IOCES 2026 Conference
The 1st International Online Conference on Education Sciences
15–17 June 2026 | Online
MDPI

Supporting Didactic Evaluation of Mathematics and Science Concepts Through Automatic Short-Answer Grading

Miguel Ángel González Maestre, Javier Cubero Juánez, Alejandro de la Hoz Serrano and Lina Melo

Department of Experimental Science and Mathematics Teaching Area, University of Extremadura, 06006 Badajoz, Spain;
maestre@unex.es (M.Á.G.M.), jcubero@unex.es (J.C.J.), alexdlhoz@unex.es (A.d.l.H.S.), lvmelo@unex.es (L.V.M.N.)

INTRODUCTION & AIM

Short open-ended responses are widely used in STEM education to assess students' conceptual understanding. These responses provide richer evidence of learning than multiple-choice questions, but their manual evaluation can be time-consuming and difficult to scale.

Automatic Short-Answer Grading (ASAG) offers a promising solution by supporting teachers in the assessment process and has become an established research area within educational assessment [1]. However, educational datasets often present low-resource characteristics, including limited sample sizes, lexical sparsity, and class imbalance. These conditions make model evaluation challenging and may compromise the reliability and reproducibility of reported results.

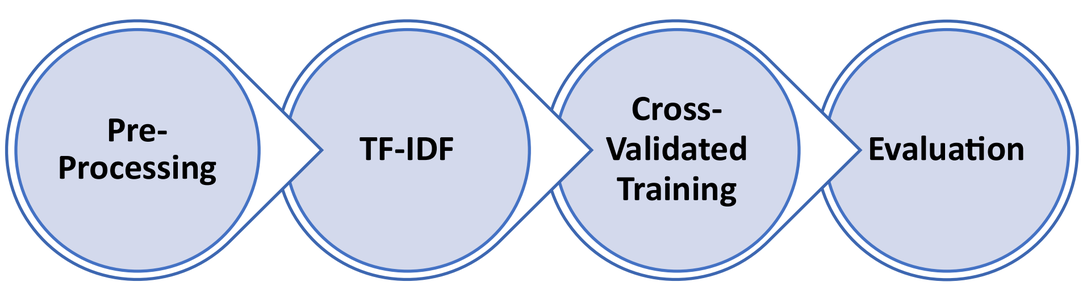
Furthermore, while recent advances in deep learning have improved performance in many natural language processing tasks, educational applications require not only accuracy but also transparency, interpretability, and methodological robustness. In classroom settings, understanding why a model produces a given prediction is often as important as the prediction itself, making interpretability a key requirement for educational AI systems [2].

The specific objectives were:

- To compare the performance of three interpretable linear classifiers (Logistic Regression, Multinomial Naïve Bayes, and Linear Support Vector Machines).
- To implement a stability-oriented evaluation framework based on adaptive stratified cross-validation.
- To analyse classifier behaviour through token-level lexical inspection.
- To support transparent and reproducible automated assessment in mathematics and science education.

METHOD

The proposed framework was evaluated using 2,940 short student responses distributed across ten concept-specific datasets from mathematics, science, and computing education. Each dataset contained manually labelled responses classified as correct or incorrect, reflecting authentic classroom assessment scenarios.



1. Preprocessing

Student responses were cleaned and standardized through basic text preprocessing procedures. This stage reduces lexical noise and prepares the responses for consistent feature extraction across educational concepts.

2. TF–IDF Representation

Responses were transformed into TF–IDF vectors, allowing each answer to be represented according to the importance of its lexical features. This representation captures discriminative vocabulary while maintaining model interpretability.

3. Cross-Validated Evaluation

Adaptive stratified cross-validation was applied to preserve class distributions and obtain reliable performance estimates under low-resource conditions. This strategy reduces evaluation bias and minimizes information leakage.

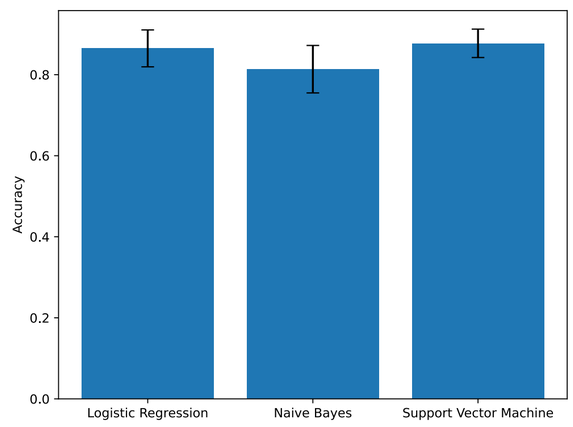
4. Training & Evaluation

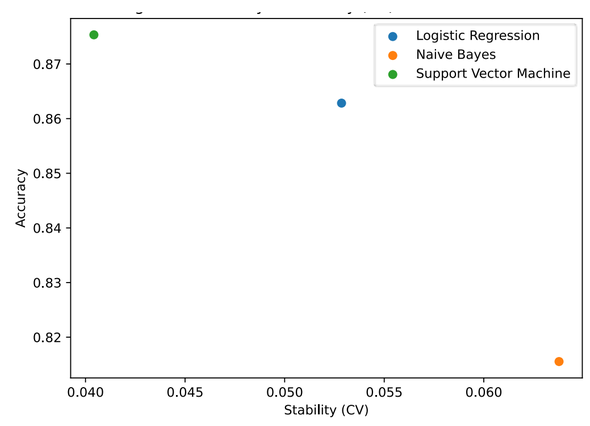
Three interpretable linear classifiers (Logistic Regression, Multinomial Naïve Bayes, and Linear SVM) were trained and evaluated using multiple performance metrics. Model behaviour was further analysed through stability-oriented evaluation procedures.

RESULTS & DISCUSSION

Global Performance

All three classifiers achieved competitive and consistent performance across the ten educational concept datasets. Although Linear SVM obtained the highest average accuracy, the differences among models remained relatively small.
The overlapping variability observed across cross-validation folds suggests that all classifiers provide reliable baseline performance under realistic classroom-scale conditions. These findings indicate that simple and interpretable linear models remain effective for low-resource Automatic Short-Answer Grading (ASAG) tasks.

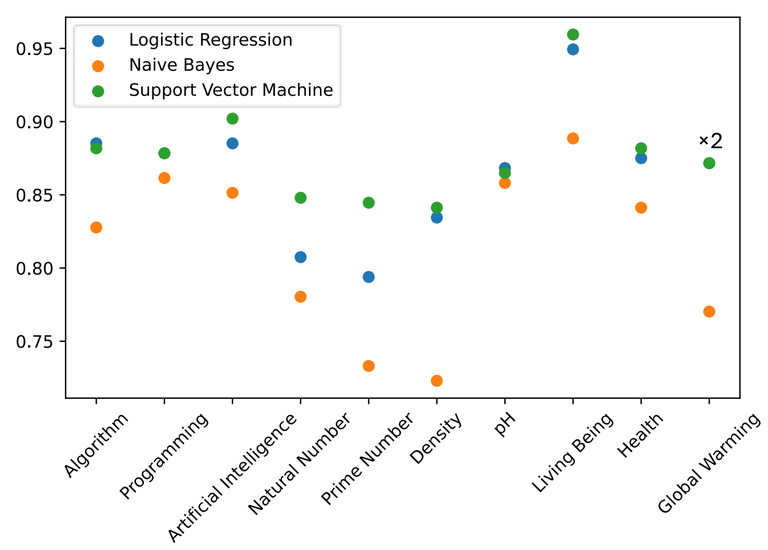




Cross-Concept Robustness

Performance varied moderately across educational concepts, reflecting differences in linguistic complexity and lexical variability among student responses. However, dispersion patterns remained broadly consistent across classifiers. This suggests that concept-specific dataset characteristics exert a stronger influence on performance than the choice of linear classification algorithm itself.
Consequently, careful dataset design and evaluation protocols may be more important than marginal differences between classifiers.

Performance remained relatively stable across the ten educational concepts. Although moderate differences appear across datasets, similar dispersion patterns were observed for all classifiers. Although moderate differences appear across datasets, similar dispersion patterns are observed for all classifiers.



This suggests that concept-specific linguistic characteristics have a greater impact on performance than the particular classification algorithm employed.

CONCLUSION

Conclusions

- The proposed ASAG pipeline provides a reproducible framework for evaluating short open-ended responses in low-resource educational settings.
- Linear classifiers achieved stable and comparable performance across ten STEM concept datasets.
- Stability-oriented evaluation complements traditional performance metrics and supports more reliable educational AI assessment.

REFERENCES

[1] Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. International Journal of Artificial Intelligence in Education, 25, 60–117.
[2] Molnar, C. (2020). Interpretable Machine Learning. Leanpub.