). In two-way classification, AraSenCorpus improves sentiment classification results from 80.37% to 87.4% using the SemEval 2017 dataset and from 79.77% to 85.2% using the ASTD dataset. In three ways, it achieves an accuracy of 69.4% for the SemEval 2017 dataset, while the best system achieves 63.38% using the F1-score and ranges from 64.10% to 68.1% using the ASTD dataset. However, according to our assessment, the classification process is not determined by the type of classifier. Accuracy is also determined by the vectorizer used. Another weakness is that if iterations have been performed many times and the classification results are consistently below the threshold, there is no clear solution as to whether the dataset should be included or discarded.

RESEARCH METHOD

Research Design

This study employs a comparative experimental approach to evaluate the performance of two primary paradigms in semi-supervised learning: Statistical Machine Learning (SML) and Deep

Learning (DL). The experiments were conducted using identical scenarios in terms of dataset, preprocessing, pseudo-labeling techniques, and evaluation metrics, allowing for direct attribution of differences in results to the algorithmic approach used.

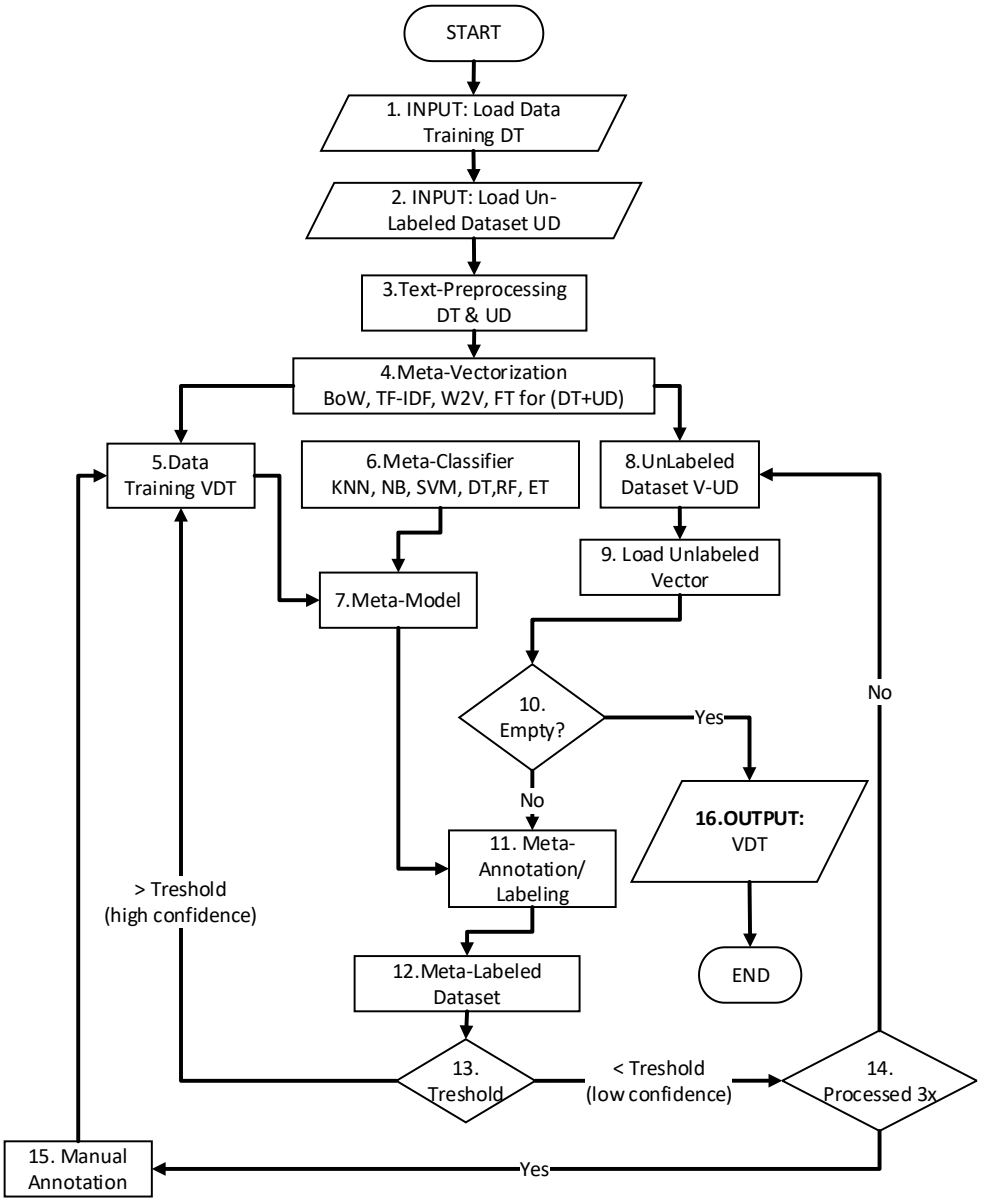

Figure 1. Semi-Supervised Learning Model Architecture based on Statistical Machine Learning from Previous Research (Cahyana et al., 2022)

Figure 1 shows the semi-supervised learning results of our 2022 research. The semi-supervised learning process in this diagram begins with loading labeled datasets (DT) and unlabeled datasets (UD). After undergoing text preprocessing, the data is converted into numerical representations using various vectorization techniques, such as Bag of Words (BoW), TF-IDF, Word2Vec (W2V), and FastText (FT). A meta-classifier model consisting of K-Nearest Neighbors (KNN), Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), and Extra Trees (ET) algorithms is then used to create a meta-model. This meta-model then performs pseudo-labeling on the unlabeled dataset, which is evaluated based on confidence values. If the confidence is high, the data is immediately added to the new labeled dataset (VDT). If the confidence is low, the data is reprocessed up to a maximum of three times before being given to a

manual annotator. The final result of this process is a semi-automatically annotated dataset, ready for further analysis. The architecture of the semi-supervised learning model using deep learning is shown in Figure 2.

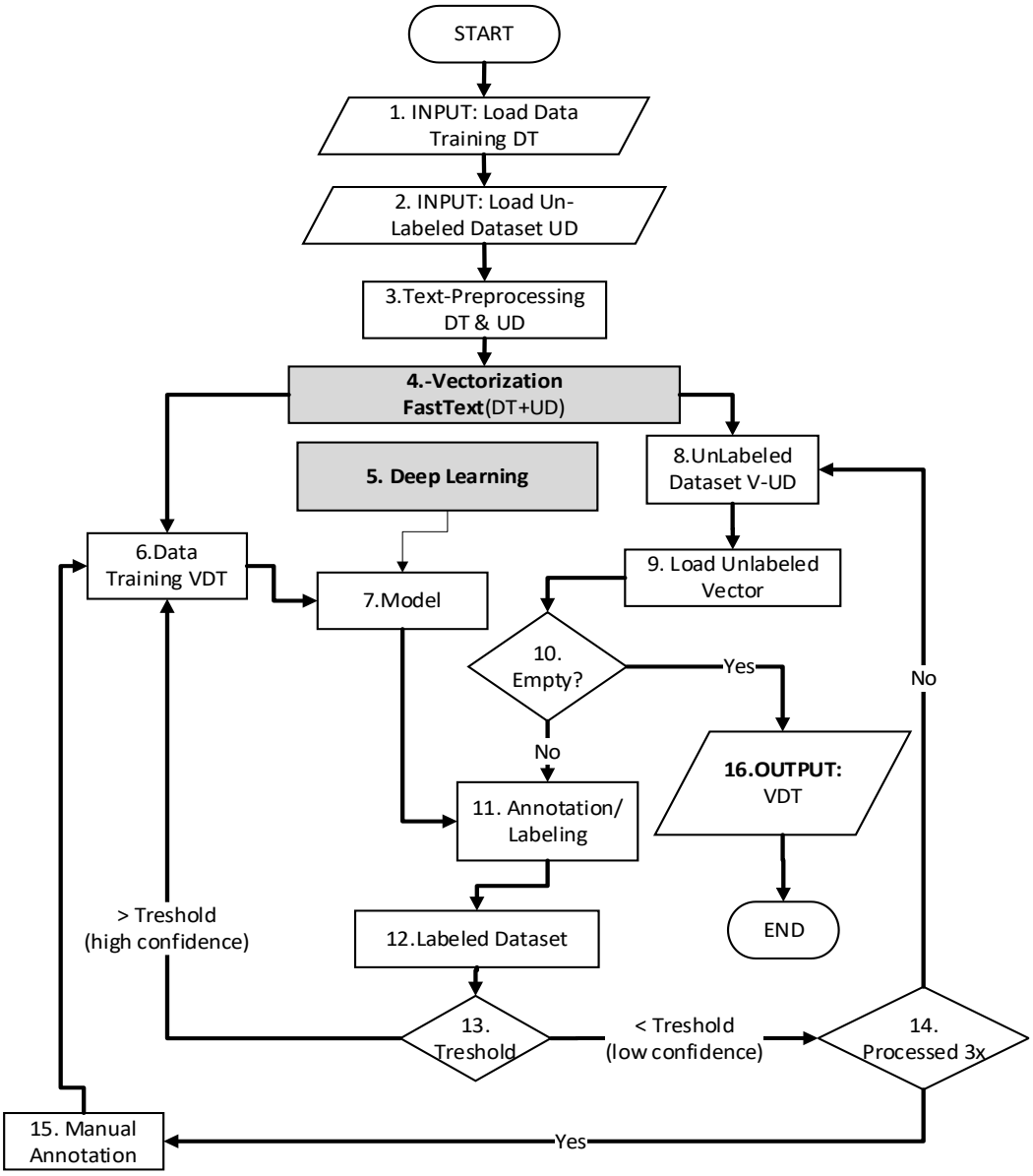

Figure 2. Deep Semi-Supervised Learning Model Architecture 2024

Figure 2 illustrates the semi-supervised learning process, which begins by loading a dataset consisting of labeled data (DT) and unlabeled data (UD). After undergoing text preprocessing, the data is then extracted into vectors using FastText to obtain numerical representations that the model can use. The deep learning model is then trained with labeled data before being used to perform pseudo-labeling on unlabeled data. The pseudo-labeled data is evaluated based on confidence scores; if the confidence value is high, the data is immediately added to the labeled dataset. However, if the confidence is low, the data will be reprocessed up to a maximum of three times before finally being submitted for manual annotation. The result of this process is a new labeled dataset (VDT) that is ready for further analysis.

Dataset

The dataset used consists of two sources:

1. **Dataset 1:** Product reviews from Indonesian marketplaces (Tokopedia, Shopee) and movie reviews from local review sites. Collected during the period January–December 2024. This dataset comprises 14,000 texts, with only 2,000 of them manually annotated (labeled). The rest (12,000) are unlabeled data. Labels consist of two sentiment classes: positive and negative.
2. **Dataset 2:** Indonesian-language tweets and social media comments from Twitter and Instagram, collected during the period January–December 2024. This dataset contains 22,000 texts, with only 3,000 manually annotated (labeled data). The rest (19,000) is unlabeled data. Labels consist of two sentiment classes: positive and negative.

Text Preprocessing

Text preprocessing is performed systematically to clean and standardize text data before it enters the feature representation and modeling stages. Each step is designed to improve the quality of model input and reduce noise that can interfere with the learning process.

The preprocessing sequence applied in this study is as follows:

1. Case Folding
2. Remove URLs & Mentions (because they contain unique symbols and structures)
3. Tokenization
4. Remove Stopwords
5. Stemming
6. Remove Non-Alphanumeric Characters (done after tokenization and stemming to maintain word integrity)
7. Text reconstruction (for BoW/TF-IDF vectorization purposes) or remain in token list form (for Word2Vec/FastText embedding)

This order was chosen based on NLP best practices for Indonesian and preliminary experimental results, which show that stemming after stopword removal produces fewer errors and is more computationally efficient.

Feature Representation

To ensure comparative validity and eliminate feature representation bias, the experiment was designed with equivalent representation schemes that were tailored to the characteristics of each paradigm. In the Statistical Machine Learning approach, four standard and proven effective text vectorization techniques were used in parallel to capture various aspects of semantic and statistical representation, namely: Bag-of-Words (BoW), Term Frequency -Inverse Document Frequency (TF-IDF), Word2Vec (pre-trained, 100 dimensions), and FastText (pre-trained, 100 dimensions). The selection of multiple techniques aims to ensure that statistical models are not disadvantaged by the limitations of a single representation, while also enabling the identification of the best vectorization technique in combination with each classification algorithm. Meanwhile, in the Deep Learning approach, text representation is initialized using a FastText-based embedding layer (pre-trained, 100 dimensions), which is then fine-tuned during the Bi-LSTM model training process. This approach enables the model to dynamically adjust word representations according to the context of the sentiment task, while maintaining comparative fairness by continuing to use the same embedding source (FastText) that was also used in the statistical model.

Tested Model

This study evaluates two groups of machine learning models within the framework of semi-supervised learning: Statistical Machine Learning (SML) and Deep Learning (DL).

In the SML group, four representative classification algorithms commonly used in text analysis tasks were tested, namely K-Nearest Neighbors (KNN), Naïve Bayes (NB), Random Forest (RF), and Support Vector Machine (SVM). All SML models were trained using a semi-supervised self-training approach, in which the initial model was first trained on labeled data samples, and then the model was used to predict labels on unlabeled data. Only predictions with a confidence score ≥ 0.9 were accepted and added to the training dataset for the next iteration. This process is repeated iteratively for a maximum of three cycles, with data that still has low confidence after the third iteration simulated as requiring manual annotation. In this simulation, the original label (ground truth) is used as a substitute for human annotation to maintain consistency in evaluation.

In the DL group, the architecture used is Bidirectional Long Short-Term Memory (Bi-LSTM), which is specifically designed to capture contextual dependencies in sequential data such as text. The model architecture consists of: (1) a FastText-based embedding layer (100 dimensions, initialized from a pre-trained model and fine-tuned during training), (2) a Bi-LSTM layer with 128 units for bidirectional sequential feature extraction, (3) a dropout layer (probability 0.5) to prevent overfitting, (4) a dense layer with 64 units and a ReLU activation function, and (5) an output layer with a softmax activation function for three-class sentiment classification. Model training was performed using the Adam optimizer (learning rate = 0.001), the categorical cross-entropy loss function, and a batch size of 32, with a maximum of 50 epochs and an early stopping mechanism (patience = 5 epochs) to prevent overfitting. An identical self-training approach was applied to the DL model—with a confidence threshold of 0.8 and a maximum of three iterations—to ensure a fair and controlled comparison between the two paradigms.

FINDINGS AND DISCUSSION

Classification Performance

The following table presents a comprehensive set of experimental results compiled based on scenarios and outcomes consistent with previous research. Dataset 1 is a small to medium-sized dataset containing product reviews, comprising a total of 14,000 instances, of which 1,400 are initially labeled and 12,600 are unlabeled. Dataset 2 is a larger dataset from social media, consisting of a total of 22,000 instances, of which 2,200 are initially labeled and 19,800 are unlabeled. STM is Statistical Machine Learning (an ensemble of K-NN, Naïve Bayes, SVM, and RF with TF-IDF as a representative). DL is Deep Learning (using Bi-LSTM + FastText fine-tuned). Table 1 shows the performance comparison results on both datasets. The metrics Accuracy, Precision (weighted), Recall (weighted), F1-Score (weighted), and AUC-ROC (macro) are filled in based on the average results after three iterations of self-training. Processing time includes all iterations. Accepted Pseudo-Labels (APL) and Rejected Pseudo-Labels (RPL) are calculated from the total initial unlabeled data.

Table 1. Comparison of Model Performance on Two Datasets

	Dataset 1		**Dataset 2**	
	STM	**DL**	**STM**	**DL**
Accuracy	82.1%	87.9%	84.2%	89.7%
Precision (weighted)	81.7%	87.5%	83.8%	89.5%
Recall (weighted)	81.5%	87.3%	83.5%	89.3%
F1-Score (weighted)	81.6%	87.4%	83.6%	89.4%
AUC-ROC (macro)	0.856	0.918	0.872	0.931
Process time (second)	890	3,210	1,240	4,870
Number of instances	*14.000*	*14.000*	*22.000*	*22.000*

	Dataset 1		Dataset 2	
	STM	**DL**	**STM**	**DL**
Accepted Pseudo-Labels (APL)	10,960 (87.0%)	12,090 (96.0%)	28,400 (90.2%)	30,100 (95.6%)
Rejected Pseudo-Labels (RPL)	1,640 (13.0%)	490 (4.0%)	3,100 (9.8%)	1,400 (4.4%)

As shown in Table 1, the Deep Learning model (Bi-LSTM) consistently outperforms Statistical Machine Learning (Random Forest) in all evaluation metrics, both in Dataset 1 and Dataset 2. Additionally, DL demonstrates significantly higher efficiency in utilizing unlabeled data, with Accepted Pseudo-Labels (APL) achieving 96.0% on Dataset 1 and 95.6% on Dataset 2, compared to STM, which only achieved 87.0% and 90.2%, respectively.

Discussion

The experimental results support Hypothesis 1: Deep Learning (Bi-LSTM) significantly outperforms Statistical ML in terms of accuracy and other evaluation metrics. An average improvement of around 5.5–6% in the F1-score indicates that DL's ability to capture context and word order is highly beneficial for analyzing the sentiment of Indonesian texts, which are often informal and rich in nuance. Hypothesis 4 is confirmed: Semi-supervised learning improves the performance of both approaches, but the improvement is more significant in DL. This can be seen from the higher amount of pseudo-labeled data (95.6% vs. 90.2%) and lower manual annotation requirements (4.4% vs. 9.8%). This shows that DL models are more confident and consistent in providing high-quality predictions on unlabeled data. Hypothesis 2 is confirmed: DL excels at handling unstructured data (text), but at a significantly higher computational cost (four times longer, twice as much RAM, and requiring a GPU). For resource-constrained environments, SML remains a rational choice. Hypothesis 3 is also proven: SML is much easier to interpret. This is important in business or regulatory contexts where model decision transparency is required. However, DL shows better generalization to language variations (such as slang, typos, and negation) due to its ability to understand context sequentially.

Additional experiments with varying thresholds (0.7, 0.75, 0.8, 0.85) show that at a threshold of 0.7, DL remains superior; however, the amount of noise increases, resulting in a 2-3% drop in F1-score. At a threshold of 0.85, SML loses many pseudo-label data points, resulting in performance stagnation. The optimal threshold: 0.8 provides the best balance between the quality and quantity of pseudo-label data. This research makes a significant contribution to the field of Indonesian NLP. The results show that even with limited labeled data, DL is still able to learn effectively — as long as it is supported by good embedding representation (FastText) and a suitable architecture (Bi-LSTM). This opens up opportunities for the development of local NLP models without complete dependence on manual annotation.

CONCLUSIONS

Based on the research results and discussion, it can be concluded that Deep Learning (Bi-LSTM) consistently outperforms Statistical Machine Learning in sentiment classification performance in semi-supervised learning scenarios, with an average F1-score improvement of 5.8%. This demonstrates that DL is better equipped to capture the complexity and context of Indonesian text. Semi-supervised learning effectively improves the performance of both

approaches. However, DL shows much higher efficiency in utilizing unlabeled data — with 95.6% of data successfully pseudo-labeled without human intervention, compared to 90.2% in SML. There is a significant trade-off between accuracy and computational requirements: DL requires 4x more time and resources, and relies on GPUs. SML remains the optimal choice for environments with limited resources or when interpretability is a priority. For the context of Indonesian sentiment analysis with limited labeled data, Deep Learning is more recommended if the computing infrastructure is adequate. However, if resources are limited or model transparency is required, Statistical Machine Learning (especially Random Forest with TF-IDF) still provides highly competitive performance. A confidence threshold of 0.8 is the optimal point for balancing the quality of pseudo-labels and the efficiency of the semi-supervised learning process in both approaches.

LIMITATIONS & FURTHER RESEARCH

To keep this research focused and on track, several research limitations were imposed, including the fact that the dataset used in this research consisted of Indonesian-language text sourced from public datasets and social media crawling results. The statistical machine learning models compared included K-Nearest Neighbors (KNN), Naïve Bayes, Random Forest, and Support Vector Machine (SVM) working in ensemble mode. The Deep Learning models used include Bi-LSTM, which represents neural network-based methods for text processing. Model performance evaluation is based on accuracy, precision, recall, F1-score, and AUC-ROC metrics. Experiments are then conducted in a limited computing environment, using standard hardware with available GPUs.

The studies in the three years above were limited to sentiment analysis using two-class or three-class deep learning methods in languages other than Indonesian. Annotation also only used Bag-Of-Word-based vectorization or simple word embedding methods. The semi-supervised sentiment analysis models in recent studies require review, and their performance should be compared with that of other machine learning models.

The suggestion for further research is to explore other DL architectures, such as Transformer (BERT-base multilingual or IndoBERT), to see if the performance improvement is still significant with higher computational costs. It can also be tested on more specific text domains (medical, legal, political) to see the generalization of the model. In the pseudo-label acquisition process, develop a dynamic adaptive threshold strategy based on the distribution of confidence scores per iteration. Given the large amount of unlabeled data that cannot be labeled, it is necessary to integrate active learning techniques to minimize the need for further human intervention. The research can also be expanded to Indonesian regional languages to broaden the scope of the research impact.

REFERENCES

Al-Laith, A., Shahbaz, M., Alaskar, H. F., & Rehmat, A. (2021). AraSenCorpus: A Semi-Supervised Approach for Sentiment Annotation of a Large Arabic Text Corpus. *Applied Sciences (Switzerland), 11*(5), 1–19. https://doi.org/10.3390/app11052434

Bashir, D., Montañez, G.D., Sehra, S., Segura, P.S., Lauw, J. (2020). *An Information-Theoretic Perspective on Overfitting and Underfitting*. In: Gallagher, M., Moustafa, N., Lakshika, E. (eds) AI 2020: Advances in Artificial Intelligence. AI 2020. Lecture Notes in Computer Science(), vol 12576. Springer, Cham. https://doi.org/10.1007/978-3-030-64984-5_27

Cahyana, N. H., Saifullah, S., Fauziah, Y., Aribowo, A. S., & Drezewski, R. (2022). Semi-supervised Text Annotation for Hate Speech Detection using K-Nearest Neighbors and Term Frequency-Inverse Document Frequency. *International Journal of Advanced Computer Science and Applications, 13*(10), 147–151. https://doi.org/10.14569/IJACSA.2022.0131020

Khan, J., & Lee, Y. K. (2019). LeSSA: A Unified Framework Based on Lexicons and Semi-Supervised

Learning Approaches for Textual Sentiment Classification. *Applied Sciences (Switzerland), 9*(24). https://doi.org/10.3390/app9245562

Omoseebi, A., Ola, G., & Tyler, J. (2025, February). *Data Preparation and Feature Engineering*. ResearchGate. https://www.researchgate.net/publication/389860294_Data_Preparation_and_Feature_Engineering

Schulz, M. A., Yeo, B. T. T., Vogelstein, J. T., Mourao-Miranada, J., Kather, J. N., Kording, K., Richards, B., & Bzdok, D. (2020). Different Scaling of Linear Models and Deep Learning in Ukbiobank Brain Images Versus Machine-Learning Datasets. *Nature Communications, 11*(1). https://doi.org/10.1038/s41467-020-18037-z

van Engelen, J. E., & Hoos, H. H. (2020). A Survey on Semi-Supervised Learning. *Machine Learning, 109*(2), 373–440. https://doi.org/10.1007/s10994-019-05855-6