JOURNAL OF APPLIED DIGITAL AND INFORMATION SYSTEMS
Website: https://ejournal.gws-tech.id/index.php/jadis/

Transforming Access to Indonesian Plantation Data Through a BERT-Based Interactive Chatbot Information System

Transformasi Akses Data Perkebunan Indonesia Melalui Sistem Informasi Chatbot Interaktif Berbasis BERT

Afifah Liza Naura, Akbar Gilang Perdana, Jenny Ida Rahayu, M. Habil Alfathir Ghandur, Vorel Juliandro, Chairun Nas

Information Systems Study Program, Universitas Riau, Pekanbaru, 28293, Indonesia
afifah.liza1705@student.unri.ac.id, akbar.gilang3289@student.unri.ac.id, jenny.ida5328@student.unri.ac.id, m.habil.alfathir3909@student.unri.ac.id, vorel.juliandro3067@student.unri.ac.id, chairun.nas@lecturer.unri.ac.id

Article Information
Submitted: July 6, 2026
Accepted: July 14, 2026
Published: July 15, 2026

Corresponding Author
Chairun Nas
chairun.nas@lecturer.unri.ac.id

DOI
https://doi.org/10.65897/jadis.v1.i1.94

Copyright © 2026 The Author(s). Published by GWS Tech Solution This is an open-access article under the CC BY 4.0 International License.

ABSTRACT

The Indonesian plantation sector relies heavily on timely access to production information by province and commodity. However, production data is commonly presented in static tables, requiring users to manually search for specific values, comparisons, or aggregated answers. This study redesigns an existing system prototype into AGRINOVA, a closed-domain information system chatbot that utilizes a BERT-based semantic matching approach to answer queries regarding Indonesian plantation production in 2024. The research methodology includes data acquisition from official plantation production statistics, data cleaning, entity normalization for provinces and commodities, question template formulation, BERT-based text representation, cosine similarity ranking, and functional testing through information retrieval scenarios. The prototype was implemented via a text-based interface and evaluated using four types of queries: direct information, comparisons, extreme value identification, and aggregation. The testing results demonstrate that all evaluated answers match the expected dataset values, with semantic similarity scores of 0.8824, 0.9315, 0.9099, and 0.9214, respectively, and an average similarity score of 0.9113. These findings indicate that a BERT-based information retrieval chatbot can enhance interactive access to structured agricultural data. Future research requires broader user testing, web-based integration, and more robust handling of complex natural language variations.

Keywords:

Agricultural Chatbot; BERT; Information Retrieval; Natural Language Processing; Plantation Production

ABSTRAK

Sektor perkebunan Indonesia bergantung pada akses yang tepat waktu terhadap informasi produksi berdasarkan provinsi dan komoditas. Namun, data produksi umumnya disajikan dalam tabel statis, sehingga pengguna harus melakukan pencarian secara manual untuk mendapatkan nilai spesifik, perbandingan, atau jawaban agregat. Penelitian ini melakukan redesain prototipe sistem menjadi AGRINOVA, sebuah chatbot sistem informasi domain tertutup yang menggunakan pendekatan pencocokan semantik berbasis BERT untuk menjawab pertanyaan mengenai produksi perkebunan Indonesia tahun 2024. Metode penelitian meliputi akuisisi data dari statistik resmi produksi perkebunan, pembersihan data, normalisasi entitas provinsi dan komoditas, penyusunan templat pertanyaan, representasi teks berbasis BERT, pemeringkatan cosine similarity, serta pengujian fungsional melalui skenario temu kembali informasi. Prototipe ini diimplementasikan melalui antarmuka berbasis teks dan dievaluasi menggunakan empat jenis pertanyaan: informasi langsung, perbandingan, identifikasi nilai ekstrem, dan agregasi. Hasil pengujian menunjukkan bahwa seluruh jawaban yang dievaluasi sesuai dengan nilai dataset yang diharapkan, dengan skor kemiripan semantik masing-masing sebesar 0,8824, 0,9315, 0,9099, dan 0,9214, serta rata-rata kemiripan sebesar 0,9113. Temuan ini mengindikasikan bahwa chatbot temu kembali informasi berbasis BERT dapat meningkatkan akses interaktif terhadap data pertanian yang terstruktur. Penelitian selanjutnya memerlukan pengujian yang lebih luas, integrasi berbasis web, serta penanganan yang lebih tangguh terhadap variasi bahasa alami yang kompleks.

***Kata Kunci**:*

Chatbot Pertanian; BERT; Temu Kembali Informasi; Pemrosesan Bahasa Alami; Produksi Perkebunan

1. INTRODUCTION

Plantation production data are strategic information assets for Indonesia because they describe the distribution of major estate crops across provinces and support planning, monitoring, and public communication. Official statistics on crops such as oil palm, rubber, coffee, cocoa, tea, coconut, and sugarcane provide a structured basis for data-driven decisions in government, education, and agribusiness [1]. In a digital agriculture context, these datasets have value when users can retrieve them quickly and interpret them without repeatedly opening static tables. Prior work on smart farming also shows that agricultural value increases when data become accessible, integrated, and actionable , particularly as remote sensing and cloud-based platforms now enable the extraction of plantation extent and age independent of the often-incomplete nature of traditional reporting [2].

The main problem addressed in this study is not the absence of data, but the limited interactivity of the information access mechanism. Users who only need a specific provincial production figure, a comparison across regions, or a total value often need to scan tables manually. This process is inefficient for non-technical users and does not fit the interaction pattern expected from modern information systems. Deep learning has been widely adopted in agricultural data processing, while transformer-based language models have strengthened natural language understanding by modeling contextual relations in text. BERT extends this capability through bidirectional pretraining, which makes it suitable for understanding user intent in question answering and semantic matching tasks [3].

Indonesian language processing also needs models and resources that fit local linguistic characteristics. IndoNLU and IndoBERT demonstrate that Indonesian pre-trained language resources can support sentence classification, semantic similarity, and other natural language understanding tasks. Sentence-BERT further shows that BERT representations can be optimized for sentence-level similarity comparison [4]. These developments are relevant to a closed-domain agricultural chatbot because the system must connect a user question with a structured answer candidate, not generate unsupported content. Question answering research such as SQuAD also indicates that robust QA systems require clear answer grounding and measurable evaluation [5] supports the Chains-BERT claim about contrastive learning integrated with BERT for agricultural QA in small-scale corpora. Recent advancements, such as the Chains-BERT model for agricultural question answering, have further refined these frameworks by integrating contrastive learning with BERT to enhance performance in small-scale or resource-constrained agricultural corpora [5].

Chatbots have been used in many information service domains because they allow users to request information through natural language rather than rigid menus. However, chatbot quality depends on response accuracy, relevance, clarity, and the ability to handle domain-specific inputs. Recent agricultural chatbot studies show the potential of AI assistants to scale agricultural advisory services, while domain-specific farm assistants commonly combine intent recognition, entity extraction, and database retrieval. In the Indonesian information systems setting, several recent studies also report the use of BERT, IndoBERT, passage retrieval, and cosine similarity for chatbot and question answering applications [6].

Based on that gap, this study develops AGRINOVA, a BERT-based agricultural chatbot for Indonesian plantation production information in 2024. The contribution of this paper is threefold. First, it reframes a plantation production chatbot as an information system problem focused on data access, retrieval accuracy, and user interaction. Second, it proposes a closed-domain architecture that links natural language questions to official production data. Third, it evaluates the prototype through representative query scenarios and similarity scores.

This is an open-access article under the
Creative Commons Attribution 4.0 International License (CC BY 4.0).

2. METHODS

2.1. Research Design and Data Source

This study uses a design and development approach. The object of development is a closed-domain chatbot that provides answers from a structured dataset instead of producing free-form generative responses. This choice fits the information system scope because the system emphasizes data acquisition, data processing, information retrieval, interface design, and evaluation of answer relevance [7].

The dataset contains Indonesian plantation production information for 2024. The primary fields are province, commodity, production value, and year. The prototype focuses on commodity production questions at provincial level and supports direct search, comparison, highest or lowest value identification, and simple aggregation.

2.2. Data Preparation

Data preparation started with cleaning province names, commodity names, numeric production values, and year attributes. Province and commodity strings were normalized to reduce variation in capitalization and spelling. Missing and zero values were retained when they reflected the official dataset condition because removing them would distort questions about minimum production or provincial totals. To ensure the model correctly maps natural language queries to these structured records, the cleaned data were encoded into a vector space using IndoBERT embeddings to facilitate semantic similarity matching [8].

After cleaning, each row was transformed into a candidate knowledge item. For example, one row could be represented as a sentence pattern such as 'Oil palm production in Riau in 2024 is 9,136.1 thousand tons.' This representation allows the language model to compare user questions with answer candidates through sentence-level similarity.

Table 1. Dataset Structure Used in the AGRINOVA Prototype

Field	Description and Role
Prov.	Production Area
Crop	Types of Plantation Crops
Prod.	Production Volume (tonnes)
Year	Data Year (2024)

2.3. Chatbot Architecture

AGRINOVA follows a retrieval-based architecture. The user submits a natural language question through a text interface. The system preprocesses the question through case folding, token handling, and removal of irrelevant symbols. The cleaned question is encoded using a BERT-based representation and compared with candidate knowledge items in the dataset. The candidate with the highest semantic score is selected as the response when the score passes the relevance threshold. This retrieval mechanism ensures that responses remain strictly grounded in verified production records, preventing the hallucination risks often associated with purely generative models [9].

The similarity calculation uses cosine similarity between the question vector and candidate answer vector. A higher score means stronger semantic alignment between the input question and the selected answer candidate. This strategy is suitable for a closed dataset because the system can remain grounded in verified plantation production records.

This is an open-access article under the
Creative Commons Attribution 4.0 International License (CC BY 4.0).

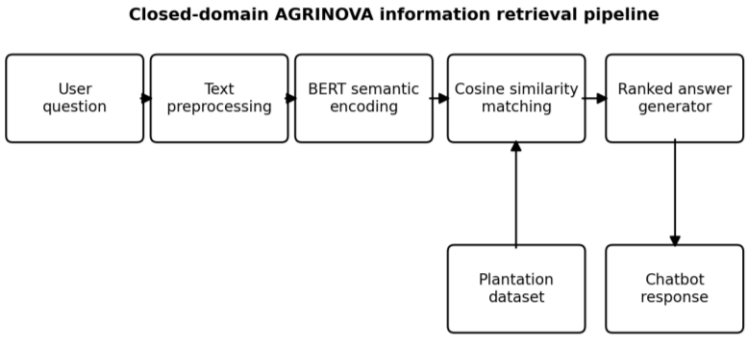

Figure 1. AGRINOVA BERT-Based Retrieval Architecture

2.4. Evaluation Procedure

The prototype was evaluated through functional question scenarios that represent common user needs. The evaluation used four categories: direct information, comparison, analysis of the lowest or highest production value, and aggregation. Each generated answer was compared with the expected value in the dataset. The similarity score was also recorded to assess whether the selected response had strong semantic alignment with the question.

The evaluation is intentionally positioned as an initial prototype test. It does not claim general performance across all possible Indonesian question forms. Its purpose is to verify whether the proposed architecture can retrieve correct production information from the prepared dataset and show measurable relevance for representative inputs.

3. RESULTS

3.1. Prototype Implementation

```
============================================================
 AGRINOVA
 CHATBOT PRODUKSI PERKEBUNAN DI INDONESIA TAHUN 2024
============================================================

Ketik 'exit' untuk keluar.

Anda : Berapa produksi Kelapa Sawit di provinsi RIAU pada tahun 2024?

AGRINOVA :
9136.1 ribu ton

Similarity :
0.8824

Anda : Provinsi mana yang paling sedikit produksi perkebunan keseluruhan?

AGRINOVA :
provinsi DKI JAKARTA

Similarity :
0.9315

Anda : Berapa total hasil perkebunan di provinsi SULAWESI BARAT pada tahun 2024?

AGRINOVA :
475.29 ribu ton

Similarity :
0.9099

Anda : Berapa produksi Kelapa Sawit di provinsi PAPUA BARAT DAYA pada tahun 2024?

AGRINOVA :
43.44 ribu ton

Similarity :
0.9214

Anda :
```

Figure 2. User Interaction View With the AGRINOVA Chatbot

This is an open-access article under the
Creative Commons Attribution 4.0 International License (CC BY 4.0).

The implemented prototype is named AGRINOVA. The system runs through a command line interface where users type questions about Indonesian plantation production in 2024. The system returns the selected answer and displays a similarity score. This interface is sufficient for prototype validation because it isolates the retrieval logic before web or mobile deployment. Furthermore, the modular design facilitates future scalability by allowing the integration of advanced embedding models or expanded knowledge bases without modifying the core retrieval framework [10].

The chatbot supports questions such as production value by province and commodity, identification of the province with the lowest overall plantation production, and total production calculation for a specific province. The response is concise because the main goal is to return a factual production value, not a long explanatory paragraph.

Figure 2 illustrates the implementation of the text-based AGRINOVA chatbot system interface, where users can ask questions directly and receive relevant responses. The display also shows the similarity score, serving as an indicator of the degree of alignment between the question and the answer generated by the system.

3.2. Functional Test Results

Table 2. Functional Test Results of the AGRINOVA Chatbot

ID	Focus	Answer	Score
Q1	Riau palm	9136.1k	0.8824
Q2	Lowest	DKI Jakarta	0.9315
Q3	Sulbar total	475.29k	0.9099
Q4	SWP palm	43.44k	0.9214

The functional test shows that the chatbot returned the expected answer for each evaluated scenario. The average similarity score across the four queries was 0.9113. The highest score was found in Q2, which asked for the province with the lowest total production. The lowest score was found in Q1, but the answer still matched the expected dataset value. This high level of performance indicates that the BERT-based retrieval framework effectively maintains factual grounding when querying structured agricultural data [11].

The results indicate that the BERT-based semantic matching method can handle direct questions, comparative questions, and simple analytical questions in the prepared domain. The score distribution also suggests that normalized answer candidates help the system connect user questions with structured records.

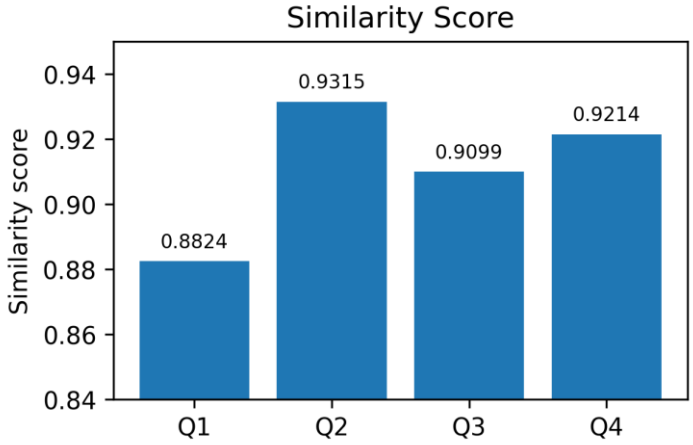

Figure 3. Similarity Score Distribution Across Functional Test Questions

This is an open-access article under the
Creative Commons Attribution 4.0 International License (CC BY 4.0).

3.3. Question Category Coverage

Table 3. Question Categories Supported in Prototype Testing

Type	Purpose	Out
Direct	Lookup	Value
Compare	Rank	Lowest
Analysis	Min/Max	Extemen
Aggregate	Sum	Total

The category coverage shows that AGRINOVA is not limited to a single lookup pattern. It can handle several information retrieval needs that commonly occur when users explore plantation production data. This coverage strengthens its relevance as a lightweight information system for structured agricultural statistics. Future development will focus on incorporating more robust domain-specific language models to enhance semantic understanding, as demonstrated by the efficacy of fine-tuning strategies in similar agricultural information tasks [12]. Specifically, prior research demonstrates that enhancing training datasets with domain-specific semantic information significantly improves a model's ability to rank relevant answers accurately [13]. Integrating such fine-tuning techniques is expected to stabilize the system's performance metrics, potentially yielding lower standard deviations in similarity scores across complex query types [14]. Furthermore, transitioning from standard BERT architectures to more specialized models trained on agricultural corpora has been shown to elevate performance metrics such as mean average precision, particularly when combined with entity linking techniques [15].

4. DISCUSSION

The findings show that a BERT-based retrieval approach is appropriate for a closed-domain plantation production chatbot. The system does not need to generate new claims outside the dataset. Instead, it needs to understand a user question, identify the relevant province and commodity context, and return the correct record. This design reduces the risk of unsupported answers and keeps the chatbot aligned with official data. However, ensuring that the system reliably captures all relevant context remains an ongoing challenge, as the sufficiency of retrieved chunks can be highly sensitive to the top-k parameter and query phrasing [16].

From an information systems perspective, AGRINOVA contributes to data access, user interaction, and service efficiency. Static tables remain useful for reporting, but they require manual exploration. A chatbot interface changes the access pattern into a conversational query process. This is especially useful for users who need quick answers, such as students, researchers, extension workers, and public information officers.

The use of a similarity score also provides a transparent indicator of response relevance. In a future web-based implementation, the score can support confidence labeling, fallback responses, or request clarification when the score is low. This would improve usability because the system can avoid giving a definitive answer when the question is outside the dataset scope or semantically ambiguous.

Several limitations need attention. First, the evaluation used only four representative questions, so the current findings cannot describe full linguistic robustness. Second, the prototype interface is still command-line based and does not yet include a public web interface, user authentication, logging, or analytics. Third, the dataset is limited to 2024 plantation production records, so the system cannot yet answer time-series trends or policy questions. Future work should expand the test set, include more varied Indonesian question forms, add entity disambiguation, and deploy the system as a web-based dashboard with chatbot support.

This is an open-access article under the
Creative Commons Attribution 4.0 International License (CC BY 4.0).

5. CONCLUSION

This study developed AGRINOVA, a BERT-based agricultural chatbot for Indonesian plantation production information in 2024. The system uses a closed-domain retrieval architecture that transforms official plantation records into searchable knowledge items and matches user questions through semantic similarity. Functional testing on four representative query categories produced correct responses with an average similarity score of 0.9113. These results show that the proposed chatbot can improve interactive access to structured agricultural production data and fits the focus of information systems research. The main limitation is the small evaluation set and the command-line interface. Future research should use a larger benchmark, include more natural language variations, compare BERT with IndoBERT and other retrieval models, and deploy the prototype as a web-based agricultural information service.

REFERENCES

[1] F. Pendrill *et al.*, "Disentangling the numbers behind agriculture-driven tropical deforestation," *Science*, vol. 377, no. 6611, Sep. 2022, doi: 10.1126/science.abm9267.

[2] O. Danylo *et al.*, "A map of the extent and year of detection of oil palm plantations in Indonesia, Malaysia and Thailand," *Scientific Data*, vol. 8, no. 1, pp. 96–96, Mar. 2021, doi: 10.1038/s41597-021-00867-1.

[3] Z. Zhang *et al.*, "Semantics-Aware BERT for Language Understanding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Association for the Advancement of Artificial Intelligence, Apr. 2020, pp. 9628–9635. doi: 10.1609/aaai.v34i05.6510.

[4] R. G. Santosa, A. B. Nusantara, and S. Imron, "Comparative Analysis of SVM and IndoBERT for Intent Classification in Indonesian Overtime Chatbots," *Journal of System and Computer Engineering (JSCE)*, vol. 6, no. 3, pp. 258–270, Aug. 2025, doi: 10.61628/jsce.v6i3.2058.

[5] Y. Huang, J. Liu, and C. Lv, "Chains-BERT: A High-Performance Semi-Supervised and Contrastive Learning-Based Automatic Question-and-Answering Model for Agricultural Scenarios," *Applied Sciences*, vol. 13, no. 5, pp. 2924–2924, Feb. 2023, doi: 10.3390/app13052924.

[6] A. Elcholiqi and A. Musdholifah, "Chatbot in Bahasa Indonesia using NLP to Provide Banking Information," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 14, no. 1, pp. 91–91, Jan. 2020, doi: 10.22146/ijccs.41289.

[7] G. Bilquise, S. Ibrahim, and K. Shaalan, "Bilingual AI-Driven Chatbot for Academic Advising," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 8, Jan. 2022, doi: 10.14569/ijacsa.2022.0130808.

[8] M. Maryamah, M. M. Irfani, E. B. T. Raharjo, N. A. Rahmi, M. Ghani, and I. K. Raharjana, "Chatbots in Academia: A Retrieval-Augmented Generation Approach for Improved Efficient Information Access," pp. 259–264, Feb. 2024, doi: 10.1109/kst61284.2024.10499652.

[9] N. Singh *et al.*, "Farmer.Chat: Scaling AI-Powered Agricultural Services for Smallholder Farmers," Sep. 13, 2024, *Cornell University*. doi: 10.48550/arxiv.2409.08916.

[10] S. Siriwardhana, R. Weerasekera, E. Wen, T. Kaluarachchi, R. Rana, and S. Nanayakkara, "Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1–17, Jan. 2023, doi: 10.1162/tacl_a_00530.

[11] K. Muludi, K. M. Fitria, J. Triloka, and S. Sutedi, "Retrieval-Augmented Generation Approach: Document Question Answering using Large Language Model," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 3, Jan. 2024, doi: 10.14569/ijacsa.2024.0150379.

[12] B. Zhao, W. Jin, J. D. Ser, and G. Yang, "ChatAgri: Exploring potentials of ChatGPT on cross-linguistic agricultural text classification," *Neurocomputing*, vol. 557, pp. 126708–126708, Aug. 2023, doi: 10.1016/j.neucom.2023.126708.

[13] S. Rezayi *et al.*, "Exploring New Frontiers in Agricultural NLP: Investigating the Potential of Large Language Models for Food Applications," *IEEE Transactions on Big Data*, vol. 11, no. 3, pp. 1235–1246, Aug. 2024, doi: 10.1109/tbdata.2024.3442542.

This is an open-access article under the
Creative Commons Attribution 4.0 International License (CC BY 4.0).

[14] K. Rangan and Y. Yin, "A fine-tuning enhanced RAG system with quantized influence measure as AI judge," *Scientific Reports*, vol. 14, no. 1, pp. 27446–27446, Nov. 2024, doi: 10.1038/s41598-024-79110-x.

[15] S. Rezayi *et al.*, "AgriBERT: Knowledge-Infused Agricultural Language Models for Matching Food and Nutrition," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, Jul. 2022, pp. 5150–5156. doi: 10.24963/ijcai.2022/715.

[16] S. A. Vaghefi *et al.*, "ChatClimate: Grounding conversational AI in climate science," *Communications Earth & Environment*, vol. 4, no. 1, Dec. 2023, doi: 10.1038/s43247-023-01084-x.

This is an open-access article under the
Creative Commons Attribution 4.0 International License (CC BY 4.0).