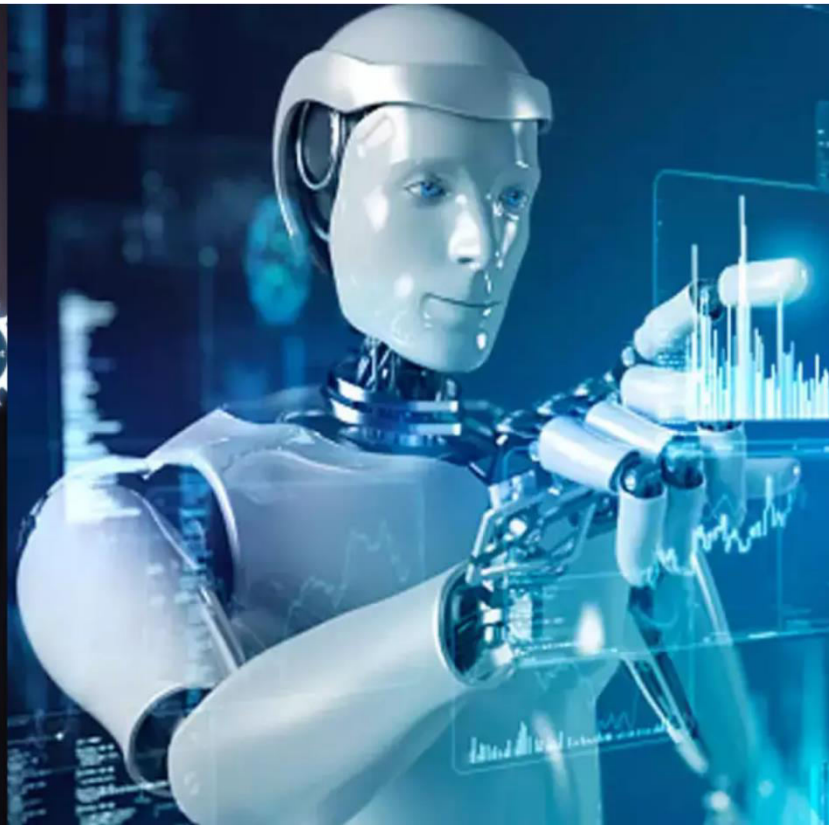
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 15, Issue 1, January 2026

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 1, January 2026

|DOI: 10.15680/IJIRSET.2026.1501036|

Intent-Based Data Preprocessing using Natural Language Understanding

Poreddy Vara Sreeja

M. Tech (Data Science), Department of Data Science, GITAM University, Hyderabad, India

ABSTRACT: This study presents a framework for data preprocessing based on user-defined analytical intent expressed in natural language, which adapts preprocessing strategies dynamically. By utilizing sentence embeddings [1] and cosine similarity [3], the system connects free-text intents to established preprocessing pipelines, facilitating automated and transparent data preparation. The methodology incorporates techniques for imputing missing values [5], encoding categorical data [4], and scaling adaptively, all customized to various analytical goals, such as prediction, comparison, visualization, and outlier detection. Evaluations conducted with the Iris dataset [10] highlight the proposed framework's adaptability and efficiency in data analytics focused on human needs.

KEYWORDS: Intent-Based Preprocessing, Natural Language Understanding, Sentence Embeddings, Explainable Data Processing, Machine Learning

I. INTRODUCTION

Data preprocessing serves as a crucial element in data science and machine learning processes, significantly impacting model efficacy and analytical precision [6], [7]. Traditional preprocessing methods are predominantly fixed and necessitate manual setup, which requires both specialized knowledge and a considerable investment of time [8]. Furthermore, these methods usually overlook the user's analytical goals, such as whether the task is focused on prediction, comparison, visualization, or anomaly detection [5].

Recent progress in natural language processing (NLP) allows systems to grasp user intentions conveyed in natural language [2], [23]. Inspired by this ability, this paper introduces an intent-driven preprocessing framework that automatically determines and implements suitable preprocessing steps by understanding user intent through sentence embeddings and cosine similarity [1], [3]. The aim is to minimize manual effort while enhancing adaptability and clarity in data preparation processes.

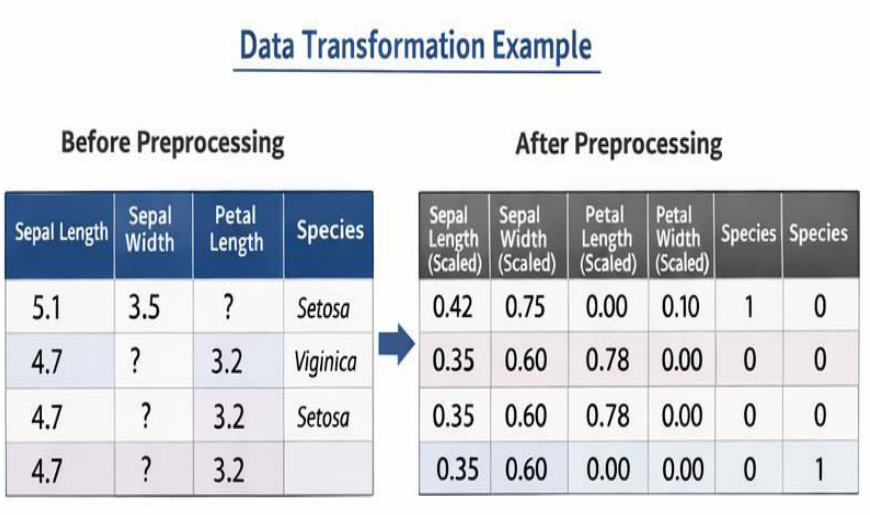


Fig 1: Data Transformation

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 1, January 2026

|DOI: 10.15680/IJIRSET.2026.1501036|

II. SYSTEM OVERVIEW

The suggested system takes in two types of inputs: a structured dataset and a user-specified analytical intent conveyed in natural language. This intent is converted into a semantic representation with the help of a pre-trained sentence embedding model [1]. The resulting representation is then compared with the embeddings of known analytical intents using cosine similarity [3] to find the closest match. Once an intent is matched, a corresponding preprocessing pipeline is executed on the dataset. Furthermore, the system keeps a record of the transformations applied to ensure transparency and interpretability [13].

III. PROPOSED METHODOLOGY

A. REPRESENTING INTENT WITH SENTENCE EMBEDDINGS.

User intent is transformed into dense vector representations using a pre-trained Sentence-BERT model (all-MiniLM-L6-v2) [1]. In contrast to keyword-based methods, sentence embeddings encapsulate semantic meaning, enabling the system to adapt to diverse phrasing of analytical objectives.

B. MECHANISM FOR INTENT MATCHING.

Cosine similarity is employed to assess the semantic similarity between the user intent embedding and the embeddings of a set of predefined known intents [3]. The known intent that exhibits the highest similarity score is chosen as the corresponding analytical objective.

C. MAPPING INTENT TO PREPROCESSING STEPS.

Each known intent is linked to a specific preprocessing pipeline [4], [5]. This mapping allows the system to autonomously select transformations that align with the desired analytical goals.

TABLE 1: TECHNIQUES USED IN VARIOUS RESEARCH

INTENT TYPE	PREPROCESSING STEPS
Prediction	Mean imputation, Min-Max scaling [5], [18]
Comparison	Mean imputation, Standard scaling, One-hot encoding [4], [5]
Outlier Detection	Mean imputation, Robust scaling [5]
Visualization	Zero imputation

D. EXECUTION OF AUTOMATED PREPROCESSING.

Following the identified intent, the system methodically implements the necessary preprocessing operations in order. Numeric features undergo suitable imputation and scaling methods [5], [18], while categorical features are transformed using one-hot encoding [4]. Every step taken is documented to enhance transparency [13].

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 1, January 2026

|DOI: 10.15680/IJIRSET.2026.1501036|

IV. ALGORITHM

Step 1: Provide the dataset D along with the user intent I.
Step 2: Transform intent I into a sentence embedding.
Step 3: Calculate the cosine similarity with established intent embeddings.
Step 4: Identify the intent that has the highest similarity score.
Step 5: Obtain the preprocessing steps corresponding to the chosen intent.
Step 6: Sequentially execute the preprocessing steps on dataset D.
Step 7: Deliver the modified dataset along with an explanation log.

V. EXPERIMENTAL SETUP

A. Dataset

The Iris dataset [10] served as the basis for experimental evaluation. This dataset consists of 150 instances, encompassing 4 numerical attributes (sepal length, sepal width, petal length, and petal width) and 1 categorical class label representing three species of flowers.

B. Experimental Configuration

Four distinct free-text analytical intents were assessed. For each intent, the framework produced an intent embedding [1] and compared it with 7 known intents using cosine similarity [3]. Preprocessing pipelines were carried out in sequence based on the identified intent [4], [5]. All experiments were performed in Python using Pandas, Scikit-learn, and Sentence-Transformers [4].

VI. RESULTS AND DISCUSSION

A. Performance in Intent Matching.

The framework achieved a 100% success rate in intent matching across all evaluated intents, reflecting strong semantic comprehension of user intent [1], [3].

B. Applied Preprocessing Operations.

Table II provides a summary of the quantity and types of preprocessing operations utilized for each analytical intent.

TABLE 2: NUMERIC SUMMARY OF PREPROCESSING OPERATIONS

User Intent	No. of Steps	Preprocessing Operations
Predict future species distribution	2	Mean imputation, Min-Max scaling [5], [18]
Compare petal lengths across species	3	Mean imputation, Standard scaling, One-hot encoding [4], [5]
Detect unusual measurements	2	Mean imputation, Robust scaling [5]
Visualize data for report	1	Zero imputation

C. Impact of Dataset Transformation.

After preprocessing:

- All missing numeric values were completely addressed through imputation [5].
- Numerical features were normalized to suitable ranges according to the analytical goals [18].
- Categorical variables were converted into binary features as needed [4].
- The explanation log documented all the preprocessing steps undertaken, ensuring clarity [13].

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 1, January 2026

|DOI: 10.15680/IJIRSET.2026.1501036|

D. Analysis.

The numerical outcomes demonstrate that the framework adjusts the complexity of preprocessing in relation to the analytical goals. Tasks focused on comparison necessitated the greatest number of preprocessing actions, whereas visualization tasks needed minimal alterations. This behavior corresponds with data analytics best practices and underscores the efficacy of preprocessing that considers intent [5], [13].

VII. CONCLUSION

This paper introduces an intent-based data preprocessing framework that dynamically adapts data transformation techniques in real-time based on analytical goals articulated in natural language [1], [3]. By utilizing sentence embeddings and similarity-driven intent matching, the system facilitates automated, adaptable, and interpretable preprocessing workflows. Quantitative assessment using the Iris dataset [10] confirms the framework's efficacy and versatility. Future work could focus on learning new intents dynamically and incorporating domain-specific preprocessing strategies.

REFERENCES

[1] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," Proc. EMNLP, pp. 3982–3992, 2019.
[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," Proc. NAACL-HLT, pp. 4171–4186, 2019.
[3] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," Proc. ICLR, 2013.
[4] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
[5] S. García, J. Luengo, and F. Herrera, Data Preprocessing in Data Mining, Springer Briefs in Computer Science, 2015.
[6] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, 2016.
[7] K. P. Murphy, Machine Learning: A Probabilistic Perspective, MIT Press, 2012.
[8] M. Z. H. Khan and S. U. Khan, "A survey of data preprocessing techniques in data mining," Int. J. Computer Applications, vol. 65, no. 19, pp. 28–35, 2013.
[9] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd ed., O'Reilly Media, 2019.
[10] R. A. Fisher, "The use of multiple measurements in taxonomic problems," Annals of Eugenics, vol. 7, no. 2, pp. 179–188, 1936.
[11] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval, Cambridge Univ. Press, 2008.
[12] C. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.
[13] R. Guidotti et al., "A survey of methods for explaining black box models," ACM Computing Surveys, vol. 51, no. 5, pp. 1–42, 2018.
[14] S. Amershi et al., "Software engineering for machine learning: A case study," Proc. ICSE, pp. 291–300, 2019.
[15] L. Rokach, "Data mining with decision trees: Theory and applications," World Scientific, 2014.
[16] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed., Morgan Kaufmann, 2012.
[17] B. Liu, Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, Springer, 2011.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details