International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. #, No.# (Mon-20..)

Intent Classification in Artificial Intelligence-Based Customer Service Chatbot for E-Wallet Service Providers

Christopher Owen[1], Derwin Suhartono[2]

[1] Computer Science Department, BINUS Graduate Program – Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia
[2]Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia

E-mail address: christopher.owen001@binus.ac.id, dsuhartono@binus.edu

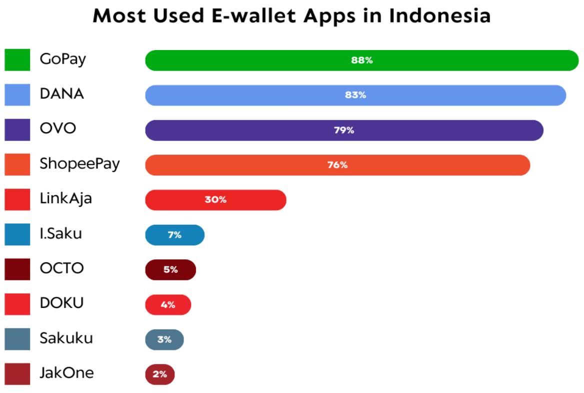
Received ## Mon. 20##, Revised ## Mon. 20##, Accepted ## Mon. 20##, Published ## Mon. 20##

Abstract: The rapid expansion of e-wallet services in Indonesia necessitates robust customer service to maintain competitiveness and user satisfaction. Current customer service chatbots in Indonesian e-wallet services primarily rely on rule-based approaches, limiting their adaptability to user needs and resulting in communication issues and negative feedback. The transition to AI-based chatbots is challenging, particularly in accurately classifying user intents in Indonesian-language due to the language's complexities and the absence of specialized models and datasets. This research proposes a customized intent classification model for AI-based customer service chatbots in e-wallet services, employing transformer-based embedding methods, specifically IndoBERT and Sentence-BERT (SBERT), with TextConvoNet classification model. Comparative analysis is conducted against conventional transformer models and the original TextConvoNet framework. The findings consistently showcase the superiority of the proposed models across various metrics, demonstrating significant advancements compared to baseline approaches. Notably, SBERT embeddings with TextConvoNet classification achieved the highest accuracy (86.90%), precision (84.81%), recall (86.90%), and F1-score (85.11%) with a learning rate of 0.001, indicating its potential to enhance customer service chatbots in e-wallet platforms. These findings not only advance AI-driven customer service within the financial sector but also offer valuable insights into the broader application of natural language processing technologies for addressing real-world challenges.

Keywords: Intent Classification, E-wallet, Customer Service, Chatbot, IndoBERT, TextConvoNet

1. Introduction

Nowadays, advancements in technology and information have driven changes in many areas of society. One notable advancement in the financial sector is the emergence of digital wallet services, commonly referred as e-wallet. E-wallet are defined as payment tools usable for online transactions via computers or smartphones [1]. As per Bank Indonesia, there are about 48 legally recognized businesses in Indonesia that provide e-wallet services. According to a survey conducted by Populix in [2] as shown in Figure 1, GoPay emerges as the leading e-wallet platform in Indonesia, commanding 88% of users. Following closely behind are DANA with 83%, OVO with 79%, ShopeePay with 76%, and LinkAja with 30%. Intense competition among e-wallet providers drives continuous improvement in service quality, highlighting the growing importance of customer service as a critical feature. As part of their customer service strategy, some businesses have introduced chatbots that function as virtual assistants and are trained through artificial intelligence.



Source: Populix (2022)

Figure 1. Most used e-wallet apps in Indonesia

E-mail: christopher.owen001@binus.ac.id, dsuhartono@binus.edu

Chatbots, defined as intelligent conversational agents capable of communicating with users in natural language [3], serve various purposes within e-wallet customer service. They address user issues, provide information and recommendations, and help manage service queues [4]. Two common types of chatbots used as customer service tools are rule-based chatbots and artificial intelligence-based chatbots. Rule-based chatbots operate on pre-defined rules, being sensitive to spelling and language conventions. They usually provide predefined inputs, and users respond based on specific keywords provided [5]. On the other hand, artificial intelligence-based chatbots often use free-text inputs, allowing users to initiate conversations using everyday language. These chatbots identify user intent using machine learning and natural language processing, providing contextually appropriate responses [6].

While rule-based chatbots offer structured interactions, they lack flexibility and can be limited in question options. In contrast, AI-based chatbots offer more flexible interactions but may sometimes lead to contextually irrelevant conversations and interpretation errors [6]. A potential solution is combining rule-based and AI-based chatbots to allow users to express themselves in natural language while the system provides multiple response options to maintain context [7], [8], [9]. However, most Indonesian e-wallet providers currently only offer rule-based customer service chatbot, leading to limitations in user communication and negative feedback on social media, potentially eroding user trust. Indonesian e-wallet companies cite difficulties in accurately classifying user intents in specific Indonesian customer service language as a major reason for not adopting AI-based chatbots. The need for specialized training models and the associated time, effort, and cost, along with the limited availability of datasets, pose challenges in achieving accurate intent classification and understanding user needs [10].

In artificial intelligence-based chatbots, a crucial component of Conversational Natural Language Processing (NLP) typically comprises two main elements: Natural Language Understanding (NLU) and Natural Language Generation (NLG). NLU involves the computer's interpretation of human language to comprehend user requests, while NLG enables the computer to generate responses understandable to users. Within NLU in a chatbot, there is often a model for intent classification. Intent classification is a text classification task aimed at understanding actions or behaviors within a given text [11]. Various models have been employed for intent classification, such as CapsNet-based artificial neural networks and transformer-based BERT models [12], IndoBERT for knowledge-based chatbots [13], and Random Forest, Linear SVC, Multinomial Naïve Bayes, and Logistic Regression for categorizing intents in product review tweets on Twitter [14]. Despite these studies, there is a lack of a specific model for classifying user intents in Indonesian for customer service in the context of e-wallet providers. Therefore, this research proposes building a model that can classify user intents in Indonesian specifically for AI-based customer service in e-wallet services, aiming to transition from rule-based to AI-based customer service chatbots.

To address this gap, this study aims to develop a research model, modifying the TextConvoNet model [15]. The proposed model will utilize transformer-based embedding methods and the same classification model as TextConvoNet. Two transformer-based models, namely IndoBERT and Sentence-BERT (SBERT), will be tested for embedding purposes. Pre-trained IndoBERT, RoBERTa, and TextConvoNet models will be used as benchmarks for comparison. The researcher anticipates that this study will enhance the flexibility of e-wallet customer service chatbots, creating human-like interactions between the chatbot and users. Additionally, the resulting model can serve as a benchmark for future research in this domain.

2. Related Works

Several studies have been conducted in the field of intent classification using various types of models. For example, [14] conducted a study comparing the performance of Random Forest, Linear SVC, Multinomial Naïve Bayes, and Logistic Regression algorithms in classifying intent categories of customer product reviews on Fitbit's official Twitter account. The results showed that Logistic Regression achieved the highest accuracy, reaching 97%.

In a separate study, [16] investigated intent detection in Lithuanian using FastText and BERT word embedding methods along with LSTM, BiLSTM, and CNN classification models. The study utilized a dataset of question-answer pairs from "Tildes Biuras" products, covering 41 intents. The findings indicated that the CNN classification model using BERT embeddings performed the best with an accuracy of approximately 71.5%.

Moreover, in the field of Customer Service, previous research has explored the implementation of intent classification using social media datasets. For example, [17] conducted a study where they fine-tuned an unsupervised learning model using data collected from PTT, one of the biggest social intercourse platforms in Taiwan. To assess the model's effectiveness, they utilized a dataset consisting of 7,459 recorded customer service interactions from a credit card bank in Taiwan. The study concluded that fine-tuning the model with social media data resulted in an accuracy improvement from 56.5% to 61.3%.

In the context of the Indonesian language, although limited, there are some studies on intent classification. [12] conducted research comparing Capsule Network

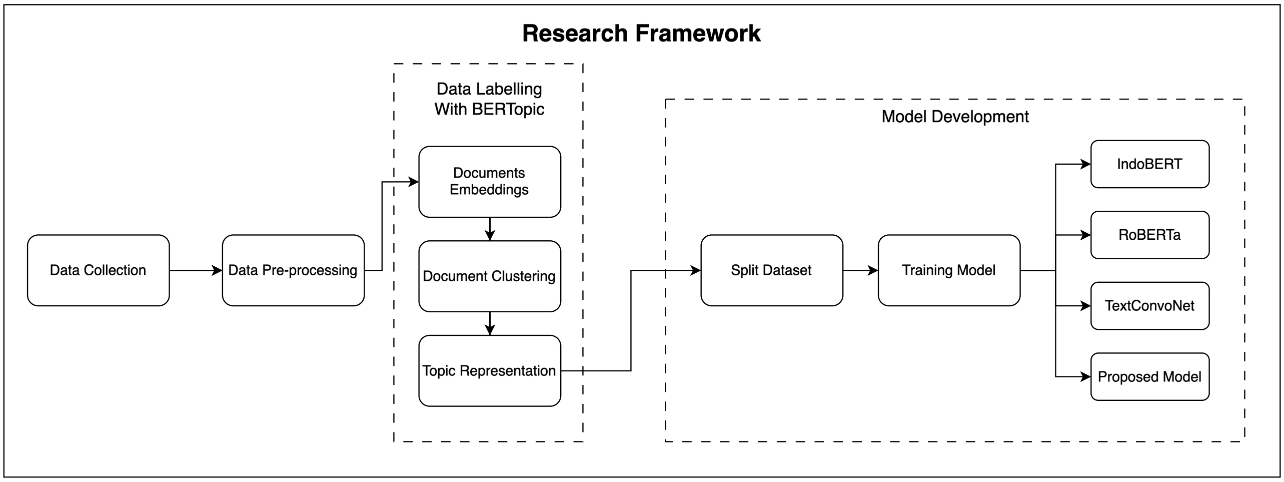


Figure 2. Research Framework

(CapsNet) and BERT models for intent detection in conversations with customers at PT. Kazee in Indonesian. The results showed that CapsNet had faster execution time, but lower accuracy compared to the BERT model. On another study, [13] used IndoBERT to build an intent classification model, distinguishing between in-scope and out-of-scope intents for a knowledge-based chatbot in Indonesian. The dataset included 548 automatically generated in-scope intent queries and out-of-scope intents from Twitter. The paper proposed an algorithm to generate an intent-specific training dataset from a knowledge base and suggested approaches to detect out-of-scope queries using BERT and Bayesian uncertainty estimation, achieving maximum F1-scores of 100% for in-scope intents and 86% for out-of-scope intents. Intent classification is a subset of text classification, and models trained for text classification can be applied to intent classification. A recent CNN-based text classification model, TextConvoNet, uses 2-dimensional convolutional filters, enabling the extraction of n-gram features within a sentence and across sentences in the text. This differs from other CNN-based classification models that typically use 1-dimensional convolutional filters. However, in the TextConvoNet model, the word embedding method still employs Word2Vec which lacks contextual information and word sequence capture during embedding. Hence, this study explores the use of transformer-based embedding methods to capture context and sequence during embedding, providing distinct representation vectors for words based on context and order. Two transformer-based models, namely IndoBERT and SBERT, will be tested for embedding. After the embedding process, the results will be fed into a CNN-based classification model with the same architecture as TextConvoNet.

3. METHODOLOGY

As depicted in Figure 2, the development of the model involves a series of crucial steps aimed at achieving the research objectives. There are four key steps, namely data collection, data pre-processing, data labelling using BERTopic, and model development.

A. Dataset

The dataset designed for model training consists of data that has been collected, subjected to pre-processing, and labelled. The collected data consists of user comments in the Indonesian language from the social media platforms of the top five prominent e-wallet service providers in Indonesia, namely GoPay, DANA, OVO, ShopeePay, and LinkAja. To ensure uniform treatment, the names of these companies will be standardized as {ewallet}. The collected data will then undergo pre-processing to generate clean data, resulting in approximately 14,000 data points after this stage. Subsequently, the data will be labelled using the topic modelling method with the BERTopic model. BERTopic will cluster similar texts together and assign labels to those clusters based on representative words with weights reflecting the entire text in each cluster. The number of labels in the dataset will depend on the outcomes of the BERTopic model.

In the data collection phase, comments from the social media platforms of an e-wallet service company will be gathered. Specifically, the focus will be on gathering comments from Instagram and Twitter, chosen due to the significant presence of e-wallet users expressing their concerns through comments on the company's accounts. The comment extraction from social media platforms will employ scraping methods, performed ethically and in

E-mail: christopher.owen001@binus.ac.id, dsuhartono@binus.edu compliance with the platforms' terms of service. This approach allows researchers to swiftly accumulate a substantial amount of data from social media. A total of 25,123 data entries were successfully obtained from these platforms and will be consolidated into a CSV file.

B. Data Pre-processing

After collection, the data will undergo pre-processing. The primary objective of pre-processing is to clean and prepare the data before using it in the model, ensuring that the data applied to the model is of high quality and relevance. The pre-processing process involves several stages, as depicted in Figure 3.

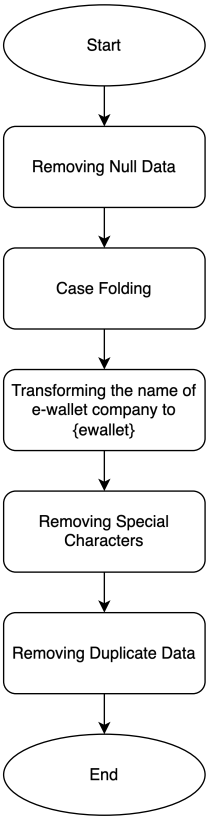

Start
Removing Null Data
Case Folding
Transforming the name of e-wallet company to {ewallet}
Removing Special Characters
Removing Duplicate Data
End

Figure 3. Pre-processing stages

In the process of collecting data, encountering null or empty entries is a common occurrence that can disrupt subsequent model training and analysis stages, potentially introducing unintended biases. To mitigate this, null data will be removed before proceeding, ensuring a more robust and unbiased dataset for analysis and model training. Another critical step in data pre-processing is case folding, where all characters in a text are converted to lowercase to normalize the data, promoting consistency, and reducing irrelevant variations. By transforming the entire text to lowercase, the model can treat words with uppercase or lowercase letters as the same form, thereby enhancing its ability to better understand textual content. Furthermore, e-wallet company names such as DANA, OVO, GoPay, ShopeePay, and LinkAja will be transformed into {ewallet}, with the goal of avoiding errors and ambiguities that may arise from varied company name representations. This process also aids in reducing bias during model training. Subsequently, irrelevant elements like usernames, symbols, punctuation, emojis, and special characters are systematically removed from the dataset, ensuring cleaner data for analysis and intent classification. This step is crucial as these elements often lack significant meaning and can disrupt the model training process. Following data cleaning, the occurrence of duplicate data is addressed to prevent bias in the model and eliminate redundancy in the dataset. The removal of duplicate entries ensures a cleaner, more balanced dataset for model training, free from the influence of excessive repetitions, thereby contributing to the overall effectiveness of the analysis and training processes. Following these pre-processing stages, the dataset was refined, resulting in a reduction to 14,916 records.

C. Data Labelling

Intent classification is a component of supervised learning that requires incorporating labelled data in the dataset to train the model. In the context of this research, the proposed approach involves employing the BERTopic model to assign labels to the dataset. This model involves three key stages: document embeddings, document clustering, and topic representation [18]. The document embeddings stage employs the pre-trained SBERT model to transform existing documents into vector forms, facilitating semantic comparisons. Following this, the document clustering phase groups documents based on their semantic similarities. Dimensionality reduction using the Uniform Manifold Approximation and Projection (UMAP) method is implemented prior to clustering, as it has proven to retain more local and global features compared to other methods, such as PCA and t-SNE. HDBSCAN is then utilized for clustering, employing a soft-clustering approach to minimize errors in distributing clusters and improve accuracy. After obtaining clustering results, the next step is topic representation. In this stage, topics will be assigned to each cluster using the class-based Term Frequency-Inverse Document Frequency (c-TF-IDF) method. This involves calculating the weight or importance level of a word within a specific cluster, rather than across all documents. Initially, all texts within a cluster are merged to be considered as a single document. Subsequently, the weight of term t in a class c can be calculated using the equation:

$$W_{t,c} = tf_{t,c} \times \log(1 + \frac{A}{tf_t}) \qquad (1)$$

TABLE I. Data Labelling Result

Label	Label Name	Top 5 Keywords	Total Data
-1	sydn_kupu_mhmd_love	[‘sydn’, ‘kupu’, ‘mhmd’, ‘love’, ‘not’]	60
0	saya_dm_saldo_di	[‘saya’, ‘dm’, ‘saldo’, ‘di’, ‘ewallet’]	12,398
1	promo_yang_dengan_sesuai	[‘promo’, ‘yang’, ‘dengan’, ‘sesuai’, ‘tidak’]	596
2	makan_orang_jangan_lain	[‘makan’, ‘orang’, ‘jangan’, ‘lain’, ‘org’]	395
3	biar_yuk_viral_bantu	[‘biar’, ‘yuk’, ‘viral’, ‘bantu’, ‘apa’]	361
4	min_kasih_infonya_selamat	[‘min’, ‘kasih’, ‘infonya’, ‘selamat’, ‘selalu’]	338
5	tai_makin_kek_kaya	[‘tai’, ‘makin’, ‘kek’, ‘kaya’, ‘bangsat’]	246
6	dompet_biasa_bukan_digital	[‘dompet’, ‘biasa’, ‘bukan’, ‘digital’, ‘luar’]	205
7	kau_otak_kelen_gak	[‘kau’, ‘otak’, ‘kelen’, ‘gak’, ‘pak’]	170
8	aku_ra_ku_for	[‘aku’, ‘ra’, ‘ku’, ‘for’, ‘job’]	147

Here, $tf_{t,c}$ represent the term frequency of term t in class c and $\log\left(1+\frac{A}{tf_t}\right)$ denotes the inverse class frequency of term t. In this calculation, inverse document frequency is replaced with inverse class frequency to measure how much information a term provides within a class. The inverse class frequency is computed using the logarithm of the average word count per class A divided by the frequency of term t across all classes. To ensure a positive result, one is added to the division within the logarithm. After identifying the most influential words, a certain number of these impactful words are combined into a sentence to generate a topic that represents the document. The labelling results are presented in Table I, revealing the presence of 10 distinct label types in the dataset, with each label name derived from the four keywords that best encapsulate its content.

D. Model Development

After completing the pre-processing and labelling stages of the dataset, the next step involves constructing a model to classify user intentions in e-wallet services. The model development process consists of two stages: dataset splitting and model training.

In the dataset splitting stage, the dataset will be divided into three parts: the training dataset, validation dataset, and testing dataset. The objective is to prevent overfitting in the model under development. The dataset will be divided using the stratify split method, ensuring that the distribution of classes or target variables remains proportional between the resulting sets [19]. Initially, the dataset will be split into 80% training data and 20% testing data [20]. Subsequently, the training dataset will be further divided into 80% training data and 20% validation data. The training dataset will be utilized to train the model, enabling it to recognize patterns within the data. The validation dataset will be employed to determine hyperparameters in the model and assess the model's performance during the training process. Meanwhile, the testing dataset will be used to evaluate the model after the training process is complete, aiming to test the accuracy of the model in predicting previously unknown data.

To train the model, the divided dataset will undergo further processing using the embedding method. Several combinations of embedding methods and classification models will be experimented with to generate a model capable of classifying user intents in the Indonesian language. This model will be specifically applied to artificial intelligence-based customer service for e-wallet service providers. Four types of models will be employed, namely IndoBERT, RoBERTa, TextConvoNet, and a model with a proposed method.

1) IndoBERT

The first model to be utilized is IndoBERT, an Indonesian version of BERT trained as a masked language model using the HuggingFace framework. IndoBERT has been trained on over 220 million words from various sources, including Indonesian Wikipedia, news articles from Kompas, Tempo, and Liputan6, as well as the Indonesian Web Corpus [21]. This model employs contextualized embeddings, employing the WordPiece tokenization method to break down words into smaller pieces or subwords in the dataset. Special tokens such as [CLS] (classifier token) at the beginning of a sentence, [SEP] (separation token) at the end of a sentence, and [PAD] (padding token) are added. Following tokenization, encoding takes place where all tokens are converted into unique integer IDs based on a predefined vocabulary. After encoding, various embeddings are incorporated to furnish insights into the intricate relationships among different segments within a text. Segment embeddings enhance comprehension of text segment relationships, while position embeddings convey crucial information about token position and sequence within a sentence. Moreover, attention mask embeddings discern between [PAD] tokens and others, contributing to the model's precision. Following the completion of the embedding process, the subsequent stage involves the training of the IndoBERT model, with a specific focus on fine-tuning the pre-trained "indobenchmark/indobert-large-p2" model [22].

2) *RoBERTa*

The next model to be employed is the RoBERTa model. RoBERTa (A Robustly Optimized BERT Pretraining Approach) is a transformer-based model that is a modification of the BERT model. This modification eliminates Next Sentence Prediction (NSP), removing the need for additional pre-training, and the model achieves better results [23]. RoBERTa also dynamically replaces the masking patterns in the training data, trains the model for a longer duration, uses larger batches, longer sequences, and utilizes more data. Like the BERT model, RoBERTa also employs embedding methods and transformer-based classification models. The embedding method used is contextualized embeddings, involving tokenization with Byte-level Byte-Pair Encoding (BPE) tokenization to break words in the dataset into byte form, then forming subwords based on the combination of the most frequently occurring byte fragments. Following this, several special tokens are added: the <s> token to indicate the beginning of a sentence and serve as a classifier token, the </s> token to separate sentences and indicate the end of a sentence, and the <pad> token (padding token) used to adjust the length of sentences shorter than the predefined length. Subsequently, an encoding process takes place where all tokens are converted into unique integer IDs based on a predefined vocabulary. After the encoding process, position embeddings layers are added to provide information about the position and sequence of tokens in a sentence, and attention mask embeddings layers are added to distinguish between the <pad> token and other tokens. After the embedding process is complete, the next step involves training the RoBERTa model using the resulting embeddings. Fine-tuning will be conducted on the pre-trained RoBERTa model, namely "flax-community/indonesian-roberta-base" which has been trained using the OSCAR dataset with a TPUv3-8 VM.

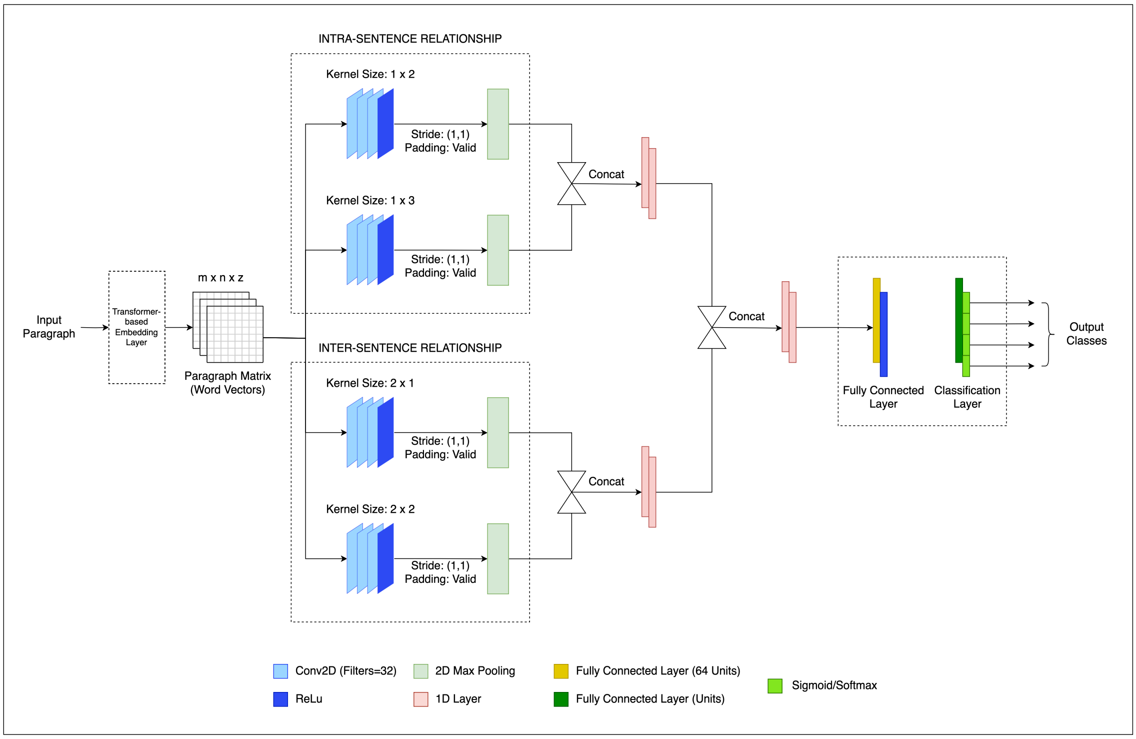


Figure 4. Proposed model architecture

3) *TextConvoNet*

The next model to be employed is the TextConvoNet model, an architecture based on Convolutional Neural Network (CNN) used for text classification. This architecture utilizes 2-dimensional convolutional filters to extract n-gram features within a sentence (intra-sentence) and between sentences (inter-sentence) in the text data. The model incorporates Word2Vec embedding and a CNN-based classification model. Through Word2Vec embedding, words in the dataset are transformed into their vector representations without considering context and sequence. The resulting embeddings are used to train the CNN-based classification model. The classification model consists of four convolutional layers, with the first two layers having 32 filters each and sizes of 1x2 and 1x3. The outputs of these two layers are combined to extract intra-sentence n-gram features. Meanwhile, the other two layers also have 32 filters each, with sizes of 2x1 and 2x2. The outputs of these two layers are combined to extract inter-sentence n-gram features. The results from both intra-sentence and inter-sentence layers are merged and fed into a fully connected layer with 64 neurons for dataset classification [15].

4) *Proposed Model*

The final model to be employed is a proposed method utilizing transformer-based embedding and a CNN-based classification model, as illustrated in Figure 4. This model is an adaptation of the TextConvoNet model, wherein TextConvoNet utilizes Word2Vec embedding that fails to capture context and word sequence in a dataset during embedding. This results in the representation vector of a word remaining constant. Therefore, the use of transformer-based embedding methods will be explored to capture context and sequence during embedding. Two transformer-based models, namely IndoBERT and SBERT, will be tested for embedding. These models have undergone pretraining on Indonesian language datasets to gain a comprehensive understanding of the language's nuances and context. This pretraining ensures that the models capture diverse information based on word context and sequence, resulting in variable representation vectors. Once the embedding process is complete, the results will be fed into the CNN-based classification model with the same architecture as TextConvoNet.

4. Results and Discussion

All models underwent training under identical conditions, with parameters set at 10 epochs, a batch size of 16, and employing two distinct learning rates: 0.001 and 0.0003. In evaluating the trained models, various metrics, including accuracy, precision, recall, and F1-score, were employed, utilizing weighted averaging to address imbalanced label distributions within the dataset. The training results are presented in the Table II.

In the evaluation with a learning rate of 0.001, the combination of IndoBERT with the TextConvoNet (IndoBERT + TextConvoNet) model exhibited notable performance improvements over the baseline IndoBERT model, with enhancements of 3.42% in accuracy, 15.50% in precision, 3.42% in recall, and 9.21% in F1-score. Similarly, compared to RoBERTa, the IndoBERT + TextConvoNet model demonstrated improvements of 1.17% in accuracy, 2.45% in precision, 1.17% in recall, and 1.93% in F1-score. Moreover, compared to the original TextConvoNet model, the IndoBERT + TextConvoNet model manifested enhancements of 0.23% in accuracy, 0.56% in precision, 0.23% in recall, and 0.06% in F1-score.

TABLE II. Comparison of Model Performance in Classifying the Dataset

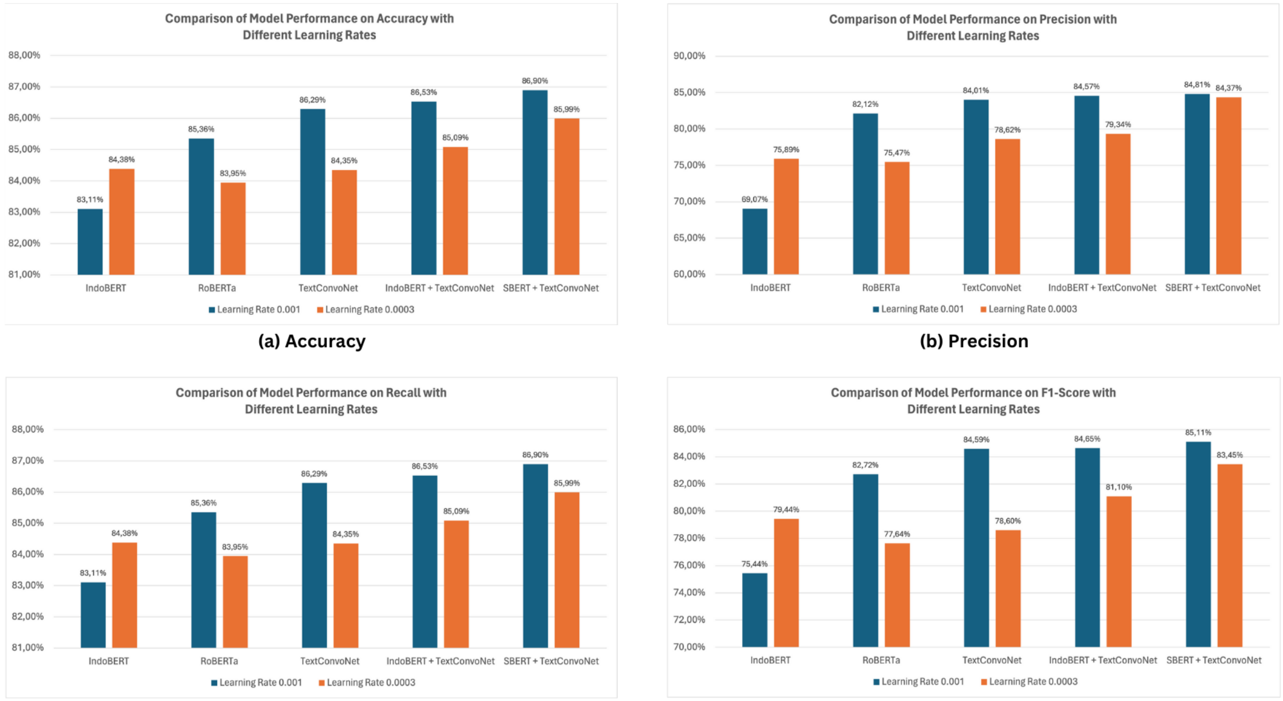
Model	Learning Rate	Accuracy	Precision	Recall	F1-score
IndoBERT	0.001	83.11%	69.07%	83.11%	75.44%
RoBERTa		85.36%	82.12%	85.36%	82.72%
TextConvoNet		86.29%	84.01%	86.29%	84.59%
IndoBERT + TextConvoNet		86.53%	84.57%	86.53%	84.65%
SBERT + TextConvoNet		**86.90%**	**84.81%**	**86.90%**	**85.11%**
IndoBERT	0.0003	84.38%	75.89%	84.38%	79.44%
RoBERTa		83.95%	75.47%	83.95%	77.64%
TextConvoNet		84.35%	78.62%	84.35%	78.60%
IndoBERT + TextConvoNet		85.09%	79.34%	85.09%	81.10%
SBERT + TextConvoNet		85.99%	84.37%	85.99%	83.45%

Likewise, the combination of SBERT with the TextConvoNet (SBERT + TextConvoNet) model surpassed all baseline models when trained with the same learning rate, exhibiting significant enhancements over IndoBERT, RoBERTa, and the original TextConvoNet in all metrics. Specifically, compared to IndoBERT, the SBERT + TextConvoNet model showcased an improvement of 3.79% in accuracy, 15.74% in precision, 3.79% in recall, and 9.66% in F1-score. Similarly, compared to RoBERTa, the SBERT + TextConvoNet model exhibited an improvement of 1.54% in accuracy, 2.70% in precision, 1.54% in recall, and 2.39% in F1-score. Additionally, compared to the original TextConvoNet, the SBERT + TextConvoNet model demonstrated enhancements of 0.60% in accuracy, 0.80% in precision, 0.60% in recall, and 0.52% in F1-score.

Similar results can also be observed when the existing models are trained with a lower learning rate of 0.0003, where all suggested techniques also outperform IndoBERT, RoBERTa, and the original TextConvoNet model in accuracy, precision, recall, and F1-score metrics. Notably, when compared with the original TextConvoNet model, IndoBERT + TextConvoNet showcased enhancements of 0.74% in accuracy, 0.72% in precision, 0.74% in recall, and 2.50% in F1-score. Similarly, SBERT + TextConvoNet exhibited enhancements of 1.64% in accuracy, 5.75% in precision, 1.64% in recall, and 4.85% in F1-score.

Based on the training results, the model utilizing SBERT embedding with TextConvoNet classification achieved the highest scores across all metrics, with accuracy, precision, recall, and F1-score reaching 86.90%, 84.81%, 86.90%, and 85.11%, respectively, using a learning rate of 0.001. Consequently, it can be concluded from the training evaluation that SBERT embedding with TextConvoNet classification models are superior in classifying user intention within the dataset, surpassing both the original TextConvoNet model and traditional transformer-based models like IndoBERT and RoBERTa.

Furthermore, the impact of learning rate selection on model training outcomes is prominently illustrated in Figure 5. It is evident that the choice of learning rate significantly influences model performance. While a learning rate of 0.001 proved to be optimal for the majority of models evaluated in this study, it is notable that the IndoBERT model demonstrates superior performance with a lower learning rate of 0.0003. This superiority is reflected in the higher values of accuracy, precision, recall, and F1-score achieved when using a learning rate of 0.0003 compared to 0.001. This indicates that the ideal learning rate may differ based on variables like model structure, dataset attributes, and training goals. These findings underscore the importance of hyperparameter tuning in achieving optimal model performance and generalization.



(a) Accuracy

(b) Precision

(c) Recall

(d) F1-score

Figure 5. Comparison of model performance with different learning rates on various metrics. (a) Accuracy. (b) Precision. (c) Recall. (d) F1-score.

5. Conclusion and Future Works

This study presents a comparative analysis of various models for classifying user intents in Indonesian, with a specific focus on AI-based customer service in e-wallet services. Two novel approaches, utilizing transformer-based embedding techniques namely SBERT and IndoBERT in combination with the TextConvoNet classification approach, were evaluated alongside conventional transformer-based models such as IndoBERT and RoBERTa, as well as the original TextConvoNet model. The dataset employed for training encompasses user comments in the Indonesian language from the social media platforms of the top five prominent e-wallet service providers in Indonesia, namely GoPay, DANA, OVO, ShopeePay, and LinkAja.

Some conclusions can be drawn from this study:

a. The proposed model, leveraging SBERT embeddings combined with the TextConvoNet classification approach, outperforms other models in terms of accuracy, precision, recall, and F1-score, achieving remarkable results of 86.90%, 84.81%, 86.90%, and 85.11%, respectively, with a learning rate set at 0.001. These findings emphasize the superiority of the proposed model in precisely classifying user intention within the dataset, outperforming both conventional transformer-based models like IndoBERT and RoBERTa, and the original TextConvoNet model.

b. This study emphasizes the crucial influence of learning rate selection on model training outcomes. Although a learning rate of 0.001 was found to be optimal for most models examined in this study, it is noteworthy that the IndoBERT model exhibited better performance with a reduced learning rate of 0.0003. This underscores the importance of customizing hyperparameter tuning to optimize model performance.

To evaluate the model's performance in practical situations, further research will aim to integrate the developed model with the customer service chatbot in e-wallet services. Additionally, advanced methods will be explored for more accurate and consistent labelling of the dataset.

Acknowledgment

This research was supported by Bina Nusantara University. The authors also appreciate the valuable feedback and suggestions provided by the reviewers, which have enhanced the clarity of the presentation.

References

[1] A. Sikri, S. Dalal, N. P. Singh, and D.-N. Le, “Mapping of e-wallets with features,” in *Cyber Security in Parallel and Distributed Computing*, John Wiley & Sons, Ltd, 2019. doi: https://doi.org/10.1002/9781119488330.ch16.

[2] Populix, “Consumer preference towards banking and e-wallet apps,” 2022. [Online]. Available: https://info.populix.co/articles/report/digital-banking-survey/

[3] B. Luo, R. Y. K. Lau, C. Li, and Y.-W. Si, “A critical review of state-of-the-art chatbot designs and applications,” *WIREs Data Mining and Knowledge Discovery*, vol. 12, no. 1, 2022, doi: https://doi.org/10.1002/widm.1434.

[4] C. V. Misischia, F. Poecze, and C. Strauss, “Chatbots in customer service: their relevance and impact on service quality,” *Procedia Comput Sci*, vol. 201, no. C, 2022, doi: 10.1016/j.procs.2022.03.055.

[5] E. Adamopoulou and L. Moussiades, “An overview of chatbot technology,” in *Artificial Intelligence Applications and Innovations*, I. Maglogiannis, L. Iliadis, and E. Pimenidis, Eds., Cham: Springer International Publishing, 2020.

[6] J. J. Y. Zhang, A. Følstad, and C. A. Bjørkli, “Organizational factors affecting successful implementation of chatbots for customer service,” *Journal of Internet Commerce*, vol. 22, no. 1, 2023, doi: 10.1080/15332861.2021.1966723.

[7] I. K. F. Haugeland, A. Følstad, C. Taylor, and C. A. Bjørkli, “Understanding the user experience of customer service chatbots: an experimental study of chatbot interaction design,” *Int J Hum Comput Stud*, vol. 161, 2022, doi: https://doi.org/10.1016/j.ijhcs.2022.102788.

[8] C.-H. Li, S.-F. Yeh, T.-J. Chang, M.-H. Tsai, K. Chen, and Y.-J. Chang, “A conversation analysis of non-progress and coping strategies with a banking task-oriented chatbot,” in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, in CHI ’20. New York, NY, USA: Association for Computing Machinery, 2020. doi: 10.1145/3313831.3376209.

[9] W. Maeng and J. Lee, “Designing a chatbot for survivors of sexual violence: exploratory study for hybrid approach combining rule-based chatbot and ml-based chatbot,” in *Asian CHI Symposium 2021*, in Asian CHI Symposium 2021. New York, NY, USA: Association for Computing Machinery, 2021. doi: 10.1145/3429360.3468203.

[10] B. Jang, M. Kim, G. Harerimana, S. Kang, and J. W. Kim, “Bi-lstm model to increase accuracy in text classification: combining word2vec cnn and attention mechanism,” *Applied Sciences*, vol. 10, no. 17, 2020, doi: 10.3390/app10175841.

[11] S. Rizou *et al.*, “Efficient intent classification and entity recognition for university administrative services employing deep learning models,” *Intelligent Systems with Applications*, vol. 19, 2023, doi: 10.1016/j.iswa.2023.200247.

[12] F. Fatharani, K. P. Kania, J. Hutahaean, and S. R. Wulan, “Deteksi intensi chatbot berbahasa indonesia dengan menggunakan metode capsule network,” *Journal of Information System Research (JOSH)*, vol. 3, no. 4, Jul. 2022, doi: 10.47065/josh.v3i4.1821.

[13] L. P. Manik, “Out-of-scope intent detection on a knowledge-based chatbot,” *International Journal of Intelligent Engineering and Systems*, vol. 14, 2021, doi: 10.22266/ijies2021.1031.39.

[14] T. N. R. Kumar, G. Shidaganti, P. Anand, S. Singh, and S. Salil, “Analyzing and automating customer service queries on twitter using robotic process automation,” *Journal of Computer Science*, vol. 19, no. 4, Mar. 2023, doi: 10.3844/jcssp.2023.514.525.

[15] S. Soni, S. S. Chouhan, and S. S. Rathore, “TextConvoNet: a convolutional neural network based architecture for text classification,” *Applied Intelligence*, vol. 53, no. 11, pp. 14249–14268, 2023, doi: 10.1007/s10489-022-04221-9.

[16] J. Kapočiute-Dzikiene, “Intent detection-based lithuanian chatbot created via automatic dnn hyper-parameter optimization,” *Frontiers in Artificial Intelligence and Applications*, vol. 328, 2020, doi: 10.3233/faia200608.

[17] J. Huang, Y.-R. Liou, and H.-H. Chen, “Enhancing intent detection in customer service with social media data,” in

Companion Proceedings of the Web Conference 2021, in WWW '21. New York, NY, USA: Association for Computing Machinery, 2021. doi: 10.1145/3442442.3451377.

[18] M. Grootendorst, "BERTopic: neural topic modeling with a class-based tf-idf procedure." 2022. [Online]. Available: http://arxiv.org/abs/2203.05794

[19] M. Merrillees and L. Du, "Stratified sampling for extreme multi-label data," in *Advances in Knowledge Discovery and Data Mining*, K. Karlapalem, H. Cheng, N. Ramakrishnan, R. K. Agrawal, P. K. Reddy, J. Srivastava, and T. Chakraborty, Eds., Cham: Springer International Publishing, 2021, pp. 334–345.

[20] V. R. Joseph, "Optimal ratio for data splitting," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 15, no. 4, 2022, doi: https://doi.org/10.1002/sam.11583.

[21] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "Indolem and indobert: a benchmark dataset and pre-trained language model for indonesian nlp," in *COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference*, Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020. doi: 10.18653/v1/2020.coling-main.66.

[22] S. Cahyawijaya *et al.*, "Indonlg: benchmark and resources for evaluating indonesian natural language generation," in *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, K.-F. Wong, K. Knight, and H. Wu, Eds., Suzhou, China: Association for Computational Linguistics, Dec. 2021, pp. 8875–8898. doi: 10.18653/v1/2021.emnlp-main.699.

[23] M. A. Aleisa, N. Beloff, and M. White, "Implementing airm: a new ai recruiting model for the saudi arabia labour market," *J Innov Entrep*, vol. 12, no. 1, p. 59, 2023, doi: 10.1186/s13731-023-00324-w.

Christopher Owen obtained a Bachelor's degree from Bina Nusantara University in Jakarta, Indonesia, in 2022, followed by a Master's degree from the same university in 2024. His research interests include text analysis and artificial intelligence.

Derwin Suhartono received the Ph.D. degree in computer science from Universitas Indonesia, in 2018. He is currently a Faculty Member of Bina Nusantara University, Indonesia. His research interest includes natural language processing. Recently, he is continually doing research in argumentation mining and personality recognition. He actively involves in Indonesia Association of Computational Linguistics (INACL), a national scientific association in Indonesia, IndoCEISS, and Aptikom. He has his professional memberships in IEEE, ACM, INSTICC, and IACT. He also takes role as reviewer in several international conferences and journals.