Explainable models for feedback design: An argumentative writing example

Antonette Shibani*
University of Technology Sydney, Sydney, Australia
antonette.shibani@uts.edu.au

Ratnavel Rajalakshmi†
Vellore Institute of Technology, Chennai, India
rajalakshmi.r@vit.ac.in

Srivarshan Selvaraj
Vellore Institute of Technology, Chennai, India
srivarshan.2019@vitstudent.ac.in

Faerie Mattins
Vellore Institute of Technology, Chennai, India
faeriemattins.r2019@vitstudent.ac.in

Dhivya Chinnappa
JPMorgan Chase and Co.
dhivya.infant@gmail.com

ABSTRACT

Recent works in educational data mining emphasize the need for producing practical insights that enhance learning. There is particular interest in supporting *student writing* by generating actionable feedback using machine learning algorithms. While algorithmic efficiency is generally sought after in machine learning, it might not be the most important aspect to consider for 'explainability'. This paper presents a predictive model for argumentative writing feedback where explanations supported by Local Interpretable Model-agnostic Explanations (LIME), SHapley Additive exPlanation (SHAP), and logic are derived to generate insights for designing student feedback on argumentative writing. It discusses the computational trade-offs and insights derived that inform writing feedback in practice, with lessons transferable to other contexts.

Keywords

explainable, feedback, predictive models, argumentation, writing, educational data mining, learning analytics, black box

1. INTRODUCTION

A common usage of data in education involves the development of *machine learning models* that can provide predictions, recommendations, and personalised support for learners, connecting fields such as Educational Data Mining (EDM), Artificial Intelligence and EDucation (AIED), and Learning Analytics (LA) [10]. Yet, the complex algorithms in these models create a 'black-box' effect, making the variables that contribute to the final prediction unclear (*Intrinsic opacity*) [2] [6]. This phenomenon is challenged by the emergence of Explainable Artificial Intelligence (XAI) as a field of research for models that offer interpretability and trustworthiness [3] [8].

The need for explainability becomes even more eminent when designing *feedback* for student-facing tools where impact on learning is at the forefront. Feedback-based LA systems generally include the provision of automated feedback to learners that closes the loop from the analytics generated [17] [5]. Automated tools can provide additional feedback to learners in a quick, consistent way at a scale that humans can't provide, although noting that students may engage with it in different ways based on their automated feedback literacy and critical engagement skills [12]. For actionable feedback to be provided by LA tools and to increase learner trust, the foundation lies in explainable LA that can help provide appropriate explanations for the decisions by machine learning models [4].

Argumentation is a critical skill for humans as they routinely engage with conflicting information and inconsistencies arising out of them [1]. Teaching argumentation is often integrated into writing curricula through the use of argumentative essays, with recent efforts in analyzing and providing automated feedback on these essays [15] [16]. While progress has been made in identifying and analyzing argumentation in data sets, for instance using argument mining [7], there is a need for more work on providing actionable feedback to learners to improve their argumentation skills. This can be expanded by the work in *writing analytics* that supports the provision of automated feedback to improve writing skills, where feedback to improve students' higher order competencies such as argumentation has been a recent focus [11].

In this study, we present an approach to designing an explainable machine learning model that supports the provision of feedback to learners in argumentative writing. We discuss the specific case of building a predictive model for argumentative writing quality and explain our approaches and findings examining what works best for explainability and feedback design. We demonstrate exemplary methods for developing explainable models for learner feedback and how it can impact educational practitioners who design this feedback and point out avenues for future work.

*Antonette Shibani
†Ratnavel Rajalakshmi

2. OUR APPROACH

Data for this study came from the Dagstuhl-15512 ArgQuality Corpus [14] - a standard annotated corpus commonly used for argumentation studies. The corpus contained 320 arguments manually coded for 15 dimensions of argumentation quality by three annotators with the overall score metrics: Low, Average, or High. The corpus consisted of 16 different issues (topics for arguments), with a for and against stance for each issue. The data set distribution across the different quality metrics is highly imbalanced, reflecting how this data occurs in the real world. Table 1 shows examples from the dataset.

Our approach to building the prediction model for argumentation quality is as follows. To start with, the arguments were pre-processed by filtering out the non-arguments, removing stop words and punctuation, and stemming the words. The ground truth was established by consolidating the annotations for argumentation quality (low, average, high) only considering rows where at least two annotators agreed on the quality. This process removed inconsistencies in the coding, reducing the number of arguments to 261. The four dimensions identified by authors of the data set as key quality indicators: overall quality, cogency, effectiveness, and reasonableness [14] were taken for modeling as the other sub-dimensions were too fine-grained for automated analysis. The data, vectorized using bag-of-words, was then used to build predictive models for argumentation quality, using two approaches discussed next.

In the baseline approach, the vectorized arguments were used to train Logistic Regression, Decision Tree, Random Forest classifiers, and a Neural Network to predict the overall quality. Hyperparameter tuning was performed using an exhaustive grid search on the Logistic Regression, Decision Tree and Random Forest models, and model parameters used are shown in Table 2. While this approach would likely work for evaluating the overall quality of arguments, it will only be able to provide minimal insight for generating feedback (which is often an end goal when opting for more explainable models).

In the proposed approach, we introduce a *two-stage model* to predict the overall quality of the argument along with the other underlying dimensions (cogency, effectiveness, and reasonableness) to enhance model explainability. Three classifiers (Models 1, 2, and 3) individually trained on the vectorized arguments to predict the three underlying dimensions constituted the first stage of the model. The four machine learning algorithms used in the previous approach were also employed in this context to find the best-performing classifier. The second stage of the model used a single classifier (Model 4) trained on a vector formed by augmenting the one-hot encoding of the underlying dimensions with the vectorized argument to add further context (Training stage 2). This classifier predicts the overall quality. For the final two-stage model, the argument vector was passed to the stage 1 classifiers, and the best-performing models were used for predicting each of the three dimensions (Table 4). These predicted dimensions were encoded and augmented to the original argument text vector, which was then fed to the stage 2 classifier to predict overall quality. The steps are shown in (Figure 1)

We use two existing tools to interpret the models in this study for explainability. The first, Local Interpretable Model-Agnostic Explanations (LIME) offers local explanations by explaining the classifier for a single instance [9]. We used LIME to extract explainable features from the Logistic Regression model predicting argumentation quality in our work. The second, SHapley Additive exPlanation (SHAP) uses Shapely values for finding values of the features that influence the model's scoring. SHAP was used to provide explanations for the Decision Tree model predicting argumentation quality.

3. FINDINGS AND DISCUSSION

A weighted average has been taken for precision, recall, and F_1 score to account for class imbalance to evaluate the results of the baseline model (Table 3). The Decision Tree model, though not the best-performing model, is rule-based and can easily provide explanations for the decisions it makes, hence demonstrated in this study for better explainability. The Bag-of-Words representation was chosen as it provides information on the occurrence of words in the argument and can provide insight into the overall quality of the argument, thus enhancing the explainability of the system. In this model, the decision taken in each node is based on the presence or absence of a particular token in the argument. Using the nodes of the tree one can arrive at a rule-based system to provide feedback to the learner. For instance, a node in the decision tree can indicate as follows: if the argument contains any word containing the token "discov" (discover, discovery, etc.) or "found", then the argument is most likely to be of higher quality. An explanation for these rules might be that the arguments based on discoveries and findings of others are higher quality because they include validated claims. This feedback can then be used to suggest adding evidence or links to supporting research to strengthen the argument made.

The proposed 2-stage model improves the explainability of results using the additional underlying dimensions. The chosen classifiers for each model and their results are displayed in Table 4. Some classifiers like the Logistic Regression classifier were chosen to predict the overall quality as it offers better model explainability. This trade-off for explainability where an easier-to-interpret model is used even if it yielded lower scores than the black box model is a way to tackle the intrinsic opacity in algorithmic decision making [2] [6]. The final two-stage model, after integrating stages one and two, achieves a weighted F_1 score of 0.59. Further exploration of model results can identify insights into words and dimensions that indicate better quality argumentation for improved feedback. This was explored using logistic regression results from the 2-stage model.

The logistic regression model's feature coefficients can reveal the impact of individual words on predicting argument quality. Table 5 shows sample words and their coefficients with the three coefficients corresponding to the three levels of qualities. The word 'found' had the highest coefficient, correlating with average overall quality, suggesting its presence impacted the argument's average quality coding. An argument example with 'found' coded as average is in Table 1, and similar impactful words can be studied for providing feedback. Since the model was trained on the augmented

Table 1: Examples from the dataset with selected rows and columns

id	argument	issue	stance	overall quality
arg219206	Americans spend billions on bottled water every year. Banning their sale would greatly hurt an already struggling economy...	ban-plastic-water-bottles	no-bad-for-the-economy	3 (High)
arg219259	Bottled water is somewhat less likely to be found in developing countries, where public water is least safe to drink...	ban-plastic-water-bottles	no-bad-for-the-economy	2 (Average)
arg219213	Estimates variously place worldwide bottled water sales at between $50 and $100 billion each year, with the market expanding at the startling annual rate of 7 percent...	ban-plastic-water-bottles	yes-emergencies-only	1 (Low)

Table 2: Hyperparameter tuning for baseline models

Model	Parameters
Logistic Regression	'C':1.0, 'dual': False, 'fit_intercept':True, 'penalty':none, 'solver':'sag', 'max_iter':5000
Decision Tree	'criterion': 'gini', 'max_features': 'log2', 'splitter': 'best'
Random Forest	'bootstrap':True, 'class_weight':'balances', 'criterion':gini, 'max_features':none,'n_estimators':300, 'oob_score':False, 'warm_start':False

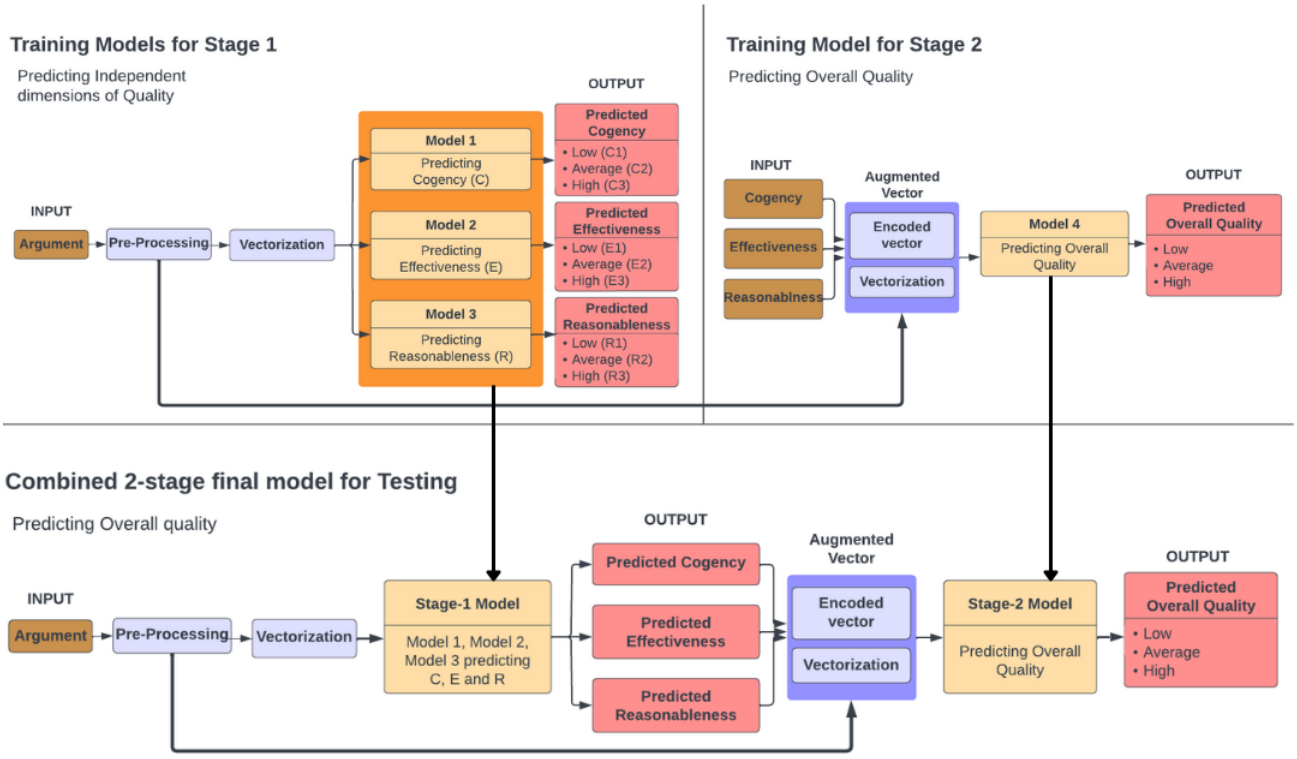

Figure 1: Proposed 2-stage model to predict the overall quality of an argument

vector containing the three underlying dimensions, the same coefficient method can be extended to examine the dimensions as well. From Table 6, we see that if the argument had average effectiveness, then the overall quality of the argument is more likely to be average. Similarly, Reasonableness has the highest positive and negative coefficients, implying its greater impact on overall quality than other dimensions. Thus the feedback provided can be to improve the reasonableness of arguments by explaining the reason behind a stance by using words like "reason", "explain", and "because" (derived from the arguments with high reasonableness). Table 6, also displays that Low Cogency contributes the most to Low Overall Quality. Feedback can thus suggest avoiding uncertain language (Words like 'would' and 'think';

Table 3: Performance of the different classifiers in the baseline model for predicting overall quality

Classifier	Accuracy	Precision	Recall	F_1 Score
Logistic Regression	0.62	0.57	0.62	0.59
Decision Tree	0.59	0.58	0.59	0.58
Random Forest	0.62	0.56	0.62	0.58
Neural Network	0.61	0.60	0.61	0.60

Table 4: Performance of the chosen intermediate classifiers

Predicted Dimension	Best Model	Metrics			
		F_1 score	Precision	Recall	Accuracy
Cogency	Neural Network	0.56	0.55	0.58	0.58
Effectiveness	Neural Network	0.56	0.54	0.59	0.59
Reasonableness	Neural Network	0.56	0.55	0.58	0.58
Overall Quality	Logistic Regression	0.87	0.87	0.87	0.87

Table 5: Feature coefficients for the word tokens in logistic regression in the 2-stage model

Word	Coef 1 (Low)	Coef 2 (Average)	Coef 3 (High)
discov	-0.024	0.035	-0.011
found	-0.015	0.019	-0.004
although	-0.036	-0.077	0.044

Table 6: Feature coefficients for the underlying dimensions in logistic regression in the 2-stage model.

Dimension	Low	Average	High
Low Cogency	0.335	-0.162	-0.173
Average Cogency	-0.146	0.434	-0.288
High Cogency	-0.189	-0.272	0.461
Low Effectiveness	0.199	-0.090	-0.109
Average Effectiveness	0.045	0.329	-0.374
High Effectiveness	-0.245	-0.239	0.484
Low Reasonableness	0.268	-0.135	-0.132
Average Reasonableness	-0.126	0.513	-0.386
High Reasonableness	-0.141	-0.378	0.519

derived from low cogency arguments) for higher-quality argumentative writing.

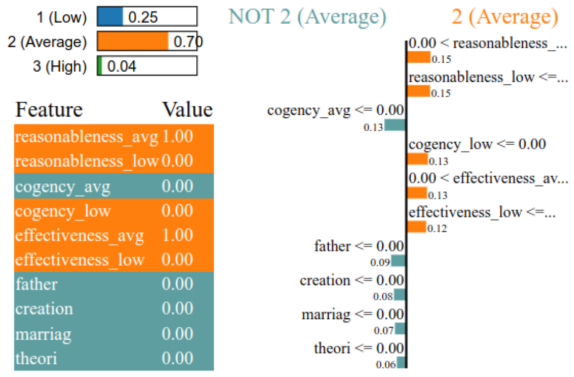

Figure 2: A sample testing instance using LIME for Logistics Regression classifier

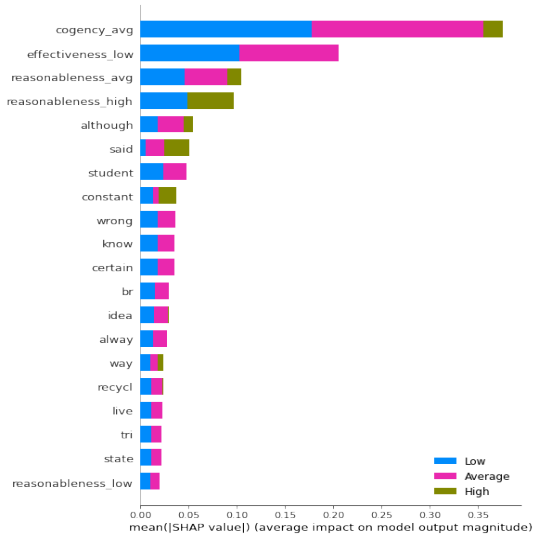

Figure 3: SHAP summary plot for the Decision Tree classifier

Figure 2 demonstrates how LIME can be used to derive explanations for a sample instance, using the 2-stage model to predict the overall quality. The figures display the features and their weights as a table (left) and a bar chart (right), in decreasing order of relevance. The feature 'reasonableness_avg' having a weight of 0.15, is the most significant attribute that supports the instance's average overall quality. The absence of topic-related words (as per the argument's context) such as "father", "creation", "marriag" and "theori" (weights are 0) suggest NOT average overall quality - the presence of such relevant words might indicate higher quality arguments instead. A useful feedback can then be to include more in-depth content related to the topic for higher argumentation quality.

SHAP's summary plot (Figure 3) illustrates the features and their shapely values which attribute more to each target class. The main feature contributing to the prediction of overall argument quality as average is cogency_avg. Similarly, the word 'said' supports the overall quality to be high or average. The word 'idea' contributes to the overall qual-

ity being majorly average, possibly pointing to a plan, suggestion, course of action, opinion, or belief, which enhances the argument's overall quality. These frameworks and explanations when evaluated and incorporated into a tool can help generate automated feedback on writing for improving argumentation.

4. CONCLUSION

Our study demonstrates using explainable predictive models for designing feedback for learners. We used a 2-stage model to predict argumentation quality in writing, considering sub-dimensions of quality along with the argument text to enhance explainability. We demonstrated different methods to tackle the intrinsic opacity of algorithms such as the selection of easier-to-interpret models, tailoring the models for particular purposes, choosing features that contribute to better feedback, and decoding model results at different stages to provide actionable feedback. The contribution is hence in presenting an example of a generalisable approach to develop explainable models for feedback. Our methods for using explainable models to inform feedback design apply to various contexts with algorithmic decision-making. These approaches can improve the design of machine learning-based feedback tools that provide learners with interpretable and actionable feedback.

The study is a proof of concept for building explainable models to generate feedback using a small size argumentative writing data set and demonstrated feedback design for the specific context. Future work can build on this work by expanding to larger data sets and examining finer-grained details in the models to provide actionable feedback. While the analysis of the corpus provided insights into argumentation, getting input from educators and co-designing with them is required for a more deliberate design of feedback. This can help validate findings from the model to translate to feedback for classroom practice [13].

5. REFERENCES

[1] P. Besnard and A. Hunter. *Elements of argumentation*, volume 47. MIT press Cambridge, 2008.

[2] J. Burrell. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big data & society*, 3(1):1–12, 2016.

[3] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang. Xai—explainable artificial intelligence. *Science Robotics*, 4(37):eaay7120, 2019.

[4] H. Khosravi, S. B. Shum, G. Chen, C. Conati, Y.-S. Tsai, J. Kay, S. Knight, R. Martinez-Maldonado, S. Sadiq, and D. Gašević. Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 3:100074, 2022.

[5] K. Kitto, M. Lupton, K. Davis, and Z. Waters. Designing for student-facing learning analytics. *Australasian Journal of Educational Technology*, 33(5):152–168, 2017.

[6] B. Lepri, N. Oliver, E. Letouzé, A. Pentland, and P. Vinck. Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology*, 31(4):611–627, 2018.

[7] A. Lytos, T. Lagkas, P. Sarigiannidis, and K. Bontcheva. The evolution of argumentation mining: From models to social media and emerging tools. *Information Processing & Management*, 56(6):102055, 2019.

[8] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[10] B. Rienties, H. Køhler Simonsen, and C. Herodotou. Defining the boundaries between artificial intelligence in education, computer-supported collaborative learning, educational data mining, and learning analytics: A need for coherence. *Frontiers in Education*, 5:128, 2020.

[11] A. Shibani, A. Gibson, S. Knight, P. H. Winne, and D. Litman. Writing analytics for higher-order thinking skills. *Companion Proceedings of the 12th*, page 165, 2022.

[12] A. Shibani, S. Knight, and S. Buckingham Shum. Questioning learning analytics? cultivating critical engagement as student automated feedback literacy. In *LAK22: 12th International Learning Analytics and Knowledge Conference*, pages 326–335, 2022.

[13] A. Shibani, S. Knight, and S. B. Shum. Educator perspectives on learning analytics in classroom practice. *The Internet and Higher Education*, 46:100730, 2020.

[14] H. Wachsmuth, N. Naderi, Y. Hou, Y. Bilu, V. Prabhakaran, T. A. Thijm, G. Hirst, and B. Stein. Dagstuhl-15512-argquality, Apr. 2017.

[15] X. Wang, Y. Lee, and J. Park. Automated evaluation for student argumentative writing: A survey. *arXiv preprint arXiv:2205.04083*, 2022.

[16] W. Xing, H.-S. Lee, and A. Shibani. Identifying patterns in students' scientific argumentation: content analysis through text mining using latent dirichlet allocation. *Educational Technology Research and Development*, 68(5):2185–2214, 2020.

[17] R. Zhi, S. Marwan, Y. Dong, N. Lytle, T. W. Price, and T. Barnes. Toward data-driven example feedback for novice programming. *International Educational Data Mining Society*, 2019.