Comparative Analysis of Intent Classification in Indonesian Chatbots Using BERT and RoBERTa Models

Abdiansah Abdiansah
Fac. of Com. Sci. Univ. Sriwijaya
AIRLab Research Group
South-Sumatera, Indonesia
abdiansah@unsri.ac.id

Muhammad Fachrurrozi
Fac. of Com. Sci. Univ. Sriwijaya
AIRLab Research Group
South-Sumatera, Indonesia
mfachrz@unsri.ac.id

Aswin Dwiyono
Fac. of Com. Sci. Univ. Sriwijaya
AIRLab Research Group
South-Sumatera, Indonesia
aswindwiyono@gmail.com

2024 Ninth International Conference on Informatics and Computing (ICIC) | 979-8-3315-1760-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICIC64337.2024.10957286

***Abstract*—A chatbot is a software application to designed handle user inputs and generate appropriate replies based on those inputs, which are then communicated back to the user. To ensure effective response from users, need to correctly interpret the intent chatbot. This involves identifying the underlying meaning of the text provided by the user, allowing the chatbot to deliver relevant responses. This paper compares two state-of-the-art transformer-based models—BERT (Bidirectional Encoder Representations from Transformers) and RoBERTa (Robustly Optimized BERT Pretraining Approach)—for the task of intent classification in chatbot systems. Various performance metrics, including accuracy, F1-score, precision, and recall, were analyzed to determine which model performs more effectively under different conditions. Performance metrics like accuracy and F1-score were compared to assess model BERT and RoBERTa performs better in a University Chatbot Dataset in Indonesian language. The BERT model achieved an accuracy of 0.89, outperforming the RoBERTa model, which achieved 0.84.**

Keywords— Intent classification, BERT, RoBERTa, Chatbots, Transformer models, Natural Language Processing

I. Introduction

Chatbots are becoming an indispensable component of human-machine interaction in today's increasingly sophisticated digital age. They are used across various sectors such as banking, healthcare, and education, providing automated support and responsive interaction. One of the key elements in developing an effective chatbot is the ability to accurately understand and classify user intent. Intent classification is a crucial first step in guiding the conversation and understanding its context to deliver relevant and accurate responses.

There are several methods for performing intent classification in chatbots. Conventional approaches often rely on rule-based systems, but these methods are frequently inadequate when facing new or changing contexts that cannot be captured by pre-set rules, requiring continuous updates to adapt to new scenarios [1], [2]. Additionally, rule-based systems are inflexible [3], involve high development and maintenance costs due to the need for manual updates [4], and are less robust to variations in user input, often failing to recognize intent in cases of misspellings, slang, or acronyms [5].

The emergence of deep learning techniques, particularly transformer-based models like BERT (Bidirectional Encoder Representations from Transformers) and RoBERTa (Robustly Optimized BERT Pretraining Approach), has addressed many of these challenges. These models leverage deep learning to capture semantic relationships in text, improving a chatbot's ability to discern user intent. BERT has been widely adopted for intent classification due to its ability to understand context by processing text bidirectionally, offering high accuracy in intent classification tasks [6], [7], [8]. However, while BERT and similar models provide state-of-the-art capabilities, they are not without limitations, such as biases that can favor certain classes over others [9], and issues related to sub-tokenization, which can result in misalignments that affect the accuracy of label predictions [10].

Despite BERT's effectiveness, Facebook AI's development of RoBERTa has demonstrated further potential for improvement in the pretraining and data processing of BERT. Unlike BERT, RoBERTa removes the Next Sentence Prediction (NSP) objective, simplifying the training process and allowing the model to focus more on understanding context within individual sentences, leading to improved accuracy in intent classification [11], [12]. RoBERTa's architecture also enables it to capture patterns and fine-grained relationships in the data, which is crucial for effective intent recognition, particularly in multilingual and heterogeneous datasets [13].

In comparing the performance of BERT and RoBERTa for chatbot intent classification, both models demonstrate unique strengths. BERT, with its bidirectional approach and strong contextual modelling, delivers satisfactory results in natural language tasks. However, RoBERTa, with its optimized pretraining process, proves that further improvements can enhance the accuracy and efficiency of intent classification. This research aims to analyse the performance of BERT and RoBERTa in intent classification tasks using an Indonesian dataset. The results are expected to contribute valuable insights into the application of deep learning-based models for more intelligent and accurate automated conversation systems or chatbots.

II. Research Method

A. Intent Classification for Chatbots

A chatbot is a program that receives input from users and generates responses by matching the input with corresponding answers [14], [15], [16]. Intent classification is a core component of any chatbot system, enabling the chatbot to understand and interpret user questions by categorizing them into predefined intents [17], [18], [19],

[20]. For a successful conversation between the user and the chatbot, the user's intent must be accurately classified.

B. Dataset

The dataset used in this research is sourced from Kaggle, specifically the University Chatbot Dataset [21]. The file contains 38 intents, also referred to as tags. This dataset can be used for training and evaluating chatbot models. Table I shows an example of an intent dataset in English.

TABLE I. UNIVERSITY CHATBOT DATASET IN ENGLISH

Tag	Pattern	Response
Greeting	- Hi - How are you? - Is anyone there?	- Hello - Good to see you again! - Hi there, how can I help?
Goodbye	- See you - Goodbye	- Sad to see you go :(- Talk to you later

Furthermore, in this research, the dataset was translated into Bahasa Indonesia using DeepL [22], with additional human assistance. The results are shown in Table II.

TABLE II. UNIVERSITY CHATBOT DATASET IN BAHASA INDONESIAN

Tag	Pattern	Response
Salam	- Hai - Apa kabar? - Apakah ada orang di sana?	- Halo - Senang bertemu dengan Anda lagi! - Hai, ada yang bisa saya bantu?
Perpisahan	- Sampai jumpa - Selamat tinggal	- Sedih melihatmu pergi :(- Sampai jumpa lagi nanti

Table III presents the intent data used in this research. The table consists of three columns: (1) Intent, which contains 38 intent categories related to the university chatbot; (2) Count Pattern, which indicates the number of patterns for each intent; and (3) Count Response, which shows the number of responses for each intent. The total number of intent patterns is 405, and the total number of responses is 47.

TABLE III. INDONESIAN TRANLATED INTENTS

Intent	Count Pattern	Count Response
Salam (Greeting)	10	3
Perpisahan (Goodbye)	12	4
Pencipta (Creator)	16	1
Nama (Name)	13	4
Jam (Hours)	17	1
Nomor (Number)	15	1
Jurusan (Course)	27	1
Biaya (Fees)	23	1
Lokasi (Location)	14	1
Asrama (Hostel)	22	1
Acara (Event)	11	1
Dokumen (Document)	13	1
Lantai (Floors)	7	1
Silabus (Syllabus)	7	1
Perpustakaan (Library)	14	1
Infrastruktur (Infrastructure)	3	1
Kantin (Canteen)	11	1
Menu (Menu)	7	1
Karir (Placement)	9	1
Kepala Jurusan (Head of Department)	4	1
Kepala Prodi (Head of Study Program)	4	1
Sekretaris Jurusan (Department Secretary)	4	1
Rektor (Rector)	7	1
Semester (Semester)	11	1
Penerimaan (Admission)	6	1
Beasiswa (Scholarship)	26	1
Fasilitas (Facilities)	5	1
PMB (College Intake)	9	1
Seragam (Uniform)	9	1
Komite (Committee)	6	1
Acak (Random)	3	1
Umpat (Swear)	9	1
Liburan (Vacation)	12	1
Olahraga (Sports)	7	1
Salut (Salutaion)	13	1
Tugas (Task)	6	2
Pelonco (Ragging)	10	1
Dekan (Dean)	3	1

C. BERT and RoBERTa

BERT revolutionized natural language processing (NLP) by introducing a bidirectional transformer-based approach, significantly improving performance in downstream NLP tasks such as intent classification [23]. BERT is a pre-trained contextual word representation model based on the Masked Language Model (MLM) and utilizes bidirectional transformers. Its architecture consists of a multi-layer bidirectional transformer encoder-decoder structure. Transformers use stacked self-attention mechanisms and point-wise, fully connected encoders and decoders [24].

RoBERTa builds upon BERT by refining training techniques and using larger datasets [11]. Some key modifications made to enhance BERT's performance include the following [25]:

1. Train earlier models with larger datasets. While the BERT pre-trained model was trained on just 13GB of data, RoBERTa was trained on a significantly larger dataset of up to 160GB, which has been shown to improve accuracy.
2. Removed the next sentence prediction (NSP) objective.

The NSP (Next Sentence Prediction) objective trains the model to determine whether two sentences are related. In the RoBERTa model, this objective is removed to enhance performance in downstream tasks.

3. Training on longer sequences. The BERT model is trained on sequences of 256 sequences with 1M steps, whereas the RoBERTa model is trained on sequences of up to 8K sequences and 500K sequences.
4. Continuously vary the masking pattern in the training data.

D. Experimental Setup

Table IV displays various model parameter configurations used in this experiment that BERT and RoBERTa model. The table consists of two columns: (1) Parameters, which contains parameters used in the experiment; and (2) Values, which value of parameters used in the experiment. Num_train_epoch represents the number of epochs, indicating how many times the model will process the entire dataset during training. In this case, the model will go through the dataset 100 times during the training process. Per_device_train_batch_size refers to the batch size for training, indicating the number of samples the model processes in one iteration on each device. In this case, the model will process 32 samples at a time per training iteration. Per_device_eval_batch_size refers to the batch size used during evaluation when the model validates its performance on the validation data. A value of 16 means that 16 samples are evaluated at a time. Warmup_steps refer to the number of 'warm-up' steps during training. These steps help prevent abrupt changes at the beginning of training, which could make it difficult for the model to learn effectively. Weight_decay is a regularization parameter that helps prevent overfitting by gradually reducing the weight values over time. A value of 0.05 indicates that the weights will be reduced by 5% with each update to prevent them from becoming too large. Logging_steps define the frequency of logging during training. Every 50 training steps, the process will generate a log, providing information such as the current loss or accuracy. Evaluation_strategy defines the strategy used for evaluation during training. When set to 'step', the model will be evaluated every few steps rather than only at the end of each epoch. Eval_steps determine how often evaluation is performed during training. When set to 50, the evaluation will be run every 50 training steps. These variables were chosen because they have a significant influence on the performance of the BERT model on the chatbot intent classification task.

TABLE IV. PARAMETERS USED IN THE EXPERIMENT

Parameters	Values
Num_train_epoch	100
Per_device_train_batch_size	32
Per_device_eval_batch_size	16
Warmup_steps	100
Weight_decay	0.05
Logging_steps	50
Evaluation_strategy	step
Eval_steps	50

The experiments were implemented using Python and trained on Google Colab Pro A100 GPU.

E. Evaluation Metrics

To evaluate the classification performance of our model, we selected four commonly used metrics in classification tasks: Precision (P), Recall (R), F1-score (F1), and Accuracy (Acc). Higher scores indicate better classification performance. The calculations are shown in Equations (1)–(4). TP represents the number of correctly predicted samples, FP represents the number of incorrectly predicted samples, FN is the number of samples incorrectly classified into other categories, and TN is the number of samples correctly classified into other categories.

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

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

$$\textit{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

$$\textit{F1-Score} = \frac{2 * (\textit{Precision} * \textit{Recall})}{\textit{Precision} + \textit{Recall}} \quad (4)$$

III. RESULTS AND DISCUSSIONS

A. Result of BERT

After training, the model was evaluated using the test dataset. The evaluation metrics included accuracy (Acc.), F1-score (F1), precision (Prec.), and recall. The results of training and testing the BERT model are shown in Table V.

TABLE V. RESULT TRAIN AND TEST MODEL BERT

	Loss	Acc.	F1	Prec.	Recall
Train	0.015	0.997	0.998	0.998	0.999
Test	0.604	0.892	0.875	0.892	0.886

As shown in Table V, the model achieved high accuracy on the training data (0.99), but the accuracy on the testing data was lower (0.89). This indicates a bit of overfitting, where the model memorizes too many patterns in the training data, making it less capable of generalizing to new data. Similarly, the F1-score was 0.99 on the training data but dropped to 0.87 on the test data.

Fig. 1 shows the confusion matrix from the BERT model applied to the chatbot intent classification task. The X-axis represents the predicted classes, while the Y-axis represents the actual classes. The value in each cell indicates the number of samples classified into a specific class. From Fig. 1, it is evident that the BERT model performs well in classifying the "asrama" (hostel) and "beasiswa" (scholarship) intents. However, it frequently misclassifies the "kantin" (canteen) intent as "menu" (menu), indicating that the model struggles to differentiate between these two intents.

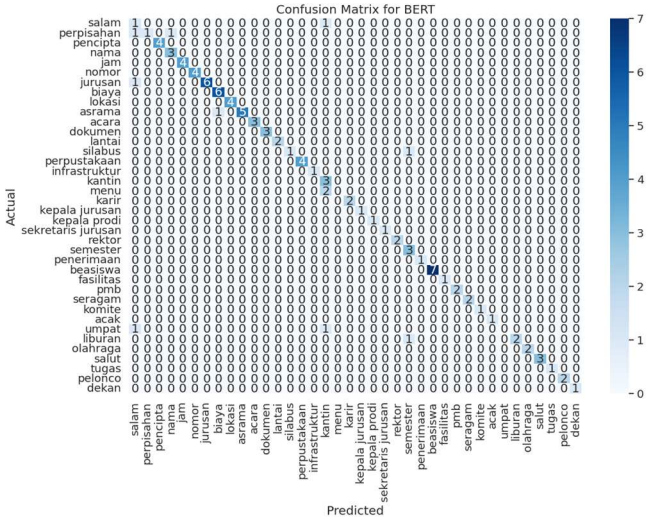

Fig. 1. Confusion Matrix Results of BERT Experiments

	precision	recall	f1-score	support
salam	0.25	0.50	0.33	2
perpisahan	1.00	0.33	0.50	3
pencipta	1.00	1.00	1.00	4
nama	0.75	1.00	0.86	3
jam	1.00	1.00	1.00	4
nomor	1.00	1.00	1.00	4
jurusan	1.00	0.86	0.92	7
biaya	0.86	1.00	0.92	6
lokasi	1.00	1.00	1.00	4
asrama	1.00	0.83	0.91	6
acara	1.00	1.00	1.00	3
dokumen	1.00	1.00	1.00	3
lantai	1.00	1.00	1.00	2
silabus	1.00	0.50	0.67	2
perpustakaan	1.00	1.00	1.00	4
infrastruktur	1.00	1.00	1.00	1
kantin	0.43	1.00	0.60	3
menu	0.00	0.00	0.00	2
karir	1.00	1.00	1.00	2
kepala jurusan	1.00	1.00	1.00	1
kepala prodi	1.00	1.00	1.00	1
sekretaris jurusan	1.00	1.00	1.00	1
rektor	1.00	1.00	1.00	2
semester	0.60	1.00	0.75	3
penerimaan	1.00	1.00	1.00	1
beasiswa	1.00	1.00	1.00	7
fasilitas	1.00	1.00	1.00	1
pmb	1.00	1.00	1.00	2
seragam	1.00	1.00	1.00	2
komite	1.00	1.00	1.00	1
acak	1.00	1.00	1.00	1
umpat	0.00	0.00	0.00	2
liburan	1.00	0.67	0.80	3
olahraga	1.00	1.00	1.00	2
salut	1.00	1.00	1.00	3
tugas	1.00	1.00	1.00	1
pelonco	1.00	1.00	1.00	2
dekan	1.00	1.00	1.00	1
accuracy			0.89	102
macro avg	0.89	0.89	0.88	102
weighted avg	0.90	0.89	0.88	102

Fig. 2. Metric Measurement for Each Intent in BERT

Next, we can calculate evaluation metrics such as accuracy, precision, recall, and F1-score for each intent. These metrics provide a more quantitative picture of model performance which can be seen in Fig. 2. Most intents have precision, recall, and F1-scores = 1.00, indicating excellent classification for these categories, such as pencipta (creator), jam (hours), nomor (number), lokasi (location), acara (event), dokumen (document), lantai (floors), perpustakaan (document), infrastruktur (infrastructure), karir (placement), kepala jurusan (head of department), kepala prodi (head of study program), sekretaris jurusan (department secretary), rektor (rector), penerimaan (admission), beasiswa (scholarship), fasilitas (facilities), pmb (college intake), seragam (uniform), komite (committee), acak (random), olahraga (sports), salut (salutaion), tugas (task), pelonco (ragging), dekan (dean). Experimental results to evaluate chatbot intent classification models using the BERT method show an accuracy of 0.89, as shown in Fig. 2.

B. *Result of RoBERTa*

After training, the model was evaluated using the test dataset. The evaluation included calculating accuracy (Acc.), F1-score (F1), precision (Prec.), and recall. The results of training and testing the RoBERTa model are shown in Table VI.

TABLE VI. RESULT TRAIN AND TEST MODEL ROBERTA

	Loss	***Acc.***	***F1***	***Prec.***	***Recall***
Train	0.009	0.997	0.998	0.999	0.998
Test	0.765	0.843	0.779	0.775	0.808

It can be shown in Table VI that the model was able to achieve high accuracy on the training data (0.99), but the accuracy on the testing data was lower (0.84). This indicates a bit of overfitting, where the model memorizes too many patterns in the training data so that it is not capable of generalizing to new data. Also, in the F1-Score value, a value of 0.99 was obtained in the training data but a value of 0.77 was obtained in the test data.

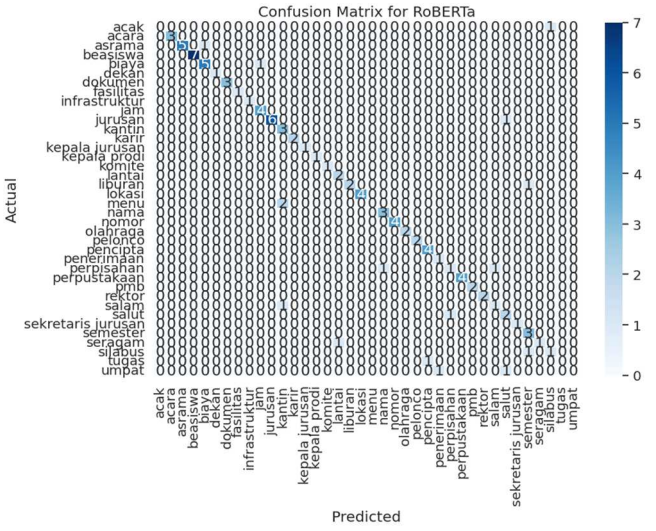

Fig. 3. Confusion Matrix Results of RoBERTa Experiments

Fig. 3 displays the confusion matrix from the RoBERTa model in the chatbot intent classification task. The X-axis represents the class predicted by the model, while the Y-axis represents the actual class. The value in each cell indicates the number of samples classified into a particular class. From Fig. 3 above, it can be seen that the RoBERTa model has quite good performance in classifying "beasiswa" (scholarship) and "asrama" (hostel) intents. However, the model still frequently misclassifies the intent "kantin" (canteen) as "menu" (menu). This indicates that the model has difficulty distinguishing between the two intents.

Next, we can calculate evaluation metrics like accuracy, precision, recall, and F1-score for each intent using RoBERTa. These metrics provide a more quantitative picture of model performance which can be seen in Fig. 4. Most intents have precision, recall, and F1-scores = 1.00, indicating excellent classification for these categories, such as acara (event), beasiswa (scholarship), dekan (dean), dokumen (document), fasilitas (facilities), infrastruktur (infrastructure), karir (placement), kepala jurusan (head of department), kepala prodi (head of study program), komite (committee), lokasi (location), nomor (number), olahraga (sports), pelonco (ragging), perpustakaan (library), pmb (college intake), rektor (rector), and sekretaris jurusan (department secretary). Experimental results to evaluate chatbot intent classification models using the RoBERTa method show an accuracy of 0.84, as shown in Fig. 4.

	precision	recall	f1-score	support
acak	0.00	0.00	0.00	1
acara	1.00	1.00	1.00	3
asrama	1.00	0.83	0.91	6
beasiswa	1.00	1.00	1.00	7
biaya	0.83	0.83	0.83	6
dekan	1.00	1.00	1.00	1
dokumen	1.00	1.00	1.00	3
fasilitas	1.00	1.00	1.00	1
infrastruktur	1.00	1.00	1.00	1
jam	0.80	1.00	0.89	4
jurusan	1.00	0.86	0.92	7
kantin	0.50	1.00	0.67	3
karir	1.00	1.00	1.00	2
kepala jurusan	1.00	1.00	1.00	1
kepala prodi	1.00	1.00	1.00	1
komite	1.00	1.00	1.00	1
lantai	0.67	1.00	0.80	2
liburan	1.00	0.67	0.80	3
lokasi	1.00	1.00	1.00	4
menu	0.00	0.00	0.00	2
nama	0.75	1.00	0.86	3
nomor	1.00	1.00	1.00	4
olahraga	1.00	1.00	1.00	2
pelonco	1.00	1.00	1.00	2
pencipta	0.80	1.00	0.89	4
penerimaan	0.50	1.00	0.67	1
perpisahan	0.50	0.33	0.40	3
perpustakaan	1.00	1.00	1.00	4
pmb	1.00	1.00	1.00	2
rektor	1.00	1.00	1.00	2
salam	0.50	0.50	0.50	2
salut	0.50	0.67	0.57	3
sekretaris jurusan	1.00	1.00	1.00	1
semester	0.60	1.00	0.75	3
seragam	1.00	0.50	0.67	2
silabus	0.50	0.50	0.50	2
tugas	0.00	0.00	0.00	1
umpat	0.00	0.00	0.00	2
accuracy			0.84	102
macro avg	0.78	0.81	0.78	102
weighted avg	0.82	0.84	0.82	102

Fig. 4. Metric Measurement for Each Intent in RoBERTa

C. Comparison Results of BERT and RoBERTa

Table VII shows that the BERT model achieved an accuracy of 0.89, compared to 0.84 for the RoBERTa model. This suggests that BERT provides better accuracy than RoBERTa when using the same Indonesian language dataset and hyperparameters. Several factors may contribute to BERT's superior performance. First, the characteristics and relatively small size of the dataset might make it more suitable for the BERT model, reducing the risk of overfitting. Second, hyperparameter tuning plays a crucial role, as different hyperparameter configurations could lead to better performance for BERT on this particular dataset.

TABLE VII. COMPARISON RESULT OF BERT AND ROBERTA

Component	BERT	RoBERTa
Accuracy	0.89	0.84
Precision	0.89	0.77
Recall	0.88	0.80
F1-Score	0.87	0.77

IV. CONCLUSIONS

The comparison of experimental results for intent classification in chatbots using BERT and RoBERTa models leads to the following conclusions. Both BERT and RoBERTa can be effectively used for intent classification in Indonesian chatbots. After translating the University Chatbot Dataset into Bahasa Indonesian, the BERT model achieved an accuracy of 0.89, outperforming the RoBERTa model, which achieved 0.84. This indicates that the BERT model provides superior accuracy compared to RoBERTa when using the same Indonesian dataset and identical hyperparameters. In the future, we plan to experiment with the IndoBERT model for chatbot intent classification, specifically using Bahasa Indonesia datasets and larger datasets. This will help us further evaluate whether BERT consistently outperforms RoBERTa in chatbot intent classification tasks. These will be the next major works.

ACKNOWLEDGMENT

This research is funded by the Directorate of Research, Technology, and Community Service under the Directorate General of Higher Education, Research, and Technology, in accordance with the contract for the implementation of the State Higher Education Operational Assistance Program (Research Program) for the 2024 fiscal year, Contract Number: 090/E5/PG.02.00.PL/2024.

REFERENCES

[1] N. Shahin and L. Ismail, "From Rule-Based Models to Deep Learning Transformers Architectures for Natural Language Processing and Sign Language Translation Systems: Survey, Taxonomy and Performance Evaluation," 2024, *arXiv*. doi: 10.48550/ARXIV.2408.14825.

[2] L. Villa, D. Carneros-Prado, A. Sánchez-Miguel, C. C. Dobrescu, and R. Hervás, "Conversational Agent Development Through Large Language Models: Approach with GPT," in *Proceedings of the 15th International Conference on Ubiquitous Computing & Ambient Intelligence (UCAmI 2023)*, vol. 835, J. Bravo and G. Urzáiz, Eds., in Lecture Notes in Networks and Systems, vol. 835. , Cham: Springer Nature Switzerland, 2023, pp. 286–297. doi: 10.1007/978-3-031-48306-6_29.

[3] D. Griol, Z. Callejas, J. M. Molina, and A. Sanchis, "Adaptive dialogue management using intent clustering and fuzzy rules," *Expert Systems*, vol. 38, no. 1, p. e12630, Jan. 2021, doi: 10.1111/exsy.12630.

[4] 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*, Yokohama Japan: ACM, May 2021, pp. 160–166. doi: 10.1145/3429360.3468203.

[5] A. Birim and M. Erden, "Robustness to Spelling Errors for Intent Detection," in *2022 30th Signal Processing and Communications Applications Conference (SIU)*, Safranbolu, Turkey: IEEE, May 2022, pp. 1–4. doi: 10.1109/SIU55565.2022.9864722.

[6] N. Boudjani, V. Colas, and A. Fotouhi, "Intent Classification: French Recruitment Chatbot Use Case," in *2023 International Conference on Computational Science and Computational Intelligence (CSCI)*, Las Vegas, NV, USA: IEEE, Dec. 2023, pp. 681–685. doi: 10.1109/CSCI62032.2023.00117.

[7] J.-H. Lee, E. H.-K. Wu, Y.-Y. Ou, Y.-C. Lee, C.-H. Lee, and C.-R. Chung, "Anti-Drugs Chatbot: Chinese BERT-Based Cognitive Intent Analysis," *IEEE Trans. Comput. Soc. Syst.*, vol. 11, no. 1, pp. 514–521, Feb. 2023, doi: 10.1109/TCSS.2023.3238477.

[8] F. Roma, G. Sansonetti, G. D'Aniello, and A. Micarelli, "A BERT-Based Approach to Intent Recognition," in *IEEE EUROCON 2023 - 20th International Conference on Smart Technologies*, Torino, Italy: IEEE, Jul. 2023, pp. 568–572. doi: 10.1109/EUROCON56442.2023.10198959.

[9] S. Sayenju *et al.*, "Quantification and Mitigation of Directional Pairwise Class Confusion Bias in a Chatbot Intent Classification Model," *Int. J. Semantic Computing*, vol. 16, no. 04, pp. 497–520, Dec. 2022, doi: 10.1142/S1793351X22500040.

[10] Y. Guo *et al.*, "ESIE-BERT: Enriching Sub-words Information Explicitly with BERT for Joint Intent Classification and SlotFilling," Feb. 02, 2023, *arXiv*: arXiv:2211.14829. Accessed: Sep. 02, 2024. [Online]. Available: http://arxiv.org/abs/2211.14829

[11] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," Jul. 26, 2019, *arXiv*: arXiv:1907.11692. Accessed: Mar. 13, 2024. [Online]. Available: http://arxiv.org/abs/1907.11692

[12] A. Souha, C. Ouaddi, L. Benaddi, and A. Jakimi, "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)*, Rabat, Morocco: IEEE, Dec. 2023, pp. 1–6. doi: 10.1109/CommNet60167.2023.10365312.

[13] K. K. Jayanth, G. Bharathi Mohan, R. P. Kumar, and M. Rithani, "Intent Recognition Leveraging XLM-RoBERTa for Effective NLU," in *2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, Salem, India: IEEE, Jun. 2024, pp. 877–882. doi: 10.1109/ICAAIC60222.2024.10575275.

[14] Rohim and Zuliarso, "Penerapan Algoritma Deep Learning Untuk Pengembangan Chatbot Yang Digunakan Untuk Konsultasi Dan Pengenalan Tentang Virus Covid-19," *PIXEL*, vol. 15, no. 2, pp. 267–278, Dec. 2022, doi: 10.51903/pixel.v15i2.777.

[15] R. C. Hutama, F. Fauziah, and R. T. Komalasari, "Aplikasi Chatbot Berbasis Teks Menggunakan Algoritma Naive Bayes Classifier FAQ GrabAds," *STRING*, vol. 6, no. 1, p. 90, Aug. 2021, doi: 10.30998/string.v6i1.9919.

[16] D. Theosaksomo and D. H. Widyantoro, "Conversational Recommender System Chatbot Based on Functional Requirement," in *2019 IEEE 13th International Conference on Telecommunication Systems, Services, and Applications (TSSA)*, Bali, Indonesia: IEEE, Oct. 2019, pp. 154–159. doi: 10.1109/TSSA48701.2019.8985467.

[17] M. Y. Helmi Setyawan, R. M. Awangga, and S. R. Efendi, "Comparison Of Multinomial Naive Bayes Algorithm And Logistic Regression For Intent Classification In Chatbot," in *2018 International Conference on Applied Engineering (ICAE)*, Batam: IEEE, Oct. 2018, pp. 1–5. doi: 10.1109/INCAE.2018.8579372.

[18] M. Menda and G. S. Keerthi, "Intent Classification in Conversational System using Machine Learning Techniques," *IJCA*, vol. 183, no. 51, pp. 6–11, Feb. 2022, doi: 10.5120/ijca2022921913.

[19] C. A. Oktavia, "Implementasi Chatbot Menggunakan Dialogflow dan Messenger Untuk Layanan Customer Service Pada E-Commerce," *JIMP*, vol. 4, no. 3, Jan. 2020, doi: 10.37438/jimp.v4i3.230.

[20] D. Liu, Z. Zhao, and L.-D. Gan, "Intention Detection Based On Bert-Bilstm in Taskoriented Dialogue System," in *2019 16th International Computer Conference on Wavelet Active Media Technology and Information Processing*, Chengdu, China: IEEE, Dec. 2019, pp. 187–191. doi: 10.1109/ICCWAMTIP47768.2019.9067660.

[21] Nirali Vaghani, "Chatbot dataset." Kaggle. doi: 10.34740/KAGGLE/DSV/5024271.

[22] DeepL, "DeepL Translator." [Online]. Available: https://www.deepl.com/

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," May 24, 2019, *arXiv*: arXiv:1810.04805. Accessed: Feb. 28, 2024. [Online]. Available: http://arxiv.org/abs/1810.04805

[24] J. H. Tandijaya and I. Sugiarto, "Klasifikasi dalam Pembuatan Portal Berita Online dengan Menggunakan Metode BERT," vol. Vol 9, No 2 (2021), 2021.

[25] R. Khusuma, W. Maharani, and P. H. Gani, "Personality Detection On Twitter User With RoBERTa," *Jurnal Media Informatika Budidarma*, vol. 7, 2023.