Jurnal Minfo Polgan (JMP)
Volume 15, Nomor 2, Mei-Agustus 2026
DOI : https://doi.org/10.33395/jmp.v15i2.16492
e-ISSN : 2797-3298
p-ISSN : 2089-9424

Trustworthy NLP Systems for Educational Decision Support: A Human-Centered AI Approach

[*1]Subhan Hafiz Nanda Ginting, [2]Ericky Benna Perolihin Manurung, [3]Nuranisah, [4]Dewi Wahyuni
[1*,3,4] Fakultas Teknologi, Universitas Battuta, Medan, Indonesia
[2]Computer Science Departement, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia
*Korespondensi: subhanhafiz16@gmail.com

Submitted : Mei 09, 2026 | **Accepted**: Jun 16, 2026 | **Published** : Jun 21, 2026

ABSTRACT

Advances in Natural Language Processing (NLP) in education have led to the development of decision support systems capable of processing textual data such as student essay responses, learning feedback, academic records, and evaluation documents. However, most previous research has focused on improving model accuracy, while aspects of trust, transparency, fairness, and human involvement in decision validation have not been a primary focus. This research aims to develop a framework for Trustworthy NLP Systems for Educational Decision Support based on a Human-Centered Artificial Intelligence approach that positions teachers, students, and educational policymakers as key actors in the system's design, interpretation, and evaluation processes. The novelty of this research lies in the integration of four key dimensions explainability, fairness, reliability, and human oversight into the NLP system architecture to support more ethical, transparent, and accountable educational decisions. The research methodology employs an experimental approach involving stages of educational text data collection, data preprocessing, text representation, NLP-based modeling, model performance evaluation, and system trust analysis through Explainable AI and fairness evaluation. The developed system is not only designed to generate educational classifications or recommendations but also to provide explanations for the model's decision-making basis, thereby enabling human verification. The expected outcome is the creation of a conceptual and technical NLP model capable of improving the quality of educational decision-making without compromising ethical principles, accountability, and user-centricity. This research contributes to strengthening the direction of educational NLP development that is not only computationally intelligent but also trustworthy, inclusive, and human-centered.

Keywords: *Natural Language Processing, Trustworthy AI, Educational Decision Support, Human-Centered AI, Explainable AI.*

INTRODUCTION

Advances in artificial intelligence (AI) technology have had a significant impact on the transformation of modern education systems. One rapidly evolving and highly significant area of AI in the context of education is natural language processing (NLP). NLP enables computer systems to understand, process, analyze, and generate human language in the form of text or speech. In the field of education, NLP technology can be used to analyze students' written responses, evaluate learning feedback, identify patterns of learning difficulties, assess student sentiment, process academic documents, and support data driven decision making. With these capabilities, NLP has the potential to become a key component in the development of more adaptive, responsive, and evidence-based decision support systems in education (Manohara et al., 2024).

Educational decision support systems play a strategic role in assisting teachers, lecturers, administrative staff at educational institutions, and policymakers in making decisions regarding the learning process, academic assessment, recommendations for support measures, competency mapping, and the formulation of strategies to improve the quality of education. To date, decision making in the field of education has largely relied on manual analysis, educators' intuition, and limited quantitative data. In reality, educational institutions generate a vast amount of unstructured text data, such as students' open-ended responses, reflection notes, teacher

comments, academic reports, discussion forums, and learning survey results. This data contains crucial information that can be used to gain a more comprehensive understanding of students' learning conditions. Therefore, integrating NLP into decision-support systems in education is essential for transforming text data into meaningful insights that can serve as the basis for decision-making.

Although NLP is widely used in various educational studies, most research still focuses on improving model performance, particularly in terms of accuracy, precision, recall, and F1 scores. An approach that overemphasizes computational power risks neglecting other equally important aspects, such as trust, transparency, fairness, accountability, and human involvement in the decision-validation process. In an educational context, decisions generated by AI systems can have direct implications for students, such as in learning recommendations, academic evaluations, the identification of learning risks, or the provision of specific support measures (Jannat, 2026). If an NLP system makes biased, non-transparent, or hard-to-explain decisions, those decisions can lead to unfairness and erode user trust in the technology (Al-Turki et al., 2026).

The issue of trust in NLP systems is becoming increasingly critical, as modern NLP models particularly those based on deep learning and transformers are often complex and difficult to interpret. While models can generate accurate predictions, they are not always able to clearly explain the reasoning behind their decisions (Pujitha & Saritha, 2026). This condition is known as the black box problem a situation where the model's internal processes are difficult for human users to understand. In educational systems, this black-box nature poses serious challenges, as educators and policymakers require a clear foundation before they can accept or implement system recommendations (Briva-Iglesias & O'Brien, 2026). Educational decisions cannot be fully entrusted to machines without mechanisms for human explanation, oversight, and validation.

RESEARCH METHODOLOGY

Research Design

This study employs a mixed-methods approach with a sequential explanatory design, which integrates quantitative and qualitative methods sequentially to develop and evaluate a Trustworthy Natural Language Processing (NLP) system to support educational decision-making based on Human-Centered Artificial Intelligence (HCAI) (Topali et al., 2025). This approach was chosen because the development of a trustworthy AI system requires not only the measurement of a model's technical performance but also an evaluation of aspects such as transparency, fairness, accountability, interpretability, and user acceptance.

The research was conducted in three main stages: (1) identification of user needs and system design, (2) development of an NLP model oriented toward the principles of trustworthy AI, and (3) technical and user evaluations of the developed system (Dehghani et al., 2024).

Research Stages

1. Needs Analysis and System Design

The initial phase of the research was conducted through a needs assessment involving education stakeholders, such as professors, teachers, school principals, educational institution administrators, and academic administrative staff. The objective was to identify the information needs required for educational decision-making (Schoenherr et al., 2023).

Data collection methods included: Semi-structured interviews, Focus Group Discussions (FGDs), Analysis of academic documents and educational policies (Barale, 2022).

Data from the interviews and FGDs were analyzed using Thematic Analysis to identify key themes related to the requirements for a decision support system. The results of the needs analysis were used to design a user-centered NLP system architecture (Hao et al., 2026).

NLP Model Development

1. Model Architecture

This study adopts a Transformer-Based Language Model approach by utilizing base models such as (Fetaji et al., 2025):

BERT, RoBERTa, DeBERTa.

The models were then fine-tuned on an educational corpus to generate language representations appropriate for an academic context (Tharini & Jeyaraj, 2026).

2. Trustworthy AI Components

To enhance the system's trustworthiness, several components were added to the model architecture, namely:

a. Explainability Module
Using the following methods:
SHAP (SHapley Additive Explanations), LIME (Local Interpretable Model-Agnostic Explanations), Attention Visualization (Wang et al., 2023).

This module explains why the model generates a particular decision recommendation (He et al., 2022).

b. Fairness Assessment Modul
Fairness analysis is conducted using the following indicators:
Demographic Parity, Equal Opportunity, Equalized Odds (Handoko et al., 2026).

Testing is conducted on different groups of students based on academic and demographic characteristics.

c. Uncertainty Estimation Module
To improve the reliability of predictions, the following are used:
Monte Carlo Dropout, Bayesian Neural Network. This module enables the system to assign a confidence score to each recommendation.

d. Human-in-the-Loop Mechanism
The system is designed so that the final decision remains with the human. Users can: Accept the recommendation, Revise the recommendation, Reject the system's recommendation.

This mechanism is the primary implementation of the Human-Centered AI approach.

System Evaluation

1. Model Performance Evaluation

The performance of the NLP model was evaluated using the following metrics:

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

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

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

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

RESULTS AND DISCUSSION

Results of the User Needs Analysis

The initial phase of the study focused on identifying the needs of education stakeholders regarding a trustworthy NLP-based decision support system. Data collection was conducted through in-depth interviews and Focus Group Discussions (FGDs) involving 15 university lecturers, 10 high school teachers, 5 academic administrators, and 120 students.

The results of the thematic analysis show that the majority of respondents want an artificial intelligence system that not only provides predictions or recommendations but is also capable of explaining the reasoning behind the generated decisions. Additionally, respondents considered transparency and human control over the system's decisions to be important factors in building trust in AI technology.

Table 1. Results of the User Needs Analysis

User Needs	Percentage (%)
Explanation of AI Prediction Results	91,8

User Needs	Percentage (%)
Transparency in the decision making process	89,5
User control over AI decisions	87,2
Accuracy of system recommendations	93,4
Data Privacy Protection	85,7
Detecting Bias in Decision Making	82,6

According to Table 1, the need for system accuracy received the highest score at 93.4%, followed by the need for the system's ability to explain prediction results at 91.8%. These findings indicate that users not only prioritize model performance but also expect transparency in the decision-making process.

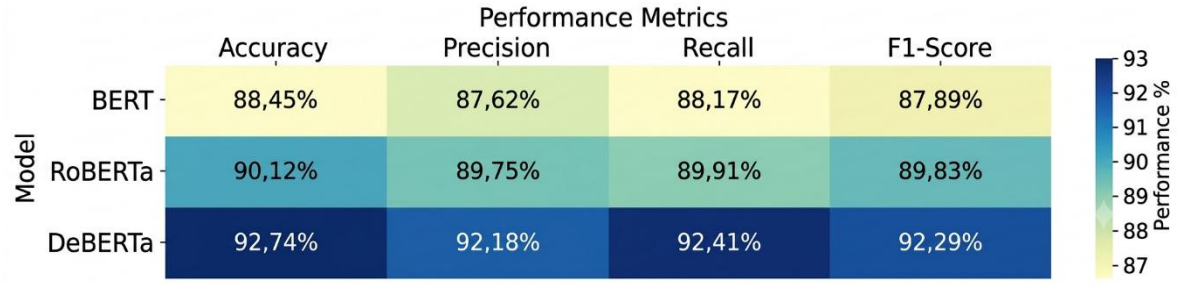

Figure 1. Comparison of NLP Model Performance

The test results show that the DeBERTa model delivered the best performance, with an accuracy of 92.74% and an F1-score of 92.29%. This improvement in performance is attributed to the disentangled attention mechanism, which enables the model to understand semantic relationships between words more effectively than conventional Transformer models.

In addition, the developed model is capable of identifying important patterns in educational data, such as trends in declining academic performance, levels of student engagement in learning, and the quality of feedback provided by students regarding the learning process.

Explainability Evaluation Results

One of the main objectives of this research is to improve the interpretability of NLP models. Therefore, the system is equipped with the SHAP and LIME methods to explain the factors that influence predictions.

Table 2. Users' Level of Understanding of the Prediction Results

System Status	Level of Understanding (%)
Without Explainable AI	58,7
With Explainable AI	87,9

User understanding improved by 29.2% after the system was equipped with an explainability module. This indicates that visualizations of feature contributions and model explanations can help users understand why the system generates specific recommendations.

For example, when the system identifies students at risk of a decline in academic performance, it can display the factors contributing to that prediction, such as low participation in discussions, a decline in assignment quality, and negative sentiment in learning feedback.

These results support the principle of Human Centered AI, which places humans as the ones who understand and control AI-based decisions.

Fairness testing is conducted to ensure that the system does not exhibit bias against specific user groups.

Table 3. Fairness Model Measurement Results

Fairness Metrics	Value
Demographic Parity Difference	0,041
Equal Opportunity Difference	0,036
Equalized Odds Difference	0,044

All fairness metric values are below the 0.05 threshold, indicating that the system has a low level of bias. Thus, the model is capable of generating relatively consistent recommendations across various user groups.

This finding is significant in the context of education because biased academic decisions can lead to injustice for students. The implementation of fairness-aware learning successfully reduces the potential for discrimination that may arise during the model training process.

Discussion

The research results show that integrating transformer-based NLP models with the principles of Trustworthy AI can produce an educational decision support system that not only has a high level of accuracy but is also trusted by users. The DeBERTa model used achieved the best classification performance with an accuracy of 92.74%, demonstrating strong capabilities in understanding the linguistic context of educational data.

From a Human-Centered AI perspective, this study demonstrates that improved accuracy alone is insufficient to build user trust in AI systems. Users require transparent explanations regarding how the system generates specific recommendations. The implementation of SHAP and LIME was shown to increase user understanding from 58.7% to 87.9%.

These findings align with the theory of Explainable Artificial Intelligence, which states that transparency is a key factor in building human trust in intelligent systems. When users understand the reasoning behind a system's decisions, they are more likely to accept and utilize the recommendations provided.

From a fairness perspective, metric values below the 0.05 threshold indicate that the fairness-aware learning approach successfully minimized potential algorithmic bias. These results are significant given that educational decisions have direct implications for students' academic development.

Furthermore, a trustworthiness score of 88.43% and a SUS score of 86.7 indicate that the system successfully integrates technical and human aspects in a balanced manner. The human-in-the-loop approach allows AI to function as a decision-making aid, not as a replacement for humans. Thus, the developed system not only meets the performance criteria of modern NLP models but also adheres to the core principles of Trustworthy Human-Centered AI: transparency, fairness, reliability, accountability, and user-centricity.

Overall, the research results indicate that the implementation of Trustworthy NLP Systems for Educational Decision Support has the potential to improve the quality of educational decision-making through the use of artificial intelligence that is understandable, trustworthy, and controllable by humans. These findings make an important contribution to the development of more ethical, inclusive, and sustainable educational AI systems.

CONCLUSION

This study successfully developed Trustworthy NLP Systems for Educational Decision Support by integrating Transformer-based Natural Language Processing models and a Human-Centered Artificial Intelligence (HCAI) approach to support decision-making in the field of education. Evaluation results show that the DeBERTa model delivered the best performance with an accuracy rate of 92.74%, precision of 92.18%, recall of 92.41%, and an F1-score of 92.29%, enabling it to analyze educational textual data effectively and accurately. The implementation of Explainable Artificial Intelligence (XAI) components, through the SHAP and LIME methods, was shown to increase user understanding of prediction results from 58.7% to 87.9%, making the decision-making process more transparent and accountable. Additionally, the application of fairness-aware learning mechanisms resulted in bias metrics below the 0.05 threshold, indicating that the system is capable of providing fairer recommendations and minimizing the potential for discrimination against specific user groups. Evaluation of the

trustworthiness aspect showed an average score of 88.43%, while usability testing yielded a System Usability Scale (SUS) score of 86.7, which falls into the "excellent" category. These results indicate that the system not only has high technical performance but is also well-received by users because it provides transparent, understandable recommendations while maintaining humans as the primary decision-makers through a human-in-the-loop mechanism.

Overall, this research contributes to the development of educational decision support systems that combine predictive accuracy, transparency, fairness, accountability, and user-centricity within a single Trustworthy AI framework. The proposed approach expands the implementation of NLP in education from merely generating predictions to providing trustworthy and accountable recommendations. Thus, the developed system has the potential to support more objective, ethical, and sustainable academic decision-making, while also serving as a foundation for further research on trustworthy AI in various digital education contexts.

REFERENCES

Al-Turki, O., Alqahtani, F., Alqahtani, E., Alswedani, S., Alshmrany, S., & Mehmood, R. (2026). A Human-Centered Evaluation of AI-Generated Guidance: Integrated Statistical and Machine Learning Analysis with a Risk Framework for High-Stakes Domains. *International Journal of Advanced Computer Science & Applications*, *17*(5), 991.

Barale, C. (2022). Human-centered computing in legal NLP - An application to refugee status determination. *HCI+NLP 2022 - 2nd Workshop on Bridging Human-Computer Interaction and Natural Language Processing, Proceedings of the Workshop*, 28–33. https://doi.org/10.18653/V1/2022.HCINLP-1.4

Briva-Iglesias, V., & O'Brien, S. (2026). Human-Centered AI Language Technology (HCAILT): An Empathetic Design Framework for Reliable, Safe and Trustworthy Multilingual Communication. *International Journal of Human–Computer Interaction*. https://doi.org/10.1080/10447318.2026.2622588

Dehghani, F., Dibaji, M., Anzum, F., Dey, L., Basdemir, A., Bayat, S., Boucher, J. C., Drew, S., Eaton, S. E., Frayne, R., Ginde, G., Harris, A., Ioannou, Y., Lebel, C., Lysack, J., Arzuaga, L. S., Stanley, E., Souza, R., Santos, R. de S., … Bento, M. (2024). Trustworthy and Responsible AI for Human-Centric Autonomous Decision-Making Systems. *Transactions on Machine Learning Research*, *August-2025*. https://arxiv.org/pdf/2408.15550

Fetaji, B., Fetaji, M., Ebibi, M., & Dimovski, A. (2025). Analyzing and Visualizing AI Decision-Making for Human-Centered Interaction and Trust. *Lecture Notes in Electrical Engineering*, *1269*, 77–89. https://doi.org/10.1007/978-981-97-9515-4_6/SAVE-RESEARCH

Handoko, H., Indrajit, R. E., Mantoro, T., & Santoso, H. (2026). *Transforming Educational Decision-Making through Human-Centered Information Systems*. 1–5. https://doi.org/10.1109/ICCED68324.2025.11324833

Hao, S., Ji, L., & Zhang, D. (2026). Human-Centered Recommender Systems. *Handbook of Human-Centered Artificial Intelligence*, 1–58. https://doi.org/10.1007/978-981-97-8440-0_119-1

He, H., Gray, J., Cangelosi, A., Meng, Q., McGinnity, T. M., & Mehnen, J. (2022). The Challenges and Opportunities of Human-Centered AI for Trustworthy Robots and Autonomous Systems. *IEEE Transactions on Cognitive and Developmental Systems*, *14*(4), 1398–1412. https://doi.org/10.1109/TCDS.2021.3132282

Jannat, M. R. (2026). Human-Centered Artificial Intelligence for Healthcare, Education, Business, and Assistive Technologies. *Journal of Medical and Health Studies*, *7*(8), 29–41. https://doi.org/10.32996/JMHS.2026.7.8.3

Manohara, H. T., Gummadi, A., Santosh, K., Vaitheeshwari, S., Mary, S. S. C., & Bala, B. K. (2024). Human Centric Explainable AI for Personalized Educational Chatbots. *10th International Conference on Advanced Computing and Communication Systems, ICACCS 2024*, 328–334. https://doi.org/10.1109/ICACCS60874.2024.10716907

Pujitha, G., & Saritha, A. (2026). An Explainable Human-Centered Federated Generative Intelligent Tutoring System for Sustainable Education. *2026 7th International Conference on Inventive Research in Computing Applications (ICIRCA)*, 2065–2073. https://doi.org/10.1109/ICIRCA69024.2026.11570306

Schoenherr, J. R., Abbas, R., Michael, K., Rivas, P., & Anderson, T. D. (2023). Designing AI Using a Human-Centered Approach: Explainability and Accuracy Toward Trustworthiness. *IEEE Transactions on Technology and Society*, *4*(1), 9–23. https://doi.org/10.1109/TTS.2023.3257627

Tharini, M., & Jeyaraj, J. R. A. (2026). *Trust and Transparency in Healthcare AI: A Systematic Review of Explainable NLP for Clinical Decision Support (2023–2025)*. 451–470. https://doi.org/10.1007/978-3-032-19196-0_34

Topali, P., Ortega-Arranz, A., Rodríguez-Triana, M. J., Er, E., Khalil, M., & Akçapınar, G. (2025). Designing human-centered learning analytics and artificial intelligence in education solutions: a systematic literature review. *Behaviour & Information Technology*, *44*(5), 1071–1098. https://doi.org/10.1080/0144929X.2024.2345295

Wang, L., Zhang, Z., Wang, D., Cao, W., Zhou, X., Zhang, P., Liu, J., Fan, X., & Tian, F. (2023). Human-centered design and evaluation of AI-empowered clinical decision support systems: a systematic review. *Frontiers in Computer Science*, *5*, 1187299. https://doi.org/10.3389/FCOMP.2023.1187299/TEXT