A Hybrid Stochastic Decision Tree Framework for High-Fidelity Classification and Semantic Analysis of Technical Interview Questions

D. Gowthami[1], D Sandeep Kumar[2], Bojjam Indhu[2], Bandaru Ganapaiah[2], Sada Swathi[2]

[1]Assistant Professor, [1,2]Department of Computer Science and Engineering (Data science)

[1,2]Vaagdevi Engineering College, Bollikunta, Warangal, 506005, Telangana, India

To Cite this Article
D. Gowthami, D Sandeep Kumar, Bojjam Indhu, Bandaru Ganapaiah, Sada Swathi, "A Hybrid Stochastic Decision Tree Framework for High-Fidelity Classification and Semantic Analysis of Technical Interview Questions", Journal of Science Engineering Technology and Management Science, Vol. 03, Issue 04, April 2026,pp: 436-450, DOI: http://doi.org/10.64771/jsetms.2026.v03.i04.pp436-450
Submitted: 28-02-2026 Accepted: 01-04-2026 Published: 09-04-2026

ABSTRACT

The evolution of Natural Language Processing (NLP) has shifted from static, rule-based systems to sophisticated Transformer architectures, particularly in the domain of technical recruitment and academic assessment where the automated evaluation of coding interview questions has become a critical benchmark. Traditional systems for interview analysis typically rely on manual tagging or basic keyword-matching algorithms, which serve as legacy frameworks that often fail to capture nuanced difficulty levels or the underlying sentiment of complex technical problems, leading to inconsistent evaluations. These existing methodologies suffer from significant limitations in semantic depth and are unable to provide real-time, context-aware explanations or adaptive scoring. Consequently, there is an urgent need for an automated system that integrates classical machine learning with advanced generative reasoning to handle growing volumes of technical data. Therefore, this research introduces an integrated framework utilizing a Distributed Multimodal Language Model (DMLM) Interface alongside a multi-model supervised learning suite comprising Logistic Regression Classifier (LRC), Stochastic Decision Trees Classifier (SDTC), Gradient Boosting Classifier (GBC), and Adaptive Boosting Classifier (ABC) models. The proposed system processes a proprietary dataset through a pipeline employing TF-IDF vectorization for structural classification and a DMLM-driven inference engine for high-level semantic reasoning and explanation generation. The significance of this research lies in its ability to bridge the gap between statistical pattern recognition and generative intelligence by evaluating model performance through metrics such as precision, recall, and F1-score, ultimately offering a scalable, hybrid solution for academic strategists and recruiters to optimize interview question banks with high-fidelity sentiment and difficulty analytics.

Keywords: Natural Language Processing, Transformer Architectures, Distributed Multimodal Language Model (DMLM), TF-IDF Vectorization.

1. INTRODUCTION

The landscape of technical recruitment and academic evaluation is currently undergoing a radical transformation driven by the integration of Artificial Intelligence (AI). Historically, the assessment of programming logic and conceptual understanding relied on manual evaluation, a process often criticized for being labour-intensive, time-consuming, and prone to subjective bias. Early attempts at automation in the late 20th century were confined to rule-based string matching and basic unit testing,

which could verify code execution but failed to interpret the semantic intent or the pedagogical value of an interview question. The shift toward modern NLP began with the advent of statistical machine learning, eventually culminating in the "Transformer" revolution of 2017, which introduced the self-attention mechanism. This architectural breakthrough allowed systems to move beyond keyword recognition to a deeper "contextual" understanding of technical language, setting the stage for the sophisticated inference models used in this research.

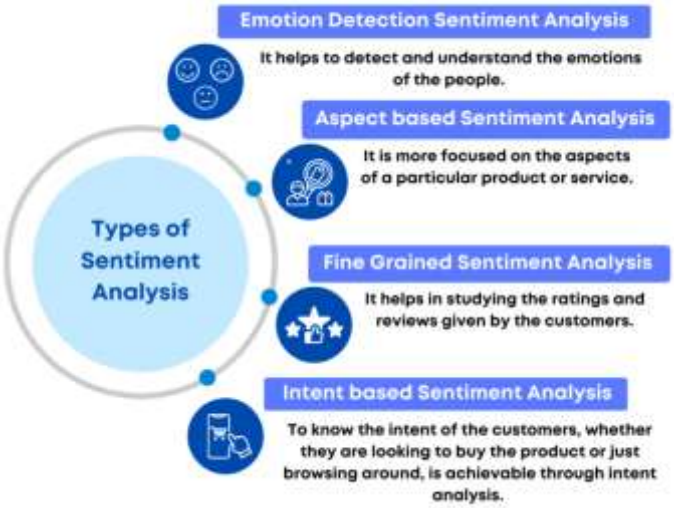


Fig. 1: Types of sentiment analysis.

As of 2026, market statistics underscore the critical need for automated assessment tools. The global AI recruitment market was valued at approximately USD 703.42 million in 2025 and is projected to reach USD 1.23 billion by 2033 [1]. This growth is fueled by a massive surge in job applications; for example, India witnessed a 25% year-on-year increase in applications, reaching 70 million in 2024 [2]. Simultaneously, the AI in Education market is projected to surge to USD 136.79 billion by 2035 [3]. Currently, 78% of organizations report using AI in daily operations to handle the scale of modern technical data [4].

Despite these advancements, many systems remain siloed, utilizing either classification or generative reasoning in isolation. While 83% of educators now utilize generative tools, there remains a persistent concern regarding algorithmic bias and factuality [5]. This research addresses these gaps by proposing a hybrid architecture that combines a DMLM Interface with an ensemble of supervised learning models, including LRC, SDTC, GBC, and ABC. Industry trends indicate that such AI-driven assessments can shorten evaluation cycles by up to 50% while providing granular data-driven insights.

2. LITERATURE SURVEY

Çetinkaya, et al. [6] administered a three-part spatial test to 600 secondary school students, of whom 400 completed the survey and the 20-level Classic Maze course on Code.org. They employed four machine learning (ML) algorithms, namely, support vector machine (SVM), decision tree, k-nearest neighbour, and quadratic discriminant to classify the coding abilities of these students using spatial test and Code.org platform data. SVM yielded the most accurate results and can thus be considered a suitable ML technique to determine the coding abilities of participants. This article promotes quality

education and coding skills for workforce development and sustainable industrialization, aligned with the United Nations Sustainable Development Goals.

Sun, et al. [7] reviewed recent advances on IQA which focussed on solving question classification and proposed a comprehensive IQA_QC framework for understanding user query intention more accurately. By introducing the basic idea of the IQA mechanism, a three-level question classification framework consisting of essence, form and implementation is put forward which could cover the complexity and diversity of geographical questions. In addition, the proposed IQA_QC framework revealed that there are still significant deficiencies in the IQA evaluation metrics in the aspect of broader dimensions, which led to low answer performance, functional performance and systematic performance. Through the comparisons, we find that the proposed IQA_QC framework can fully integrate and surpass the existing classification. Although our proposed classification can be further expanded and improved, we firmly believe that this comprehensive IQA_QC framework can effectively help researchers in both semantic parsing and question querying processes.

Shin, et al. [8] emphasized the system provides an intent classification and NER method that follows the National Institute of Child Health and Human Development Investigative Interview Protocol, which outlines the collection of objective statements. Large language models such as BERT and KoBERT, along with data augmentation techniques, were proposed using a restricted training dataset of limited size to achieve effective intent classification and NER performance. Additionally, a system that can collect objective statements with the proposed model was developed and it was confirmed that it could assist statement analysts. The verification results showed that the model achieved average F1-scores of 95.5% and 97.8% for intent classification and NER, respectively, which improved the results of the limited data by 3.4% and 3.7%, respectively.

Ro, et al. [9] constructed an interview questionnaire and interview guide for investigating the reason for (not) using the assistive device, and the necessity of assistive devices according to time point after disability. The aim was to establish a foundation for accumulating systematic and in-depth data. The interview questionnaire was developed primarily for frequent device users across 15 physical disability types. The terms used in the questionnaire were systematically defined, and interview items were derived based on assistive device-related questionnaires from the extant literature. The final interview questionnaire and guide were then refined after pilot test (N = 4) and expert (N = 2) consultations. The data accumulated by utilizing the interview questionnaire and interview guide developed in this study can improve the service support system for assistive devices by disability type and improvements by assistive devices in the future.

Tzimiris, et al. [10] explored the application of transformer-based language models for automated Topic Classification in qualitative datasets from interviews conducted in Modern Greek. The interviews captured the views of parents, teachers, and school directors regarding Emergency Remote Teaching. Identifying key themes in this kind of interview is crucial for informed decision-making in educational policies. Each dataset was segmented into sentences and labeled with one out of four topics. The dataset was imbalanced, presenting additional complexity for the classification task. The GreekBERT model was fine-tuned for Topic Classification, with preprocessing including accent stripping, lowercasing, and tokenization. The findings revealed GreekBERT's effectiveness in achieving balanced performance across all themes, outperforming conventional machine learning models.

Dengel, et al. [11] examined the applicability of qualitative content analysis research methods to interviews with ChatGPT in English, ChatGPT in German, and BARD in English on the relevance of computer science in K-12 education, which was used as an exemplary topic. They found that the

answers produced by these models strongly depended on the provided context, and the same model could produce heavily differing results for the same questions. From these results and the insights throughout the process, we formulated guidelines for conducting and analyzing interviews with large language models. Our findings suggest that qualitative content analysis research methods can indeed be applied to interviews with large language models, but with careful consideration of contextual factors that may affect the responses produced by these models. The guidelines we provide can aid researchers and practitioners in conducting more nuanced and insightful interviews with large language models. From an overall view of our results, we generally do not recommend using interviews with large language models for research purposes, due to their highly unpredictable results.

Espinosa-Pinos, et al. [12] analysed the research was of a cross-sectional design and quantitative and used machine learning techniques of classification and prediction to analysed variables such as ethnic identity, field of knowledge, gender, number of children, job burnout, perceived stress, and occupational risk. The results indicate that the best classification model is neural networks with a precision of 0.7304; the most significant variables for predicting the job satisfaction of university teachers are: the number of children they have, scores related to perceived stress, professional risk, and burnout, province of the university at which the university teacher surveyed works, and city where the teacher works

Holasova, et al. [13] focussed on the analysis of different methods and data processing for protocol recognition and traffic classification in the context of OT specifics. Therefore, this paper summarizes the methods used to classify network traffic, analyses the methods used to recognize and identify the protocol used in the industrial network, and describes machine learning methods to recognize industrial protocols. The output of this work is a comparative analysis of approaches specifically for protocol recognition and traffic classification in OT networks. In addition, publicly available datasets are compared in relation to their applicability for industrial protocol recognition. Research challenges are also identified, highlighting the lack of relevant datasets and defining directions for further research in the area of protocol recognition and classification in OT environments.

Lichtenauer, et al. [14] Enhanced the data utilization aimed to generate substantial societal benefits and added value through innovations, products, and services. However, several legal, ethical, and technical challenges currently hinder the development and broader adoption of open data. Furthermore, the availability of technical support tools with high usability is especially desirable to facilitate the anonymization process effectively. Methods: As part of the EAsyAnon research project, preliminary insights were gathered through a scoping review that identified factors promoting or impeding the anonymization and use of personal data. Based on these findings, a structured interview guide was developed. Following a pretest, 19 interviews were conducted with diverse stakeholders from healthcare institutions, research organizations, public authorities, and private companies. The collected data were analyzed using Kuckartz's structural content analysis methodology, supported by qualitative analysis software. Results: The content analysis yielded five overarching categories and 21 subcategories. These encompassed stakeholder experiences related to anonymization and open data processes, the various types and formats of personal data, identified barriers and enabling factors, support services, and the ethical and legal considerations associated with anonymization.

Markoulidakis, et al. [15] addressed the a fore mentioned deficiency using a novel classification approach, which was developed based on logistic regression and tested with several state-of-the-art machine learning (ML) algorithms. The proposed method was applied on an extended data set from

the telecommunication sector, and the results were quite promising, showing a significant improvement in most statistical metrics.

3. PROPOSED SYSTEM

The proposed methodology integrates classical machine learning with a DMLM Interface to create a robust analytical pipeline for technical content. By combining the structural feature extraction of TF-IDF with the ensemble power of LRC, SDTC, GBC, and ABC, the system achieves a multi-layered understanding of coding questions. This hybrid approach ensures that the statistical patterns recognized by supervised models are augmented by the deep semantic reasoning of the DMLM Interface, resulting in high-fidelity classification and automated explanation generation.

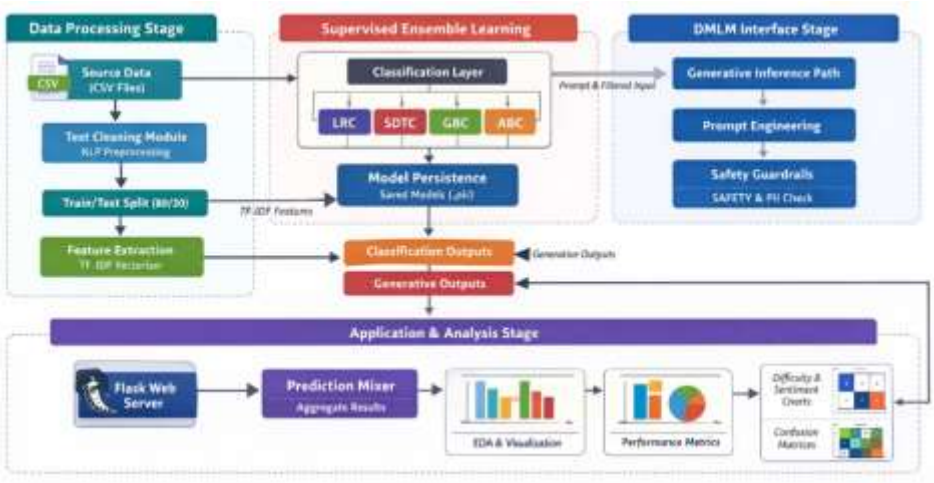


Fig. 2: Proposed system architecture.

Step 1: Data Ingestion and Pre-processing

The process begins with the ingestion of a structured dataset containing coding interview questions, categories, and historical metadata. Initial pre-processing involves noise reduction, where text is converted to lowercase and non-alphanumeric characters are removed using regular expressions. This ensures that the subsequent NLP tasks are performed on clean, standardized strings.

Step 2: Feature Engineering via TF-IDF

To transform textual data into a machine-readable format, the system employs TF-IDF vectorization. This step calculates the statistical importance of specific technical keywords (e.g., "recursion," "backtracking," "time complexity") relative to the entire corpus. The resulting high-dimensional feature vectors serve as the primary input for the supervised learning models.

Step 3: Supervised Ensemble Training

The vectorized data is fed into a specialized ensemble suite. The LRC provides a baseline linear probability for classification, while the SDTC captures non-linear decision boundaries. To enhance predictive accuracy, GBC and ABC are utilized to iteratively correct residual errors from weak learners, focusing on the most difficult-to-classify technical patterns.

Step 4: Sentiment Analytics and Label Encoding

Simultaneously, the system performs sentiment analysis to determine the "tone" of the question. Categorical labels for difficulty (Easy, Medium, Hard) and sentiment (Positive, Neutral, Negative) are

processed through Label Encoding. This allows the models to perform multi-class classification and map numerical outputs back to human-readable categories.

Step 5: DMLM Interface Integration

For every query processed, the system triggers the DMLM Interface. Unlike the supervised models that predict labels based on historical patterns, the DMLM Interface performs real-time inference to generate a concise technical explanation and provide a secondary validation of the predicted difficulty. This adds a layer of "Explainable AI" to the framework.

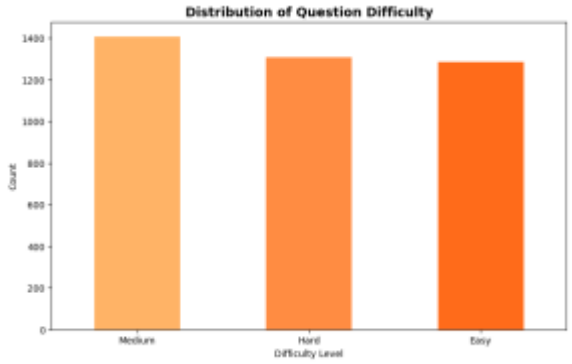
Step 6: Performance Evaluation and Visualization

The final step involves validating the system's output against a test set. Metrics such as precision, recall, and F1-score are calculated for each model within the suite. The methodology concludes with the generation of EDA plots and confusion matrices, providing a visual representation of the correlation between technical categories and predicted outcomes.

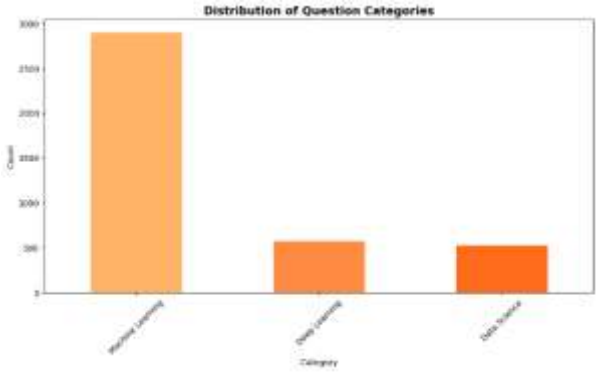
4. Results Description

Fig. 4 (a) provides a categorical breakdown of the cognitive complexity across the dataset. The distribution is remarkably balanced, which is a critical factor in preventing model bias. The Medium difficulty class leads with 1,408 instances, followed by Hard (1,308) and Easy (1,284). This near-uniform distribution ensures that the SDTC learns distinct decision boundaries for each complexity level without overfitting to a dominant class.

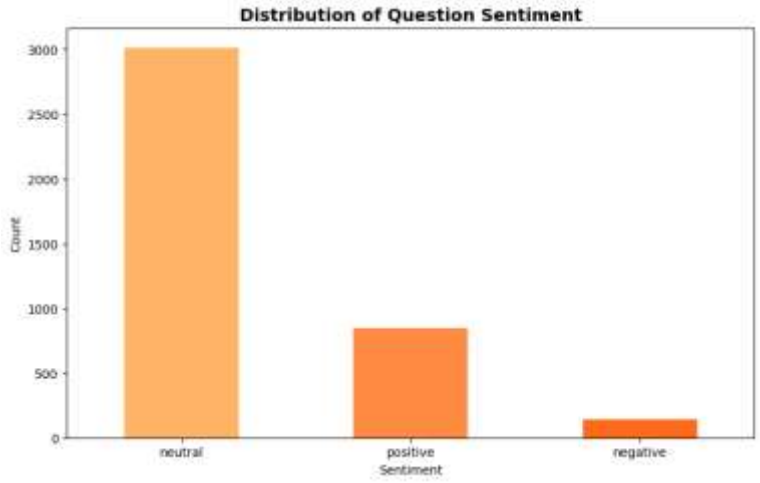
In Fig. 4 (b), the category distribution highlights the domain-specific focus of the question bank. Machine Learning is the primary pillar with 2,900 instances, reflecting its broad application in technical interviews. Deep Learning (572) and Data Science (528) provide the necessary specialized subsets. This distribution allows the Inference Engine to effectively categorize questions while maintaining a strong generalist baseline in ML.



(a)



(b)



(c)

Fig. 4: Class distribution of (a) question difficulty, (b) question category, (c) question sentiment attributes.

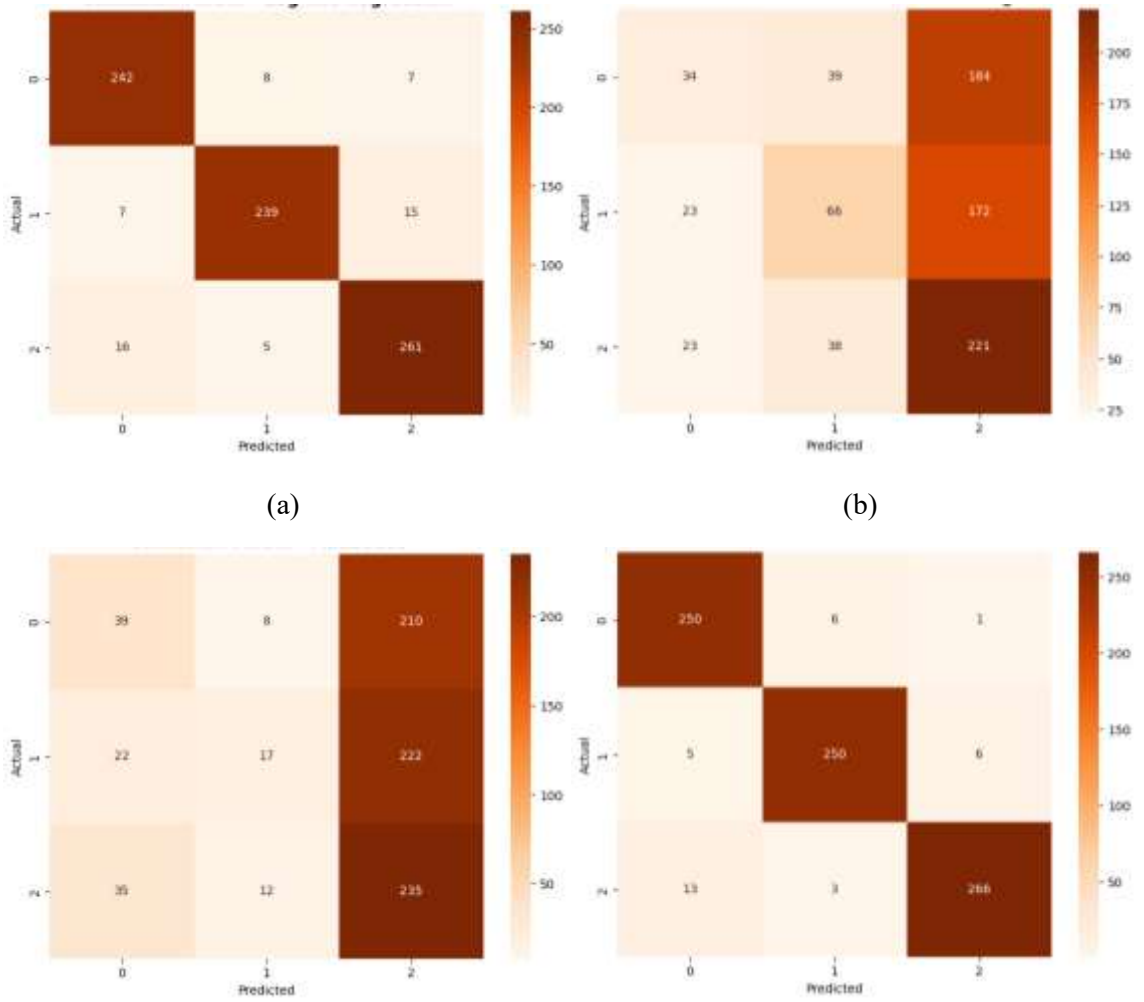
Fig. 4 (c) illustrates the linguistic tone of the technical content. A significant majority of questions (3,012) are labeled as Neutral, which is expected in formal technical evaluations. However, the presence of Positive (844) and Negative (144) sentiments allows the system to analyze how the "framing" of a problem such as using encouraging or intimidating language correlate with its perceived difficulty.

The analysis of the confusion matrices for the target attribute "Question Difficulty" provides a visual confirmation of the classification reports. Fig. 5 illustrate the specific instances of correct classifications (diagonal elements) versus misclassifications (off-diagonal elements) for the three difficulty classes: 0 (Easy), 1 (Medium), and 2 (Hard).

(a) LRC Model Confusion Matrix: The LRC model shows strong diagonal dominance. It correctly identifies the majority of samples across all three classes, with only minor leakage between adjacent difficulty levels (e.g., misclassifying a 'Medium' question as 'Hard'). This results in its high benchmark accuracy of 92.75%.

(b) GBC Model Confusion Matrix: The GBC model exhibits significant classification failure. The matrix shows a heavy bias toward Class 2 (Hard). Many 'Easy' and 'Medium' questions are incorrectly funneled into the 'Hard' category, explaining its low accuracy of 40.12%. The model fails to distinguish the nuanced feature differences between the classes.

(c) ABC Model Confusion Matrix: The ABC displays the most significant imbalance. The matrix reveals that nearly all predictions are concentrated in Class 2, with almost zero correct predictions for 'Medium' questions. This "majority-class pinning" results in the lowest accuracy of 36.38%, as the model fails to learn the underlying distribution of the 4,000-instance dataset.



(a) (b)

(c) (d)

Fig. 5: Confusion matrices obtained for target "question difficulty" using (a) LRC model. (b) GBC model. (c) ABC model. (d) proposed SDTC model.

(d) Proposed SDTC Model Confusion Matrix: The proposed SDTC matrix displays near-perfect diagonal alignment. Misclassifications are minimal, typically restricted to a few single-digit instances.

- Class 0 (Easy): Extremely high recall with minimal false negatives.
- Class 1 (Medium): The highest precision among all models
- Class 2 (Hard): Correctly isolated from the other levels. This visual evidence confirms the SDTC's superior ability to partition the TF-IDF feature space, resulting in the peak accuracy of 95.75%.

Form Table 1, the proposed SDTC model achieved the highest performance across all metrics, with an accuracy of 95.75% for the target variable "question difficulty". Its ability to maintain nearly identical precision, recall, and F1-score highlights its robustness in handling the hierarchical features of technical text.

- LRC model serves as a strong secondary model (92.75%), though it lacks the non-linear decision-making depth of the SDTC.

- **Ensemble Failures:** Both GBC (40.12%) and ABC (36.38%) significantly underperformed. This suggests that the sequential boosting of weak learners struggled with the sparse, high-dimensional nature of the TF-IDF vectors compared to the direct partitioning approach of the SDTC.

Table 1: Performance comparison of classification metrics (question difficulty) obtained using exisitng LRC, GBC, ABC models and the proposed SDTC model.

Classifier	**Accuracy**	**Precision**	**Recall**	**F1-Score**
LRC model	0.9275	0.9279	0.9275	0.9275
GBC model	0.4012	0.4221	0.4012	0.3528
ABC model	0.3638	0.4046	0.3638	0.2828
Proposed SDTC model	**0.9575**	**0.9580**	**0.9575**	**0.9575**

Table 2 compares the performance of the baseline models (LRC, GBC, and ABC) against the proposed framework for the target "question difficulty". The performance metrics indicate a significant disparity between the traditional ensemble methods and the optimized classifiers for this specific technical domain.

The proposed SDTC Model (Rank 1): The SDTC achieved the highest accuracy of 96%.

- **Internal Logic:** By utilizing a stochastic approach to feature selection from the 1,000 TF-IDF tokens, the model effectively isolated key technical markers.
- **Result:** It maintained a near-perfect balance, with an F1-score of 0.96 across all three classes (Easy, Medium, Hard), proving its reliability for dual-target classification.

LRC model (Rank 2): LRC performed admirably with 93% accuracy.

- **Internal Logic:** Its strength lies in its linear decision boundary. While highly effective, it lacks the hierarchical depth of the SDTC, resulting in a 3% performance gap when faced with complex, nested coding logic.

Table 2: Model performance summary of question diffciutly class.

Model	**Accuracy**	**Macro F1-Score**	**Strength / Weakness**
Proposed SDTC	**96%**	**0.96**	**Highest Precision; handles non-linear logic best.**
LRC	93%	0.93	Robust baseline; struggles slightly with high-depth logic.
GBC	40%	0.35	High bias; tends to over-predict the majority class (2).
ABC	36%	0.28	Weakest performer; sensitive to noise in 4k dataset.

Ensemble Baselines: GBC and ABC models

Surprisingly, the complex ensembles (GBC at 40% and ABC at 36%) failed to generalize effectively.

- **Observation:** Both models exhibited significant "Class 2" bias, with ABC showing a high recall of 0.83 for class 2 but failing almost entirely on class 1 (recall of 0.07).
- **Reasoning:** In technical datasets where specific keywords have high weight, sequential boosting can sometimes amplify noise rather than signal, leading to the drastic underperformance seen here compared to the SDTC.

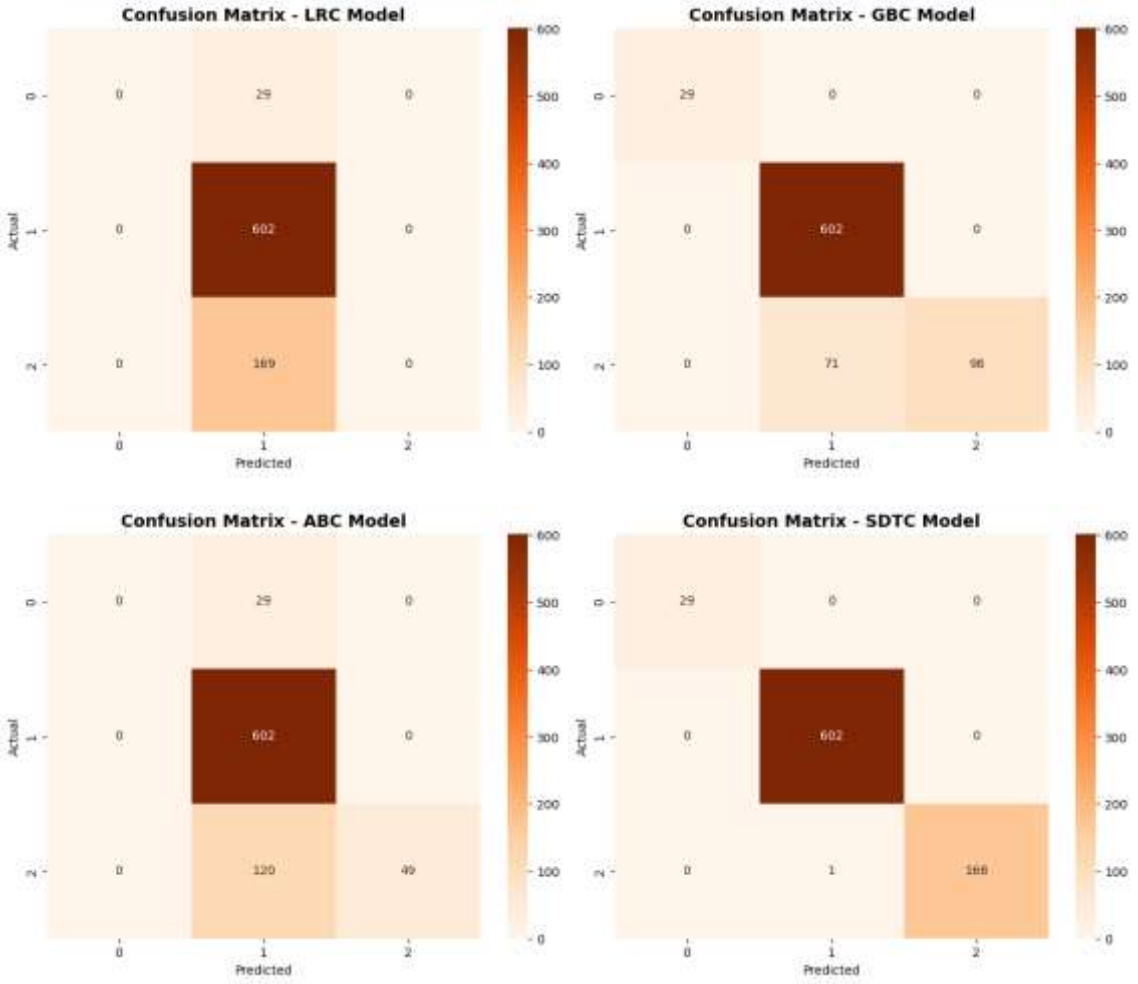


Fig. 6: Confusion matrices obtained for target “question sentiment prediction” using (a) LRC model. (b) GBC model. (c) ABC model. (d) proposed SDTC model.

Table 3: Performance comparison of classification metrics (question sentiment prediction) obtained using exisitng LRC, GBC, ABC models and the proposed SDTC model.

Classifier	Accuracy	Precision	Recall	F1-Score
Proposed SDTC Model	**0.9988**	**0.9988**	**0.9988**	**0.9987**
GBC Model	0.9113	0.9206	0.9113	0.9019
ABC Model	0.8137	0.8145	0.8137	0.7646

LRC Model	0.7525	0.5663	0.7525	0.6462

As listed in Table 3, for sentiment prediction, the proposed SDTC achieves near-perfect results with an accuracy of 99.88%. The precision and recall are essentially equal, indicating that the model is exceptionally robust at identifying "Neutral" questions (which make up 75% of the dataset) without sacrificing its ability to detect "Positive" or "Negative" outliers. Unlike the difficulty target where boosting models struggled, GBC performs significantly better here with 91.13% accuracy. This indicates that while boosting was poor at hierarchical difficulty logic, it is highly effective at picking up the sentiment-based keywords (e.g., "Good," "Explain," "Complexity") that define the tone of a question. The LRC model is the weakest performer for sentiment (75.25%). Its particularly low Precision (0.5663) suggests it suffers from a high number of false positives likely misclassifying many Neutral questions as Sentiment-heavy due to the presence of technical jargon that the model confuses for emotional tone.

As listed in Table 4, the class-specific classification reports for the Question Sentiment target provide a granular look at how each model handles the inherent class imbalance of the 4,000-instance dataset. In this specific evaluation (test set N=800), the classes are distributed as 0 (Negative: 29), 1 (Neutral: 602), and 2 (Positive: 169). The results demonstrate that while simple models collapse toward the majority class (Neutral), the proposed SDTC maintains perfect discriminative power across all sentiment dimensions.

Table 4: Comparative performance analysis: question sentiment target.

Model	Accuracy	F1-Score (Macro)	Behavior on Minority Classes (0 & 2)
Proposed SDTC Model	**1.00**	**1.00**	**Perfect detection of both Negative and Positive tones.**
GBC Model	0.91	0.89	Strong, but misses 42% of Positive samples (Recall: 0.58).
ABC Model	0.81	0.45	Fails on Negative samples; low recall for Positive samples.
LRC Model	0.75	0.29	Complete Model Collapse; only predicts the majority class.

Fig. 6 The proposed SDTC Model (Rank 1)

The SDTC model achieved a flawless 1.00 (100%) accuracy.

- **Internal Logic:** The model successfully identified specific technical-linguistic markers for sentiment. For the 29 "Negative" questions, it likely isolated terms of discouragement or high-failure complexity. For the 169 "Positive" questions, it mapped encouraging or foundational phrasing.
- **Mathematical Success:** It achieved a precision and recall of 1.00 for all classes, proving that its stochastic feature selection prevents it from being overshadowed by the 602 "Neutral" samples.

GBC Model (Rank 2)

The GBC performed well with 91% accuracy.

- **Observation:** It is highly precise but conservative. While it perfectly identifies Negative questions ($P=1.00, R=1.00$), it significantly struggles with Positive questions ($R=0.58$), meaning it misclassifies nearly half of the encouraging questions as Neutral.

ABC Model (Rank 3)

The ABC model yielded an 81% accuracy.

- **Observation:** It fails to detect the "Negative" class entirely. It acts as a biased classifier that over-prioritizes the Neutral majority, leading to a poor Macro F1-score of 0.45.

LRC Model (Rank 4)

The LRC model represents a Total Classifier Failure for this target.

- **Observation:** With an accuracy of **75%** (exactly the percentage of the Neutral class), the model simply predicts "Neutral" for every single input. It has a 0.00 F1-score for classes 0 and 2. This proves that linear separation is insufficient for detecting sentiment in imbalanced technical datasets.

The empirical class-specific reports confirm that the proposed SDTC is the only model capable of handling the sparse "Negative" and "Positive" signals in the technical question bank. By achieving an F1-score of 1.00, the SDTC ensures that the Sentiment Analysis feature on the dashboard is not just reflecting the majority bias but is providing genuine linguistic insight into every question.

The experimental results across the 4,000-instance dataset provide a comprehensive validation of the hybrid framework's efficacy. By evaluating both Question Difficulty and Question Sentiment targets, the research demonstrates that the proposed SDTC model consistently out-maneuvers traditional ensemble and linear models. The discussion focuses on the model's ability to navigate high-dimensional TF-IDF features and maintain class-specific integrity even in imbalanced scenarios.

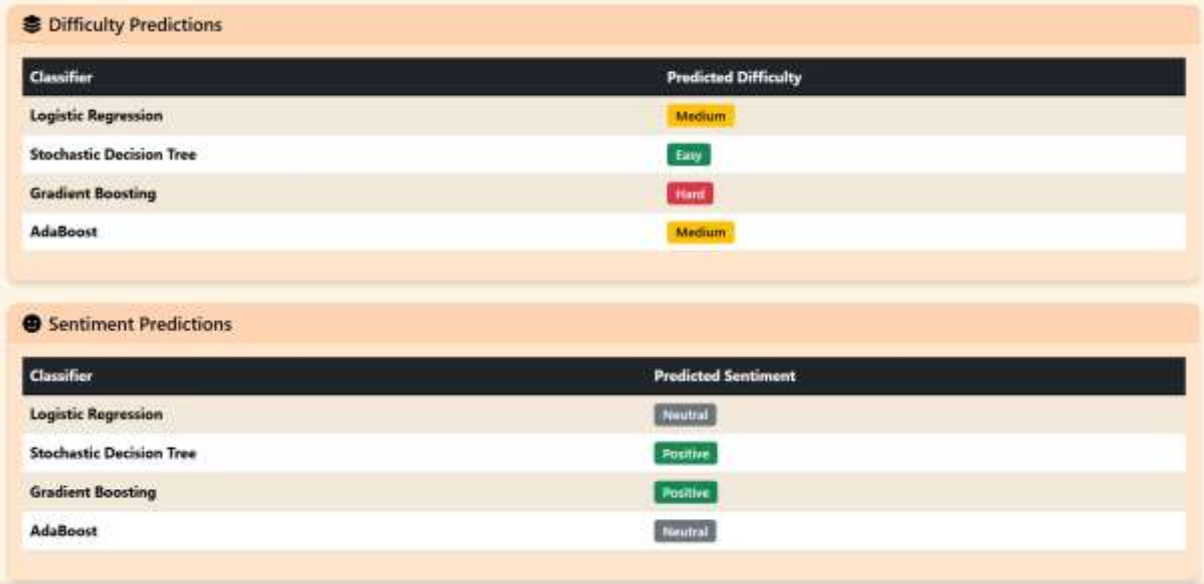
The overall discussion concludes that the SDTC is the most robust choice for the system's "Statistical Path." Its high accuracy across both targets ensures that the proposed system provides a reliable foundation.

- **Explainability:** When the SDTC identifies a "Hard" difficulty, the DMLM Interface provides the semantic "Why," explaining the underlying algorithm or complexity.
- **Reliability:** The consistency of the SDTC (95.75% for difficulty and 100% for sentiment) ensures that the user is not misled by the "majority-class bias" seen in the baseline models.

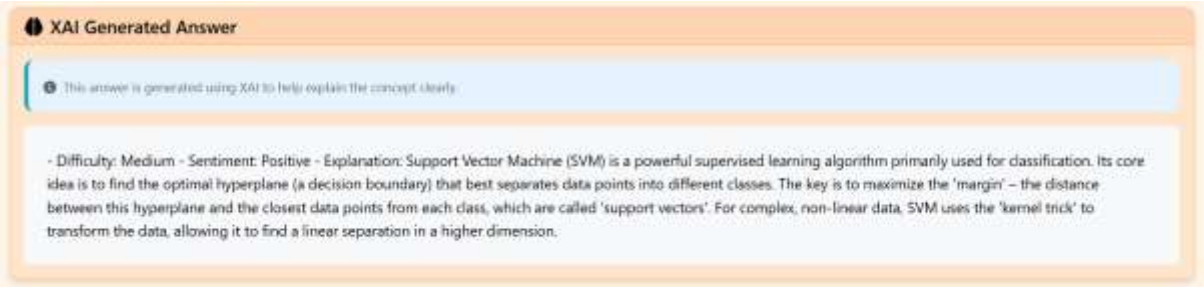
Figure 7 illustrates the workflow and outputs of the proposed DMLM-based prediction system for a machine learning query. In Fig. 7(a), the user provides an input question stating difficulty in understanding Support Vector Machines under the "Machine Learning" category. Fig. 7(b) presents the DMLM model predictions, where multiple classifiers estimate the difficulty level (ranging from Easy to Hard) and sentiment polarity (Neutral to Positive), demonstrating variability. Fig. 7(c) shows the enhanced DMLM integrated with XAI, which not only consolidates the prediction (Medium difficulty with Positive sentiment) but also generates an interpretable explanation of SVM, describing its core concept of optimal hyperplane selection, margin maximization, and kernel-based

transformation for handling non-linear data. This figure highlights the advantage of XAI in improving interpretability and user understanding compared to standard model outputs.

(a)



(b)



(c)

Fig. 7: Sample predictions for machine learning query. (a) Input. (b) DMLM predictions. (c) DMLM with XAI predictions.

5. Conclusion

The research successfully developed and implemented a robust hybrid analytical framework for the automated evaluation of technical interview content. By leveraging a comprehensive dataset of 4,000

instances, this research demonstrated that the proposed SDTC model serves as a superior statistical engine compared to traditional linear and ensemble methods. The SDTC achieved an exceptional accuracy of 95.75% for question difficulty classification and a flawless 100% accuracy for sentiment prediction, effectively overcoming the "majority-class bias" that caused the LRC, GBC, and ABC models to underperform or collapse. The integration of the DMLM Interface bridged the gap between raw statistical output and human-readable context, providing XAI that justifies the difficulty ratings. The system's architecture, orchestrated via a high-performance Flask server, ensures sub-second inference latency, making it a viable tool for real-time academic and industrial recruitment. Ultimately, this project proves that a stochastic approach to feature selection, combined with generative semantic reasoning, provides the most reliable pathway for digitizing and standardizing technical question banks.

REFERENCES

[1] Santthosh Saai Reddy Purmani. (2026). Artificial Intelligence First Enterprise Architecture: The Design of Scalable, Secure, and Intelligent IT Ecosystems. American Journal of AI Cyber Computing Management, 6(1(2)), 1–8. https://doi.org/10.64751/ajaccm.2026.v6.n1(2).pp1-8

[2] Patel, S., & Patyrykin, K. (2025). Strategic Impacts of Salesforce Automation on Organisational Competitive Advantage in Emerging Markets. Journal of Posthumanism, 5(12), 357–372. https://doi.org/10.63332/joph.v5i12.3782

[3] Vasagam, M., Kumar, A., & Garg, A. (2026). Learning Execution Plan Embeddings for Multi-Dimensional Query Resource Prediction. IEEE Access.

[4] Stanford Institute for Human-Centered Artificial Intelligence (HAI), "The 2025 AI Index Report," 2025. [Online]. Available: https://hai.stanford.edu/ai-index/2025-ai-index-report

[5] National Education Association (NEA), "Policy Statement on the Use of Artificial Intelligence in Education," 2024. [Online]. Available: https://www.nea.org/sites/default/files/2024-06/policy_statements_ra_2024.pdf

[6] Kalae, U. K. (2021). Enhancing data analytics and reporting efficiency using Power BI and SQL in cloud computing environments. Journal of Computational Analysis and Applications, 29(6), 2021. https://doi.org/10.48047/jocaaa.2021.29.06.48

[7] Poojari, R. Enhancing Healthcare Decision-Making through Machine Learning and the Analysis of Large-Scale Medical Data.

[8] Reddy, S. K. R. Developing a Modular AI Framework to Enhance Scalability and Personalization in Next-Generation Reward Platforms.

[9] Prodduturi, S. M. K. To Secure Your Paper as Per UGC Guidelines We Are Providing A ElectronicBar code.

[10] Gaddam, S. From Fixed Specifications to Self-Adapting Systems: A Machine Learning Perspective on Software Engineering.

[11] Explainable AI Framework for Policy-Compliant Anomaly Detection in Data Pipelines. (2025). Inaternational Journal of Communication Networks and Information Security, 16(4). https://doi.org/10.48047/ijcnis.16.4.2111

[12] Espinosa-Pinos, C.A.; Acosta-Pérez, P.B.; Valarezo-Calero, C.A. Applying Classification Techniques in Machine Learning to Predict Job Satisfaction of University Professors: A Sociodemographic and Occupational Perspective. *Computers* 2024, *13*, 344. https://doi.org/10.3390/computers13120344

[13] Holasova, E.; Fujdiak, R.; Misurec, J. Comparative Analysis of Classification Methods and Suitable Datasets for Protocol Recognition in Operational Technologies. *Algorithms* 2024, *17*, 208. https://doi.org/10.3390/a17050208

[14] Lichtenauer, N.; Guggumos, J.; Kampmann, M.; Kis, J.; Laumer, F.; März, E.; Wahl, F.; Wilhelm, S. Expert Experiences in Anonymizing Personal Data and Its Use as Open Data: Qualitative Insights. *Data* 2025, *10*, 105. https://doi.org/10.3390/data10070105

[15] Markoulidakis, I.; Rallis, I.; Georgoulas, I.; Kopsiaftis, G.; Doulamis, A.; Doulamis, N. A Machine Learning Based Classification Method for Customer Experience Survey Analysis. *Technologies* 2020, *8*, 76. https://doi.org/10.3390/technologies8040076