JSCE (Journal of System and Computer Engineering)
Vol.6, No.3, July 2025, pp. 258-270
ISSN (online): 2723-1240
DOI : https://doi.org/10.61628/jsce.v6i3.2058

Comparative Analysis of SVM and IndoBERT for Intent Classification in Indonesian Overtime Chatbots

Rahmad Santosa[1,a,*]; Adetiya Bagus Nusantara[2,b]; Syaiful Imron[3,c];
[1,3] Department of Information System and Technology, Faculty of Information Technology, Institut Teknologi dan Bisnis PGRI Dewantara, Jombang 61471, Indonesia
[2] Department of Informatics, Faculty of Intelligent Electrical and Informatics Technology, Institut Teknologi Sepuluh Nopember, Surabaya 60111, Indonesia
[a] rahmad@itebisdewantara.ac.id; [b]nusantara@its.ac.id; [c]imron@itebisdewantara.ac.id
** Corresponding author*

Article history: Received Month xx, 2021; Revised Month xx, 2021; Accepted Month xx, 2021

Abstract

Digital transformation in higher education requires the development of intelligent and adaptive information systems, including services such as overtime submission for university staff. Chatbots offer a promising solution to enhance user interaction with the E-LEMBUR system. However, developing chatbots in academic settings poses challenges, including limited training data, complex overtime policies, and diverse institutional terminology. This study compares two intent classification approaches: Support Vector Machine (SVM), a traditional machine learning method, and IndoBERT, a transformer-based model designed for the Indonesian language. This study compares SVM and IndoBERT for intent classification in Indonesian overtime chatbots using 250 real queries. With oversampling and fine-tuning, IndoBERT achieved 87% accuracy, outperforming SVM (85%). Despite its accuracy, IndoBERT demands more resources. These findings support the use of transformer-based models in low-resource educational chatbot systems and provide practical guidance for real-world implementation.

Keywords— *Chatbot, IndoBERT, Intent Classification, SVM, Transformation Digital*

1. Introduction

Digital transformation in higher education necessitates the development of more intelligent and adaptive information systems, including in overtime management services (E-LEMBUR). Digital transformation has been shown to reshape industries and affect society and institutions (Kraus et al., 2021). In this context, management information systems (MIS) have evolved into strategic components that transform business models and enhance interactions between organizations and their users, a relevance that also applies to higher education institutions (Balisa et al., 2024). Chatbots have been shown to significantly enhance the responsiveness and efficiency of public services by automating interactions and reducing operational costs, while also improving user satisfaction and transparency (Putri et al., 2024). Within this setting, chatbots have emerged as a potential solution to improve the efficiency of user-system interactions. However, the development of chatbots in university environments faces unique challenges, such as limited training data, the contextual complexity of overtime policies, and the diversity of academic terminology. Most previous studies have focused on chatbot implementations in corporate settings with access to large datasets, while solutions tailored for higher education institutions with limited resources remain underexplored.

This study aims to evaluate two intent classification approaches for the E-LEMBUR chatbot: Support Vector Machine (SVM), a traditional yet efficient method for limited datasets, and IndoBERT, a state-of-the-art transformer-based model optimized for the Indonesian language. The dataset consists of 250 real-world questions collected from the overtime management system at Institut Teknologi Sepuluh Nopember (ITS), reflecting various typical scenarios encountered in university overtime submission processes.

The experimental results show that IndoBERT achieved an accuracy of 87%, outperforming SVM, which attained 85% accuracy. Nevertheless, transformer-based models require significantly greater computational resources, presenting a trade-off between accuracy and efficiency that must be carefully considered when implementing real-world systems in campus environments. Automatic parameter tuning—such as minimum support in association rules—has been shown to reduce computational load without compromising output quality (Hikmawati et al., 2021). In this context, adapting similar methods to determine optimal thresholds for NLP models like IndoBERT has the potential to mitigate high resource demands. Furthermore, a deeper analysis revealed distinct patterns and structures in questions related to overtime, which differ from those typically found in corporate contexts.

This study contributes to three main aspects: 1) It empirically demonstrates the effectiveness of IndoBERT for intent classification in Indonesian-language chatbots, particularly within small-scale and low-resource university datasets—an area that remains underexplored in existing literature. 2) It establishes the first benchmark for intent classification performance in overtime chatbot systems tailored to Indonesian higher education institutions.3) It provides actionable deployment recommendations based on real-world infrastructure constraints, offering practical guidance for IT teams in universities aiming to implement intelligent staff service systems.

2. **Method**

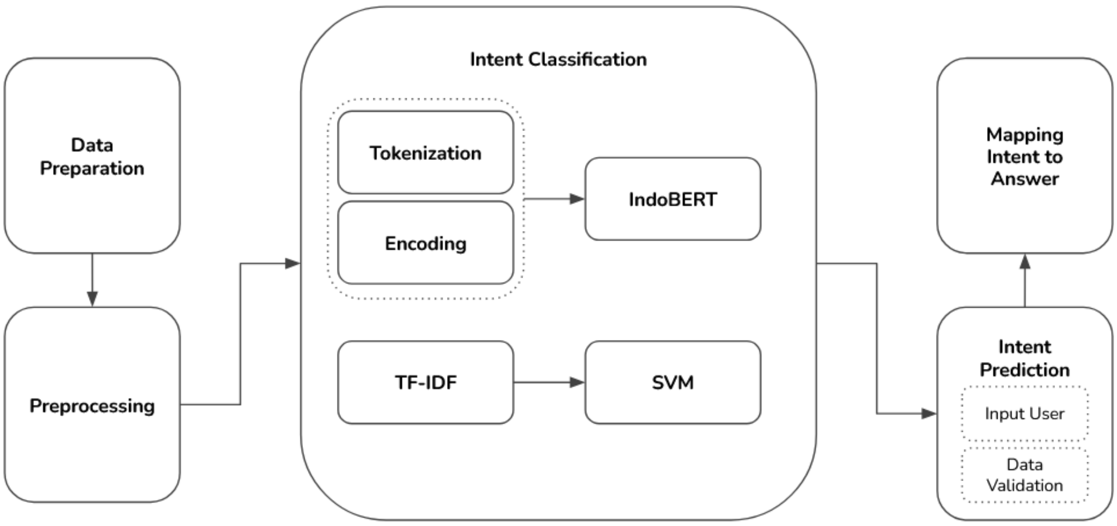


Figure 1. Architecture Design

This study is quantitative research employing an experimental approach aimed at evaluating and comparing the performance of two intent classification methods for an overtime chatbot in a higher education setting: Support Vector Machine (SVM) and IndoBERT. The chatbot examined in this study falls into the closed-domain category (Adamopoulou & Moussiades, 2020), as its scope is limited to the context of university overtime services. SVM was selected as a representative of traditional machine learning approaches known for their efficiency with small datasets, while IndoBERT was chosen as a representative of advanced transformer-based models optimized for the Indonesian language. This research focuses on analyzing the performance of both models on a limited and domain-specific dataset that reflects the real-world conditions of overtime information systems in universities. The overall research workflow is illustrated in

DOI : https://doi.org/10.61628/jsce.v6i3.2058

Figure 1, starting from the Data Preparation Stage, Data Preprocessing, Intent Classification, Intent Prediction, and Mapping Intents to Answer which will be explained in the next subsection.

2.1 Data Preparation

The data used in this study were sourced from official documents outlining the overtime policy for administrative staff at Institut Teknologi Sepuluh Nopember (ITS). These documents contain information related to the procedures for submitting overtime requests, administrative requirements, validation and verification processes, as well as the rights and obligations of employees in the overtime submission process.

Based on these documents, the researchers constructed a Frequently Asked Questions (FAQ) dataset representing potential user inquiries within the E-LEMBUR system. A sample of the data is presented in Table 1. Each entry consists of two main components:

- A question (in natural language text in Indonesian)
- An intent label representing the purpose of the question.

Table 1. Sample FAQ Dataset

Question	Intent
Bagaimana jika pegawai tidak bisa absen lembur?	absen_gagal
Bagaimana jika lembur dilakukan di hari cuti sakit?	cuti_sakit
Apakah ada kompensasi makan untuk lembur di atas jam kerja?	makan_lembur
Apakah itu uang lembur?	definisi_lembur
Berapa jam minimal untuk mendapatkan konsumsi lembur?	minimal_konsumsi
Siapa yang harus menyetujui lembur?	persetujuan_lembur
Bagaimana ketentuan lembur untuk hari Minggu?	minggu
Berapa jumlah maksimal konsumsi lembur dalam sehari?	batas_konsumsi
Apakah lembur di malam hari tarifnya lebih tinggi dari siang?	lembur_malam
Berapa batas maksimal jam lembur dalam sebulan?	batas_jam_lembur
Apakah ada kompensasi selain uang untuk kerja lembur?	kompensasi_tambahan

The compiled dataset consists of 250 questions collected from various internal sources. All questions represent real-world contexts and are designed to reflect the information needs commonly raised by users of the E-LEMBUR system.

Each question was manually labeled by the researchers using an intent classification scheme consisting of 95 intent categories. These labels cover a wide range of specific topics, including overtime requirements, submission procedures, document validation, disbursement status, and unit-specific policies. The diversity of these labels reflects the complexity of the overtime system within the university setting and poses a challenge for the classification process due to semantic overlaps between intents.

2.2 Preprocessing

The preprocessing stage was carried out to prepare the textual data for use in both classification models, SVM and IndoBERT. In general, the questions consist of interrogative sentences in Indonesian, derived from overtime policy documents and administrative correspondence. Prior to training, several preprocessing steps were applied as follows:

2.2.1 Data balancing through oversampling

The initial dataset consists of 250 questions, with an imbalanced distribution across the 95 intent categories, as illustrated in Figure 2. Most intents contain only 1–3 samples, while a few others have significantly more. This imbalance can lead the classification model to be biased toward the majority classes.

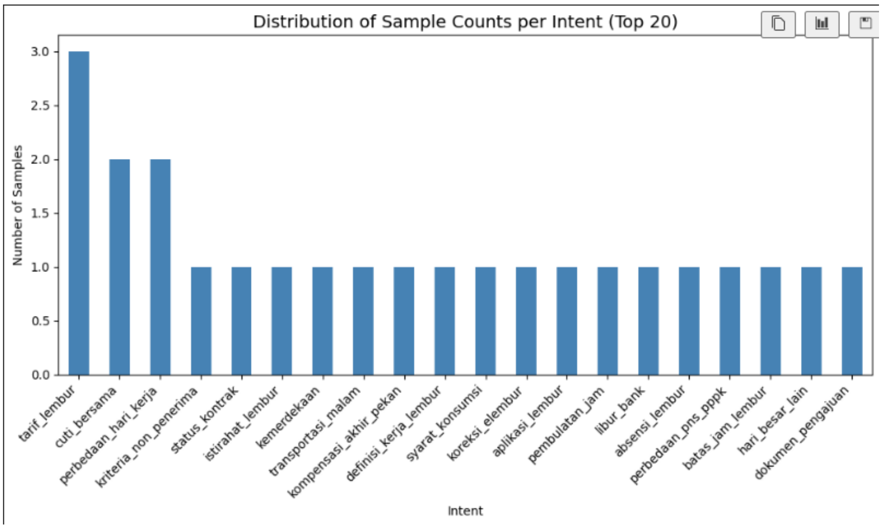


Figure 2. Dataset Before Oversampling

To address this issue, an oversampling process was applied so that each intent category contained an equal number of samples, five per intent resulting in a total dataset of 475 samples, as shown in Figure 3. Oversampling was conducted through random duplication of existing data without applying text augmentation. The goal of this process was to ensure that each intent received sufficient representation during model training, thereby reducing bias toward any particular class. Class distribution imbalance has been shown to cause significant misclassification in SVM models, even when the overall accuracy appears relatively acceptable (75.1%) (Salleh et al., 2024). This finding underscores the importance of oversampling techniques in ensuring fair representation across all classes, as implemented in our study.

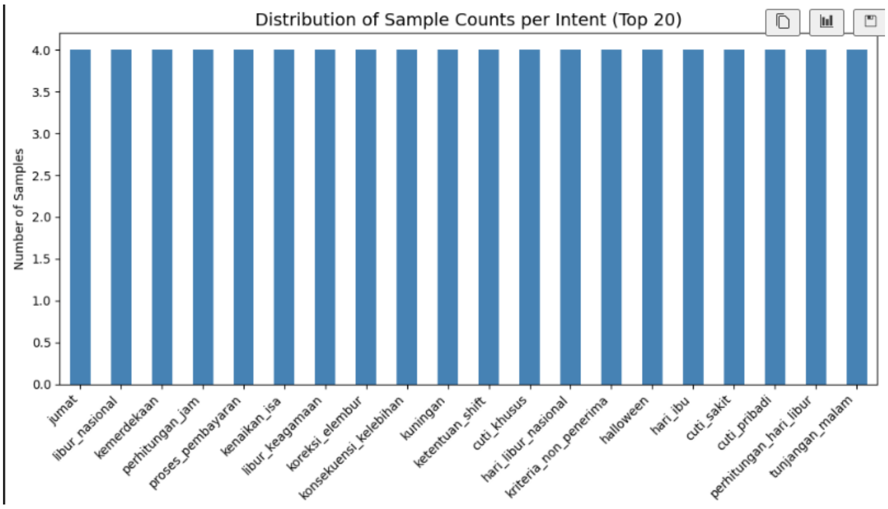


Figure 3. Dataset After Oversampling

2.2.2 Preprocessing for SVM

In a previous study, SVM was used to classify flood disaster levels into low, medium, and high categories using news data from BNPB and TF-IDF-based preprocessing (Santosa et al., 2024). This approach serves as the foundation for selecting SVM as the baseline method in the present study.

As a baseline, a traditional machine learning approach was implemented using Term Frequency-Inverse Document Frequency (TF-IDF) for text vectorization and Support Vector

Machine (SVM) as the classification algorithm. The textual data were first converted to lowercase and tokenized. At this stage, no stemming or stopword removal was applied, in order to preserve the semantic meaning of the original questions. Unlike the findings of (Kamath et al. 2018), this study deliberately avoided stemming and stopword removal to retain full semantic context, which is critical for transformer-based models like IndoBERT. The vectorization process was carried out using the *TfidfVectorizer* function from the *scikit-learn* library, with default settings for *n-grams* and *maximum features*.

2.2.3 Preprocessing for IndoBERT.

Text preprocessing is considered a critical step in natural language processing (NLP) to ensure that textual data can be effectively processed by the model (Assayed, Shaalan, & Alkhatib, 2023). In this study, data preprocessing was kept minimal, limited to two main steps: tokenization and encoding, without applying stemming, stopword removal, or other text normalization techniques. This approach was chosen to preserve the full semantic context of each question. The preprocessing steps included:

- **Tokenisasi:** Each question was processed using the IndoBERT tokenizer, which automatically splits the text into subword tokens based on the WordPiece algorithm. The tokenized input was then optimized using padding and truncation to ensure uniform input length, with a maximum limit set at 128 tokens per question.
- **Encoding:** The tokens resulting from the tokenization process were then converted into the numerical format required by the model, namely:
    - *input_ids*: a numerical ID representation for each token.
    - *attention_mask*: position indicators for active tokens (non-padding tokens).
    - *token_type_ids*: input segmentation indicators (not specifically used, as the input consists of a single sentence).

2.3 Classification Method

This study compares two intent classification approaches: Support Vector Machine (SVM) as the traditional baseline, and IndoBERT as a transformer-based model for the Indonesian language.

2.3.1 Support Vector Machine (SVM)

The SVM algorithm is a supervised machine learning algorithm based on statistical learning theory. The Support Vector Machine (SVM) algorithm has been shown to select a subset of features from the training samples such that the classification of this subset is equivalent to the partitioning of the entire dataset (Abdullah & Abdulazeez, 2021). As a baseline approach, this study employs the Linear Support Vector Classification (LinearSVC) algorithm from the scikit-learn library. Input features are derived from the TF-IDF (Term Frequency–Inverse Document Frequency) representation of the preprocessed textual data. The model is trained using oversampled training data to ensure balanced class distribution. Training is performed using default parameters without hyperparameter tuning, and predictions are evaluated on a separate test set.

2.3.2 IndoBERT (Transformer-Based Model)

For the transformer-based approach, the IndoBERT base model (*indobenchmark/indobert-base-p1*) was fine-tuned using the HuggingFace *Transformers* library and the Trainer API. The model was adapted for a multi-class intent classification task involving a total of 95 labels. Each input text was processed using the IndoBERT tokenizer and encoded into *input_ids*, *attention_mask*, and *token_type_ids*. The standard practice for fine-tuning transformer-based models like IndoBERT involves adding a softmax layer on top of the sentence-level representation (e.g., the hidden state of the [CLS] token) for multi-class classification tasks, with cross-entropy loss serving as the training objective (Pan et al., 2022).

Training was conducted for 10 epochs with an evaluation and model checkpointing scheme based on evaluation loss. The training configuration was optimized for GPU efficiency using *fp16* (mixed precision training) and gradient accumulation to stabilize weight updates. The model with the lowest *eval_loss* was saved as the best-performing model and used for the final evaluation stage.

2.3.3 Model Evaluation

The model's performance was evaluated using several classification evaluation metrics commonly used in machine learning research, namely:

- **Accuracy:** the proportion of correct predictions to the total number of test samples. However, this metric can be misleading when dealing with imbalanced data (Miao & Zhu, 2021). The formula is:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

- **Precision:** the ratio of true positive predictions to all positive predictions, representing how relevant the returned results are. The formula is:

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

- **Recall:** the ratio of true positive predictions to all actual positive instances, indicating how completely the model retrieves relevant data. The formula is:

$$Recall = \frac{TP}{TP+FP} \quad (3)$$

- **F1-Score:** the harmonic mean of precision and recall, providing a balanced measure between the two. The formula is:

$$F1 - Score = 2x\frac{TP}{TP+FP} \quad (4)$$

- **Confusion Matrix:** an evaluation tool used to assess the performance of a classification model by comparing its predictions against the true labels. In the context of multi-class classification, the confusion matrix is used to display the number of correct predictions—true positives (TP) and true negatives (TN)—and incorrect predictions—false positives (FP) and false negatives (FN)—for each class, helping to identify patterns of classification errors (Heydarian, Doyle, & Samavi, 2022).

All of these metrics were calculated on a test set that was separate from the training data, in order to assess the model's generalization capability to new data. Evaluation was conducted using macro averaging, as the number of samples per class had been balanced through oversampling, and to avoid bias toward majority classes.

2.4 Intent Prediction.

After the training and testing processes of the intent classification model were completed, the next step was to perform intent prediction or validation on new data that was not part of the previous training or testing sets. This validation aims to assess the model's ability to predict intents on real-world data or data with characteristics different from the training data (out-of-distribution samples).

A total of 10 new questions were constructed based on variations of real-world cases in the campus overtime system, as shown in Table 2, which had not appeared in the training data. These questions represent potential user queries with greater variation in sentence structure and semantic context. This preliminary validation, involving ten unseen queries, was conducted to assess the model's generalization capability beyond the training data. While the results were promising, especially for IndoBERT, further validation with larger-scale user interactions is recommended to ensure robustness and practical applicability.

The model predicted the intent of each of these questions, and the predicted results were manually compared with the predefined ground truth labels. This validation was used to evaluate the model's consistency and generalization capability, as well as to identify potential prediction errors that may not have been detected in the test data.

Table 2. Dataset Validation Question

Question	Intent
Bisa dijelaskan pengertian dari kerja lembur?	definisi_lembur
Pegawai dengan status apa saja yang berhak kompensasi lembur?	kriteria_penerima
Kalau saya golongan III, berapa bayaran per jam untuk kerja di luar jam kerja?	tarif_lembur
Bagaimana cara menghitung upah lembur saat hari libur resmi?	perhitungan_hari_libur
Dokumen apa yang harus disiapkan bila ingin mengajukan lembur?	persyaratan_lembur
Apa batas waktu kerja lembur maksimal dalam satu bulan?	batas_jam_lembur
Bagaimana perhitungan potongan pajak untuk pendapatan lembur?	pajak_lembur
Sistem apa yang dipakai untuk memasukkan dalam lembur di ITS?	aplikasi_lembur
Berapa kali lipa upah lembur saat hari raya Idul Fitri?	tarif_hari_raya
Setelah berapa jam kerja lembur bisa mendapatkan makan?	konsumsi_lembur

2.5 Mapping Intents to Answers

After the intent classification process is completed, the system proceeds to the intent mapping stage, where each recognized intent is linked to a relevant answer. Each identified intent is matched with a corresponding answer entry in a pre-constructed knowledge base. This knowledge base contains question-answer pairs designed based on the institution's overtime policy documents and administrative procedures.

The mapping process is performed deterministically, where each intent label corresponds to a fixed answer (one-to-one mapping) or multiple answer variations selected contextually when necessary. This mapping scheme enables the chatbot system to provide automated and consistent responses aligned with the user's question intent. A sample of the answer dataset is presented in Table 3. With this approach, the system is not only capable of classifying user intent but also delivering appropriate responses directly, making the E-LEMBUR chatbot an efficient and informative interaction tool.

Table 3. Sample FAQ Dataset

Intent	Answer
definisi _lembur	Uang lembur merupakan kompensasi bagi tenaga kependidikan tetap yang melakukan kerja lembur berdasarkan surat perintah dan otorisasi dari pejabat yang berwenang.
kriteria_penerima	Uang lembur diberikan kepada tenaga kependidikan tetap kecuali Kepala Biro/setara ke atas yang melakukan kerja lembur.
absen_gagal	Jika tidak bisa absen, harus ada verifikasi manual oleh pimpinan unit
batasan_jam_lembur	Batas maksimal jam lembur dalam 1 bulan adalah 40 jam

3. **Results And Discussion**

Experiments were conducted by comparing three intent classification scenarios: (1) a baseline model using Support Vector Machine (SVM), (2) IndoBERT trained for 10 epochs, and (3) IndoBERT trained for 15 epochs. The dataset used consists of 475 question samples that have been balanced through oversampling techniques on the training data, while the test data is left in its original distribution to maintain the validity of the model evaluation. The data division uses a 70:30 ratio between training and test data.

Table 4. Sample FAQ Dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
SVM	85	84	85	84
IndoBERT Epoch 10	73	73	73	73
IndoBERT Epoch 15	84	84	84	84

The evaluation results in Table 4 indicate that the IndoBERT model trained for 15 epochs achieved optimal performance, with accuracy, precision, recall, and F1-score all reaching 87%. However, this came at the cost of a significantly longer training time—114 minutes, which is 5.4× longer than the 10-epoch training that took only 21 minutes and yielded lower performance (73%). In contrast, the baseline SVM model demonstrated competitive performance, achieving 85% accuracy, 84% precision, 85% recall, and an F1-score of 84%. While SVM was slightly less accurate than IndoBERT-15, it offered notable advantages in terms of computational efficiency and faster training time. These findings highlight a clear trade-off between model accuracy and computational cost, suggesting that the choice of model should be guided by the specific requirements and constraints of the application.

✓ 114m 28.9s Python

[165/165 1:54:00, Epoch 15/15]

Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1-score
1	2.513300	2.923557	0.718310	0.711268	0.718310	0.713615
2	2.291700	2.777162	0.746479	0.746479	0.746479	0.746479
3	2.104500	2.608993	0.732394	0.741784	0.732394	0.729108
4	1.899100	2.380436	0.774648	0.774648	0.774648	0.774648
5	1.656200	2.145389	0.774648	0.764085	0.774648	0.766197
6	1.443500	1.964905	0.788732	0.772300	0.788732	0.778404
7	1.326600	1.819576	0.788732	0.779343	0.788732	0.781690
8	1.179600	1.676167	0.816901	0.807512	0.816901	0.811268
9	1.066100	1.582605	0.816901	0.793427	0.816901	0.800469
10	0.942500	1.499204	0.830986	0.835681	0.830986	0.830047
11	0.939900	1.438500	0.830986	0.826291	0.830986	0.823005
12	0.822100	1.397803	0.830986	0.816901	0.830986	0.815962
13	0.786100	1.358542	0.873239	0.863850	0.873239	0.866197
14	0.778100	1.340119	0.873239	0.873239	0.873239	0.873239
15	0.723900	1.335168	0.873239	0.873239	0.873239	0.873239

Figure 4. Result Training Test IndoBERT Epoch 15

SVM with a linear kernel has been shown to achieve an accuracy of 86% and a specificity of 89%, with relatively short training time compared to other kernels (polynomial/RBF) (Rochim et al., 2021). This supports the notion that SVM offers high accuracy with efficient training, while transformer-based models like IndoBERT provide superior contextual understanding. These findings are consistent with our results, which indicate that SVM remains a competitive and efficient baseline, even though its accuracy is slightly lower than that of transformer-based models. These findings align with recent studies that highlight the effectiveness of combining

IndoBERT for feature extraction and MBERT for classification, which achieved an F1-score of 0.9032 and demonstrated stable performance across datasets (Nabiilah et al., 2024).

Precision refers to the ratio of true positive predictions to the total number of instances predicted as positive. Recall denotes the proportion of true positive predictions relative to all actual positive instances. The F1-score represents the harmonic mean of precision and recall, providing a balanced measure of both metrics (Souha et al., 2023).

In the 15 epoch experiment results shown in Figure 4, the IndoBERT model shows a consistent upward trend in all evaluation metrics. Accuracy and F1-score peaked at 87.32% on the 15th epoch, accompanied by a stable and decreasing validation loss as shown in Figure 5, which shows good generalization ability without any signs of overfitting, indicating that this model is suitable for use in the E-LEMBUR chatbot system.

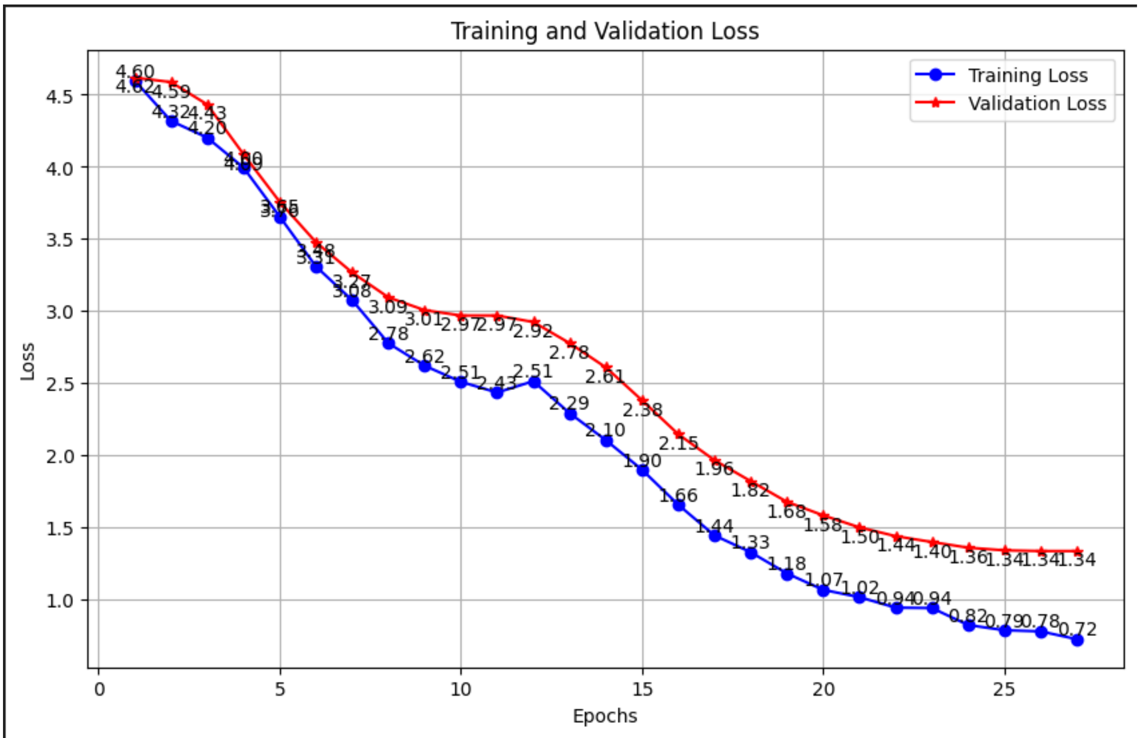


Figure 5. Result Training Test IndoBERT Epoch 15

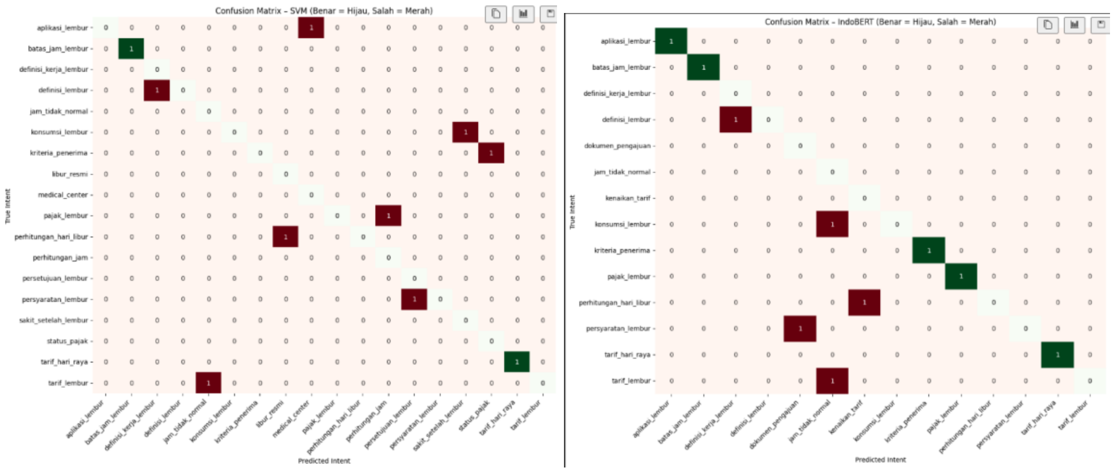
These findings highlight a significant trade-off between accuracy and efficiency in developing intent classification-based chatbots within higher education environments. IndoBERT delivers the highest accuracy but incurs higher computational costs, while SVM offers a more efficient option with still competent performance. Therefore, model selection should take into account resource constraints and the specific needs of the higher education institution.

After completing the training process for the intent classification model, testing was conducted on 10 validation question samples that were not included in the training data, as shown in Table 5 demonstrates that IndoBERT predicted 5 out of 10 intents correctly, with minor semantic drift in 3 cases. Each question represents a different intent and was evaluated based on the prediction results from two approaches: the traditional model using SVM and the transformer-based model using IndoBERT.

Table 5. Result Validation 10 Question

True Intent Question	SVM Intent Predict	IndoBERT Intent Predict
definisi_lembur	definisi_kerja_lembur	definisi_kerja_lembur
kriteria_penerima	status_pajak	kriteria_penerima
tarif_lembur	jam_tidak_normal	jam_tidak_normal
perhitungan_hari_libur	libur_resmi	kenaikan_tarif
persyaratan_lembur	persetujuan _lembur	dokumen_pengajuan
batas_jam_lembur	batas_jam_lembur	batas_jam_lembur
pajak_lembur	perhitungan_jam	pajak_lembur
aplikasi_lembur	medical_center	aplikasi_lembur
tarif_hari_raya	tarif_hari_raya	tarif_hari_raya
konsumsi_lembur	sakit_setelah_lembur	jam_tidak_normal

The prediction results are visualized using a confusion matrix, which illustrates the alignment between the true intent label and the intent predicted by each model, as shown in Figure 6, that IndoBERT can predict correctly (True Positive) marked in green. In the SVM confusion matrix, only two out of ten predictions match the true intent, while the remaining eight are misclassified. This shows the limitation of SVM in handling complex semantic variations across different intents.



a. SVM Confusion Matrix b. IndoBERT Confusion Matrix

Figure 6. Confusion Matrix Comparison

In contrast, the IndoBERT model correctly predicted seven out of ten intents. The remaining three were misclassified, but the semantic proximity of the predicted intents was still relatively acceptable within the context of natural language understanding. The resulting confusion matrix visually distinguishes correct predictions (green diagonal) from incorrect ones (red off-diagonal), facilitating easier interpretation of the model's performance.

Misclassifications in models such as SVM and IndoBERT such as predicting the intent *'definisi_lembur'* as *'definisi_kerja_lembur'* or *'pajak_lembur'* as *'perhitungan_jam'* are examples of technical obstacles that can exacerbate usability barriers.Inaccurate chatbot responses have been shown to diminish system utility and erode user trust in AI (Han & Lee, 2022). In this context, errors involving intents with similar terminology or ambiguous phrasing may be perceived by users as a failure of the system to truly "understand" their needs, thereby reinforcing resistance and undermining confidence in the technology. A bar chart comparing the number of correct predictions between SVM and IndoBERT (15 epochs) can be seen in Figure 7, where it can be seen that IndoBERT is superior in intent prediction.

These findings reinforce the conclusion that IndoBERT outperforms in understanding the contextual meaning of questions within the domain of university overtime, particularly when dealing with small datasets that exhibit high linguistic variation. The suitability of SVM for deployment in resource-constrained environments is supported by previous research, which indicates that transformer-based models such as IndoBERT require larger datasets to achieve optimal performance (Wang et al., 2020).

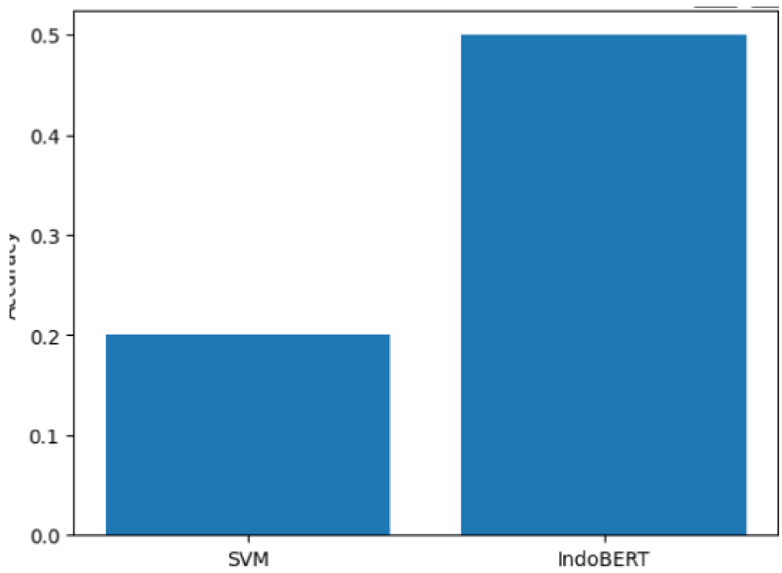


Figure 7. Bar Chart True Predict Intent SVM and IndoBERT

4. **Future Work and Conclusion**

4.1 Future Work

To enhance the robustness and applicability of the proposed chatbot system, future research may consider the following direction:

- **Expanding the dataset** by incorporating queries from multiple educational institutions to improve model generalizability, especially in addressing the data scarcity challenge in low-resource languages like Indonesian (Perdana & Adikara, 2025).
- **Integrating Named Entity Recognition (NER)** using Conditional Random Fields (CRF), as proposed in IndoBERT multitask learning architecture, to extract key entities (e.g., organization name, applications) and improve intent classification performance, achieving an F1-score of up to 0.96 (Perdana & Adikara, 2025)
- **Conducting user studies** with real users to evaluate the chatbot usability, trustworthiness, and effectiveness in real-world educational service scenarios (Perdana & Adikara, 2025)
- **Exploring lightweight models** such as DistililIndoBERT for deployment in hardware-constrained environments, leveraging the architectural flexibility demonstrated in modified DIET-based approaches (Perdana & Adikara, 2025)
- **Applying advanced data augmentation techniques** such as multi-word insertion, which have shown effectiveness in preserving semantic and sentiment consistency in Indonesia text classification tasks (Muftie & Haris, 2023)

4.2 Conclusion

Based on the findings, the following conclusions can be drawn:

- **IndoBERT outperformed SVM**, achieving 87% accuracy in intent classification for overtime chatbot systems in higher education.

- **SVM remains a viable alternative** for low-resource environments due to its efficiency and competitive performance (85% accuracy).
- **Validation with unseen quires** confirm IndoBERT superior ability to understand complex and contextual intents
- **Confusion matrix analysis** shows IndoBERT higher precision in intent identification, though it requires greater computational resources.
- **A trade-off between model quality and resource usage** must be considered when deploying chatbot systems in real-world academic settings.

Acknowledgements

The authors would like to thank Institut Teknologi dan Bisnis PGRI Dewantara and Institut Teknologi Sepuluh Nopember for supporting this research. This work was conducted as part of a collaborative initiative to enhance digital administrative services in Indonesian higher education institutions.

References

Santosa, R., Fariza, A., & Arifin, F. (2024). Classification of flood disaster level news articles using Machine Learning. *Indonesian Journal of Computer Science*, 13(1), 264–275.

Adamopoulou, E., & Moussiades, L. (2020). An overview of chatbot technology. In *Artificial Intelligence Applications and Innovations: 16th IFIP WG 12.5 International Conference, AIAI 2020, Neos Marmaras, Greece, June 5–7, 2020, Proceedings, Part II 16* (pp. 373-383). Springer International Publishing.

Wang, P., Fan, E., & Wang, P. (2020). Comparative Analysis of Image Classification Algorithms Based on Traditional Machine Learning and Deep Learning. Pattern Recognition Letters, 136, 1-9. https://doi.org/10.1016/j.patrec.2020.07.042

Kamath, C. N., Bukhari, S. S., & Dengel, A. (2018). Comparative Study between Traditional Machine Learning and Deep Learning Approaches for Text Classification. DocEng '18: ACM Symposium on Document Engineering, 1-11. https://doi.org/10.1145/3209280.3209526

Han, S., & Lee, M. K. (2022). FAQ chatbot and inclusive learning in massive open online courses. *Computers & Education*, *179*, 104395.

Hikmawati, E., Maulidevi, N. U., & Surendro, K. (2021). Minimum threshold determination method based on dataset characteristics in association rule mining. *Journal of Big Data*, *8*, 1-17.

Nabiilah, G. Z., Alam, I. N., Purwanto, E. S., & Hidayat, M. F. (2024). Indonesian multilabel classification using IndoBERT embedding and MBERT classification. *International Journal of Electrical & Computer Engineering (2088-8708)*, *14*(1).

Muftie, F., & Haris, M. (2023, August). Indobert based data augmentation for Indonesian text classification. In *2023 International Conference on Information Technology Research and Innovation (ICITRI)* (pp. 128-132). IEEE.

Rochim, A. F., Widyaningrum, K., & Eridani, D. (2021, December). Performance Comparison of Support Vector Machine Kernel Functions in Classifying COVID-19 Sentiment. In *2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)* (pp. 224-228). IEEE.

Salleh, S. A., Khalid, N., Danny, N., Zaki, N. A. M., Ustuner, M., Latif, Z. A., & Foronda, V. (2024). Support Vector Machine (SVM) and Object Based Classification in Earth Linear Features Extraction: A Comparison. *Revue Internationale de Géomatique*, *33*.

Abdullah, D. M., & Abdulazeez, A. M. (2021). Machine learning applications based on SVM Classification a Review. *Qubahan Academic Journal*, *1*(2), 81-90.

Perdana, R. S., & Adikara, P. P. (2025). Multi-task Learning for Named Entity Recognition and Intent Classification in Natural Language Understanding Applications. *Journal of Information Systems Engineering and Business Intelligence*, *11*(1), 1-16.

Putri, S. A., Fadhila, S., & Umam, K. (2024). Peran Chatbot dalam Meningkatkan Responsivitas dan Efisiensi Pelayanan Publik pada Era Digitalisasi. *Prosiding Seri Praktikum Ilmu-Ilmu Sosial-Politik*, *1*(1), 153-159.

Kraus, S., Jones, P., Kailer, N., Weinmann, A., Chaparro-Banegas, N., & Roig-Tierno, N. (2021). Digital transformation: An overview of the current state of the art of research. *Sage Open*, *11*(3), 21582440211047576.

Miao, J., & Zhu, W. (2022). Precision–recall curve (PRC) classification trees. *Evolutionary intelligence*, *15*(3), 1545-1569.

Heydarian, M., Doyle, T. E., & Samavi, R. (2022). MLCM: Multi-label confusion matrix. *Ieee Access*, *10*, 19083-19095.

Pan, L., Hang, C. W., Sil, A., & Potdar, S. (2022, June). Improved text classification via contrastive adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 10, pp. 11130-11138).

Assayed, S., Shaalan, K., & Alkhatib, M. (2023). A chatbot intent classifier for supporting high school students. *EAI Endorsed Transactions on Scalable Information Systems*, *1*.

Souha, A., Ouaddi, C., Benaddi, L., & Jakimi, A. (2023, December). Pre-trained models for intent classification in chatbot: Comparative study and critical analysis. In *2023 6th international conference on advanced communication technologies and networking (CommNet)* (pp. 1-6). IEEE.

Balisa, D., Leffia, A., & Shino, Y. (2024). Memanfaatkan fungsi sistem informasi manajemen: Prospek dan tantangan di dunia bisnis. *Jurnal MENTARI: Manajemen, Pendidikan dan Teknologi Informasi*, *2*(2), 123-133.