Automatic Explanation Quality Assessment in Online Learning Environments: new datasets and methods

Sameer Bhatnagar[1], Amal Zouaq[1], Michel C. Desmarais[1], and Elizabeth Charles[2]

[1] Ecole Polytechnique Montreal
{sameer.bhatnagar,amal.zouaq,michel.desmarais}@polymtl.ca
[2] Dawson College echarles@dawsoncollege.qc.ca

Abstract. *Asynchronous Peer Instruction* is increasingly popular in online learning environments. It relies on the principle of leveraging content that is both *generated* and *evaluated* by novices to foster learning and self-reflection. Students first respond to a question item, but they must also provide an explanation for their reasoning. They are then presented alternative explanations as written by their peers, and given the opportunity to change their initial answer choice, based on those they find most convincing. The peer-explanations that students find most convincing represent valuable data, for teachers to better grasp their students' understanding, and for the learning environment itself, as higher quality explanations can be shown to students as examples to compare their work to. This study reports on the application of text mining methods in the context of asynchronous peer instruction, with the objective of automatically identifying high quality student explanations. Our results offer the potential to inform the design of "learnersourcing" systems, such as asynchronous peer instruction. These design choices are critical as these systems scale, especially with respect to providing pedagogically insightful reports to teachers, and presenting engaging alternative explanations to students to promote higher order thinking.

Keywords: Argument mining · Learnersourcing · Peer Instruction

Must be anonymized before submission

1 Introduction

Learning environments that leverage peer submitted content, carry the great advantage of scaling up to a larger content base. Be it student-generated questions, sample solutions, feedback, or explanations and hints, peer-submitted content can easily help grow a database to meet the demand for practice items and formative assessment. Platforms built on this principle are growing in popularity [7][14], freeing the teacher from the tedious task of developing multiple variants of the same items.

But critical to the success of such environments is the capacity to automatically assess the quality of learner-generated content. There are a growing number of tools which put students at the centre of this challenge of quality assessment [22][3], wherein after students submit their own work, after which they are prompted to evaluate a subset of their peers' submissions, and sometimes even provide feedback.

To counter the drawback that lies in the varying ability of novices to evaluate and provide good feedback to their peers, these environments often use *pairwise* comparison. It is widely accepted that evaluative judgements based on ranking two objects relative to one another, are easier to make than providing an absolute score. Adaptive Comparative Judgement [21], where teachers assess student submissions by simply choosing which is better from a pair, has been shown to be a reliable and valid alternative to absolute grading. As such, there is a growing family of learning tools which, at the evaluation step, present the peer-generated content as *pairs* to the current learner, and prompt for a pairwise ranking. This data can provide feedabck to the students who originally submitted the item, but can also be used for moderating database content.

In platforms where students generate content that is part of the learning activities of future students, filtering out irrelevant and misleading material is paramount. However, while removing bad content is important, educators hope to identify, and subsequently maximize the use of, the best student generated items. Since not all students can be asked to evaluate all possible pairs, this in turn leads to the challenge of optimizing which items need evaluation by the "student-come-moderators", without hindering their learning.

This is an example of the classic trade-off of exploration-vs-exploitation from the field reinforcement learning: how do we balance

- exploiting the student submissions for which we have reliable data, and can estimate their quality, so that future students are assured of learning from this valuable content,
- while concurrently exploring the potential positive learning impact of student contributions that are newer to the database, and need to be shown and evaluated in order to get an estimate of their quality? [27]

This challenge is a hallmark of learning environments that leverage *learnersourcing*[26].

The focus of this study is on the subset of these learning environments that enable *peer instruction*[6], and are built on a specific two-stage script:

1. students are prompted to answer a multiple choice question, and provide a free-text explanation that justifies their answer;
2. without revealing the correct answer, students are then prompted to reconsider their answer, by presenting them a selection of explanations written by previous students [1]. Students can either decide that their own explanation is best, or indicate which of their peers' explanation was the most convincing.

In this instance of *learnersourcing*, as in the tools above, this "vote data" is valuable at two levels: it can then be used to determine what to present to future students, but also inform instructors of their students' understanding of the material.

We frame the explanations students produce here as *arguments* meant to persuade one's peers that their own reasoning is correct. The ultimate objective of this study is to automatically assess the quality of these arguments, as they may help future students reflect on their own reasoning surrounding different concepts. We borrow from work in the argument mining community, where argument quality can be measured along the dimension of *convincingness*, as this gives us an operational definition: in a peer-instruction setting, can we predict which peer-explanations will be selected as more convincing than the student's own work?

To our knowledge, this is the first analysis of peer instruction data through the lens of *learnersourcing* and argument *convincingness*. Thus, as a reference, we include datasets and methods from the *argument mining* research community specifically focused on pairwise preference ranking of arguments for automatic assessment of convincingness. We apply vector space models, as well as state-of-the-art neural approaches, and report on their performance on this task.

Our findings suggest that the arguments generated in learning environments centred on undergraduate science topics present a more challenging variant of the task originally proposed in the argument mining community, and that classical approaches can match neural models for performance on this task across datasets, depending on the context.

2   Related Work

2.1   Learnersourcing & Comparative Peer Assessment

The term "learnersourcing" was coined by [26], and defined as the process in "which learners collectively generate useful content for future learners while engaging in a meaningful learning experience themselves." One of the earliest examples of a learning environment centred on this mode is Peerwise [8], wherein students generate question items for the subjects they are learning, share them with their peers, so that others can use them for practice. Ripple [14] is a similarly built system, but they add a recommendation engine which adaptively selects which problems to suggest to which students.

Other tools leave the creation of question items to teachers, but call on students to generate and evaluate *explanations* for the answers. The AXIS [27] system prompts students to provide an generate explanation for their response to a short answer question, and then evaluate a similar explanation from one of their peers on a scale of 1-10 for *helpfulness*. This rating-data drives the reinforcement learning algorithm that then decides which explanations to show to which future students.

A similar class of learning platforms leverage comparative judgement for supporting how students can evaluate the work of their peers. Juxtapeer [3]

asks students to provide feedback to a single peer on their work by explicitly comparing to that of another. ComPAIR [22] asks students for feedback on each item of the pair they are presented with, with a focus on what makes one "better" than the other.

Finally, *peer instruction* question prompts are being used more frequently inside online learning assignments UBCPI [2] [4]. The student is presented with a multiple choice item, and prompted to provide an explanation of their answer choice. In order to encourage students to reflect on their own reasoning, they are then presented with a subset of alternative peer explanations to their own answer choice, as well for another choice. The student is then given the chance to revise their answer choice (or not), by indicating which of the peer explanations was most *convincing* from the subset.

2.2   Argument Quality & Convincingness

Conventional argument-mining pipelines include several successive components, starting with the automatic detection of argumentative units, classification of these units into types (e.g. major claim, minor claim, premise), and identification of argumentative relations (which evidence units support which claim). Such pipelines are essential in question-answering systems [15] and are at the heart if the IBM Project Debater initiative.

Work in the area of automatic evaluation of argument quality finds its roots in detecting evidence in legal texts [17], but has accelerated in recent years as more datasets become available in everyday contexts, and focus shifts to modelling more qualitative measures, such as *convincingness*.

Some of the earlier efforts included work on automatically scoring of persuasive essays [20] and modelling persuasiveness in online debate forums [24]. However, evaluating argument *convincingness* with an absolute score can be challenging, which has led to significant work in adopting a pairwise approach, where data consists of pairwise observations of two arguments, labelled with which of the two is most convincing.

In [11], the authors propose a feature-rich support vector machine, as well as an end-to-end neural approach based on pre-trained Glove vectors and a bidirectional Long-Short-Term Memory network for the pairwise classification task. This is extended in [10], where the authors build a Siamese network architecture, where each leg is a BiLSTM, taking as input the pair of explanations as Glove embeddings [19], in order to detect which of argument in a pair has the most convincing evidence. Finally, based on the success of transformer models such as BERT [9], the authors of [25] release a dataset of argument pairs and show that these models accurately predict the most convincing argument in a pair.

3   Data

One of the objectives of this study is to compare and contrast how text mining methods for evaluating argument quality, specifically for argument *convincing-*

ness, perform in an online learning environment with learner-generated and annotated arguments. Our primary dataset comes from a *peer instruction* learning environment, myDALITE.org. To provide context as to the performance that can be expected for this relatively novel task, we include in our study three publicly available datasets, each specifically curated for the task of automatic assessment of argument quality along the dimension *convincingness*. Table 1 provides examples of an example of an argument paoir from each of the datasets.

(a) A pair of arguments from the UKP dataset, both with the stance CON, for the prompt topic: "william-farquhar-ought-to-be-honoured-as-the-rightful-founder-of-singapore-yes-of-course-". Argument a1 is labelled as more convincing.

a1	a2
Farquar protected Singapore from Dutch attacks and attracted traders to Singapore while Raffles was away.	When Farquar was fired by Raffles, he was given a grander ceremony before he left compared to Raffles, which show how much the people loved him.

(b) A pair of arguments from the IBM dataset, both with the stance CON, for the prompt topic: "Online-shopping-brings-more-good-than-harm-(CON)". Argument a1 is labelled as more convincing.

a1	a2
it enables cheaper products by not needing to have physical store locations (which have rent), but only a warehouse.	it enables redistribution of wealth through rich people buying from poorer areas instead of their local, richer, stores and manufacturers.

(c) Example of argument pair constructed from dalite dataset

rationale	chosen_rationale
The graph shows constant positive acceleration and then constant negative acceleration. This means that the velocity-time graph should have a positive slope and then a negative slope, and graph C is the only option that satisfies those requirements.	1st, acceleration is +, indicating an increase in velocity.\r\nthen, acceleration is suddenly and without warning negative, and velocity is reduced.

Table 1: Examples of argument pairs from each dataset

3.1   UKP & IBM

UKPConvArgStrict[11], hence forth referred to as **UKP**, was the first to propose the task of pairwise preference learning for argument convincingness. The dataset consists of just over 1k individual arguments, that support a particular stance for one of 16 topics. These arguments were distributed as 11.6k pairs to annotators on a crowd-sourcing platform, where the task was to choose which of the two arguments, for the same stance regarding the same topic, was more convincing.

More recently, a second similar dataset, **IBMArgQ-9.1kPairs**[25], henceforth referred to as **IBM_ArgQ**, which is made of 3.4k individual arguments for 11 topics, assembled into 9.1k pairs labelled for which is more convincing. One of the key differences between these two is that **IBM_ArgQ** data is more strongly curated with respect to the relative length of the arguments in each pair: in order to control for the possibility that annotators may make their choice of which argument in the pair is more *convincing* based merely on the length of the text, the mean difference in word count, $\overline{\Delta wc}$, is just 3 words across the entire dataset, which is 10 times more homogeneous than pairs in **UKP**.

Finally, we include a third reference dataset, **IBM_Evi**, consisting of 1.5k individual arguments, organized into 5.2k pairs annotated for pairwise preference for convincingness [10]. The important distinction here is that the arguments are actually extracted as evidence for their respective topic from Wikipedia, and hence represent cleaner well-formed text than our other reference datasets.

Table 2 summarizes some of the descriptive statistics that can be used to compare these sources, and potentially explain some of our experimental results.

Table 2: Descriptive statistics for each dataset of argument pairs, with last rows showing **dalite** data split by discipline. N_{args} is the number of individual arguments, distributed across N_{pairs} revolving around N_{topics}. $\overline{wc}$ is the average number of words per argument, shown with the standard deviation (SD). $\overline{\Delta wc}$ is the average relative difference in number of words for each argument in each pair, shown with the standard deviation .

dataset	N_{pairs}	N_{topics}	N_{args}	N_{vocab}	$\overline{wc}$ (SD)	$\overline{\Delta wc}$ (SD)
IBM_ArgQ	9125	11	3474	6710	23 (7)	3 (2)
UKP	11650	16	1052	5170	49 (28)	30 (23)
dalite	8551	102	8942	5571	17 (15)	12 (7)
IBM_Evi	5274	41	1513	6755	29 (11)	3 (2)
dalite:Biology	3919	49	4116	3170	15 (14)	10 (6)
dalite:Chemistry	1666	24	1758	2062	20 (14)	12 (7)
dalite:Physics	2966	29	3068	2478	19 (15)	15 (7)

3.2 myDALITE

A data point from a peer instruction item, will have the following fields: question prompt, answer choices, student's first answer choice, student's explanation for their first answer choice, the peer explanations they are shown on the review step, their second answer choice, and the explanation they found most convincing. Quite often, students decide to keep the same answer choice, and indicate that their own explanation is te most convincing. We construct our **dalite** dataset by keeping only the observations where students chose a peer's explanation as more convincing on the review step. Each argument pair is thus made up of the one written by the current learner, while the other is an alternative which they either chose as more convincing. In this respect, our peer-instruction dataset is different than the one mentioned above, since its is the same person, the student, who wrote the explanation, and as well as evaluated the its relative convincingness in the pair.

To ensure internal reliability, we only keep arguments that were chosen by at least 5 different students. To ensure that the explanations in each pair are of comparable length, we keep only those with word counts that are within 25 words of each other.

This leaves us a dataset with 8551 observations, spanning 2216 learner annotators having completed, on average, 4.0 items each, from a total of 109 items across three disciplines, with at least 50 explanation-pairs per item.

Table 3: Observations of students choosing a peer explanation as more convincing than their own, or not, aggregated by discipline and whether they started and finished with the correct answer.

	rr	rw	wr	ww
Biology	2459	124	733	603
Chemistry	1151	51	228	236
Physics	2288	66	278	334

Table 3 highlights two key difference between the modelling task of this study, and related work in argument mining, where learner-annotators are presented pairs of arguments that are always for the same stance, in order to limit bias due to their opinion on the motion when evaluating which argument is more convincing (except in some observations of **IBM_Evi**). First, in a *Peer Instruction* learning environment, other pairings are possible and pedagogically relevant. In this dataset, the majority of students keep the same answer choice between the two steps of the prompt, and so they are comparing two explanations that are either both correct (*"rr"*) or incorrect (*"wr"*). However, 17 % of the observations in this dataset are for students who not only choose an explanation more convincing than their own, but also switch answer choice, either from the incorrect to correct, or the reverse. These pairs add a different level of complexity to the

model, but are very pertinent in the pedagogical context: what are the argumentative features which can help students remediate an initial wrong answer choice (*"wr"*)? What are the features that might be responsible for getting students to actually move away from the correct answer choice (*"rw"*)? Second, a more fundamental difference is that our study focuses on undergraduate science courses across three disciplines, wherein the notion of answer "correctness" is important. There are a growing number of ethics and humanities instructors using the peer instruction platform, where the question prompts are surrounding topics more like a debate, as in our reference datasets. We leave this comparison for future work.

4   Methods

Choosing which argument is more convincing from a pair, is a binary ordinal regression task, where the objective is to learn a function that, given two feature vectors, can assign the better argument a rank of $+1$, and the other a rank of -1. It has been proven that such a binary ordinal regression problem, can be cast into an equivalent binary classification problem, wherein the model is trained on the *difference* of the feature vectors of each argument in the pair [12]. Referred to as *SVM-rank*, this method of learning pairwise preferences has been used extensively in the context of information retrieval (e.g. ranking search results for a query based on past clickthrough data) [13], but also more recently in evaluating the journalistic quality of newspaper and magazine articles [16], and for predicting which argument is more convincing [11].

4.1   Vector Space Models

We follow-up on this work, building simple "bag-of-words" vector space models to represent our argument text. We take all of the individual arguments for a particular topic in our training set (known as "explanations" for a particular question item in the case of the **dalite** data), lemmatize the tokens, and build term-document matrices for each topic. We then take the arithmetic difference of these normalized term frequency vector representations of each argument to train Support Vector Machine classifiers on. We refer to this model as **ArgBow**. We do not, for this study, include any information related to the topic prompt in our document representation

4.2   Pre-trained word embeddings

A limitation of vectors space models is the exclusion of words that are "out-of-vocabulary" between the test set and the training data. This has led to experiments for this task[11][10] with language models that offer pre-trained word-embeddings that have already learned from massive corpora of text. In our **ArgGlove** model, we encode each token of each argument using 300-dimensional GloVe vectors [19], and represent each argument as the average of its token

embedding vectors. We then feed these into the same SVM-rank architecture described above.

4.3   Transfer Learning

Finally, in order to leverage recent advances in transfer-learning for NLP, the final model we explore is **ArgBert**. We begin with a pre-trained language model for English built using Bi-directional Encoder Representation from Transformers, known as *BERT*[9], trained on large bodies of text for the task of masked token prediction and sentence-pair inference. As proposed in [25], we take the final 768-dimensional hidden state of the base-uncased BERT model, feed it into a binary classification layer, and fine-tune all of the pre-trained weights using our argument-pair data. As in other applications involving sentence pairs for BERT, each argument pair is encoded as `[CLS] A [SEP] B]`, where the special `[SEP]` token instructs the model as to which token belong to argument `A` and which to argument `B`. [3].

4.4   Cross-Validation & State of the Art

The objective of this study is to apply methodology from the argument mining research community to inform the design of environment built upon learnersourcing of student explanations, so as to be able to be present student explanations that are *convincing* to students as they engage with their reasoning when they answer conceptual questions. As a baseline, we build a baseline model, **ArgLength**, which is trained on simply the number of words in each argument, as there may be many contexts where students will simply choose the longer/shorter argument, based on the prompt.

In order to get a reliable estimate of performance, we employ stratified 5-fold cross-validation for our experiments. This means that for each fold, we train our model on 80% of the available data, ensuring that each topic and output class is represented equally. In the context of peer instruction learning environments, this is meant to give a reasonable estimate of how our models would perform in predicting which explanations will be chosen as most convincing, *after* a certain number of responses have been collected.

In Table 4, we denote, to the best of our knowledge, the state-of-the-art performance for each dataset on the task of pairwise classification for *convincingness*. These values are meant to provide a reference as we present our results in the next section, but cannot be directly compared, as we employ a stratified 5-fold cross-validation scheme, while the reference studies measure cross-topic validation.

---

[3] modified from the `run_glue.py` script provided by the `tranformers` package, built by company hugging face. All code for this study provided in associated github repository

Dataset	Acc	AUC	model
UKP	0.83	0.89	ArgBERT[25]
IBM_ArgQ	0.80	0.86	ArgBERT[25]
IBM_Evi	0.73	-	EviConvNet[10]

Table 4: State of the art performance for pairwise argument classification of convincingness for three publicly available datasets, using cross-topic validation scheme

5   Results

- ArgLength sets baseline for each dataset. IBM datasets curate most aggressively for pairs with comparable lengths.
- Vector space models performance comparable to BERT.
- GloVe most stable across datasets. UKP and IBMEvi performance on ArgBoW due to wide-ranging vocabulary (lots of unknown tokens in test set), which reveals limitation of approach

- performance of in chemistry worst, physics best, given their relative proportions in data
- proportionately more errors for observations switch their answer (check if true in IBM Evi, where stances can be different)

6   Discussion

In [16], for task of pairwise ranking of newspaper articles based on "quality", the authors achieve a similar result: when comparing the performance of SVM-rank models using different input feature sets (e.g. *use of visual language*, *use of named entities*, *affective content*), their top performing models achieve "same-topic" pairwise ranking accuracy of 0.84 using a combination of content and writing features, but also a 0.82 accuracy with the content words as features alone.

7   Future Work

[18] and [16] use a combination of writing and content (BoW) features to achieve their best results, and thus this avenue must be explored more thoroughly, especially as this maybe vary across disciplines an teaching contexts. The state-of-the-art performance for pairwise classification accuracy on the **IBM_Evi** dataset employs Siamese architechture[10], where each leg of the network is a Bi-directional Long-Short-Term-Memory network, with documents represented by word2vec embeddings. These variations are left to explore in future work.

    In this study we do not ever infer which are, overall, the most convincing student explanations for any given item. Inferring a gold standard of global

(a) Accuracy

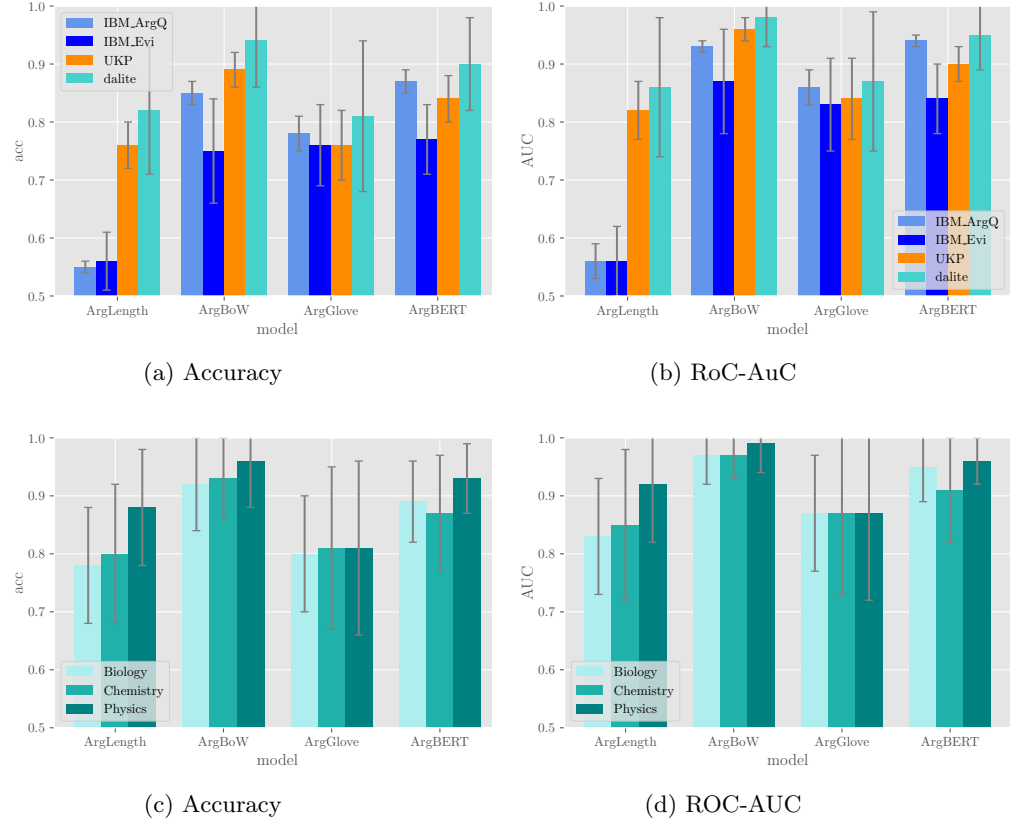
(b) RoC-AuC

(c) Accuracy

(d) ROC-AUC

Fig. 1: Pairwise ranking classification accuracy and ROC-AUC for different models in across datasets in figures 1a and 1b. Figures 1c and 1d split the performance for the **dalite** dataset, across disciplines. All results evaluated using 5-fold stratified cross-validation
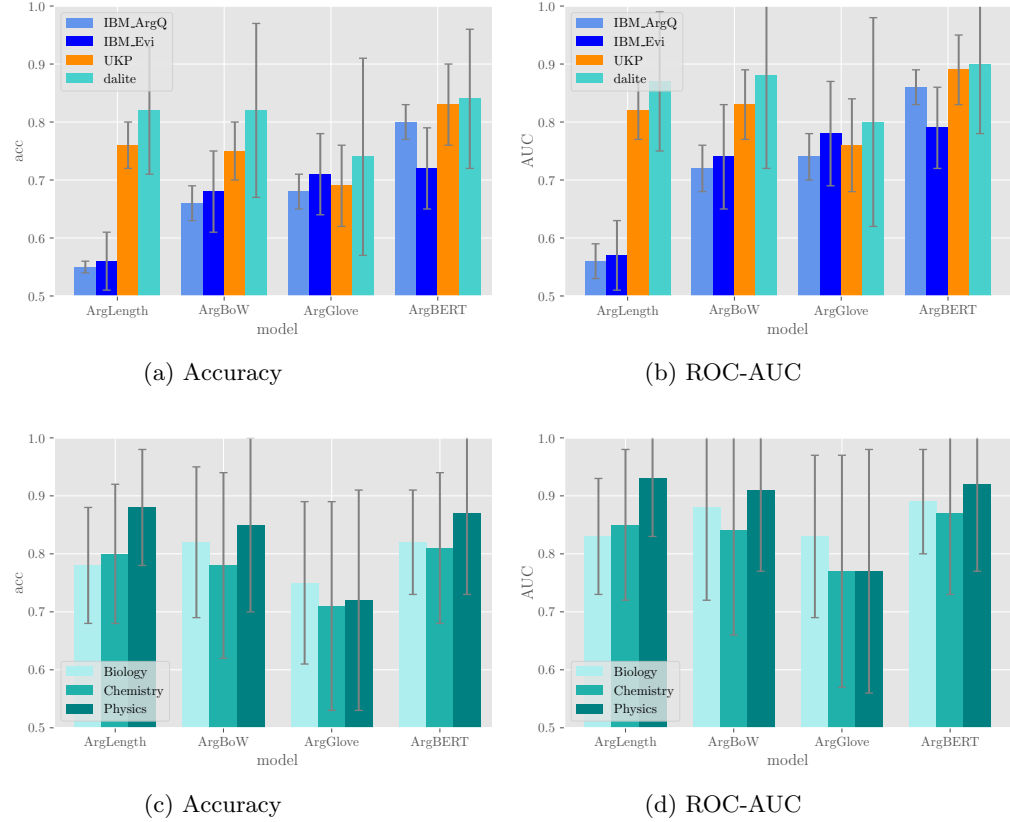.

(a) Accuracy

(b) ROC-AUC

(c) Accuracy

(d) ROC-AUC

Fig. 2: Pairwise ranking classification accuracy and ROC-AUC for different models in across datasets in figures 2a and 2b. Figures 2c and 2d split the performance for the **dalite** dataset, across disciplines. All results evaluated using cross-topic validation, where the number of folds equals the number of topics
.

rankings, starting from these pairwise preference data can be accomplished using research from the information retrieval community [5]. Work on deriving point wise scores for argument pairs is proposed as a Gaussian Process Preference Learning task by [23]. Seeing the lack of pointwise labels for overall convincingness, [25] released a dataset where they collect this data as well. A comparable source of data inside the myDALITE platform are the feedback scores teachers can optionally provide to students on their explanations.

References

1. Bhatnagar, S., Lasry, N., Desmarais, M., Charles, E.: DALITE: Asynchronous Peer Instruction for MOOCs. In: European Conference on Technology Enhanced Learning. pp. 505–508. Springer (2016)
2. Univeristy of British Columbia, T..L.T.: ubc/ubcpi (Aug 2019), `https://github.com/ubc/ubcpi`, original-date: 2015-02-17T21:37:02Z
3. Cambre, J., Klemmer, S., Kulkarni, C.: Juxtapeer: Comparative Peer Review Yields Higher Quality Feedback and Promotes Deeper Reflection. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18. pp. 1–13. ACM Press, Montreal QC, Canada (2018). https://doi.org/10.1145/3173574.3173868
4. Charles, E.S., Lasry, N., Bhatnagar, S., Adams, R., Lenton, K., Brouillette, Y., Dugdale, M., Whittaker, C., Jackson, P.: Harnessing peer instruction in- and out-of class with myDALITE. In: Fifteenth Conference on Education and Training in Optics and Photonics: ETOP 2019. p. 11143_89. Optical Society of America (2019), `http://www.osapublishing.org/abstract.cfm?URI=ETOP-2019-11143_89`
5. Chen, X., Bennett, P.N., Collins-Thompson, K., Horvitz, E.: Pairwise ranking aggregation in a crowdsourced setting. In: Proceedings of the sixth ACM international conference on Web search and data mining. pp. 193–202 (2013)
6. Crouch, C.H., Mazur, E.: Peer instruction: Ten years of experience and results. American Journal of Physics **69**(9), 970–977 (2001)
7. Denny, P.: The Effect of Virtual Achievements on Student Engagement. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 763–772. CHI '13, ACM, New York, NY, USA (2013). https://doi.org/10.1145/2470654.2470763, `http://doi.acm.org/10.1145/2470654.2470763`, event-place: Paris, France
8. Denny, P., Hamer, J., Luxton-Reilly, A., Purchase, H.: PeerWise: Students Sharing Their Multiple Choice Questions. In: Proceedings of the Fourth International Workshop on Computing Education Research. pp. 51–58. ICER '08, ACM, New York, NY, USA (2008). https://doi.org/10.1145/1404520.1404526, `http://doi.acm.org/10.1145/1404520.1404526`, event-place: Sydney, Australia
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
10. Gleize, M., Shnarch, E., Choshen, L., Dankin, L., Moshkowich, G., Aharonov, R., Slonim, N.: Are you convinced? choosing the more convincing evidence with a Siamese network. arXiv preprint arXiv:1907.08971 (2019), `https://www.aclweb.org/anthology/P19-1093/`

11. Habernal, I., Gurevych, I.: Which argument is more convincing? Analyzing and predicting convincingness of Web arguments using bidirectional LSTM. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1589–1599. Association for Computational Linguistics, Berlin, Germany (Aug 2016). https://doi.org/10.18653/v1/P16-1150, https://www.aclweb.org/anthology/P16-1150
12. Herbrich, R., Graepel, T., Obermayer, K.: Support vector learning for ordinal regression (1999), publisher: IET
13. Joachims, T.: Optimizing search engines using clickthrough data. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 133–142 (2002)
14. Khosravi, H., Kitto, K., Williams, J.J.: Ripple: A crowdsourced adaptive platform for recommendation of learning activities. arXiv preprint arXiv:1910.05522 (2019)
15. Lippi, M., Torroni, P.: Argumentation mining: State of the art and emerging trends. ACM Transactions on Internet Technology (TOIT) 16(2), 10 (2016)
16. Louis, A., Nenkova, A.: What Makes Writing Great? First Experiments on Article Quality Prediction in the Science Journalism Domain. Transactions of the Association for Computational Linguistics 1, 341–352 (2013)
17. Moens, M.F., Boiy, E., Palau, R.M., Reed, C.: Automatic detection of arguments in legal texts. In: Proceedings of the 11th international conference on Artificial intelligence and law. pp. 225–230 (2007)
18. Nguyen, D., Doğruöz, A.S., Rosé, C.P., de Jong, F.: Computational Sociolinguistics: A Survey. arXiv preprint arXiv:1508.07544 (2015)
19. Pennington, J., Socher, R., Manning, C.D.: Glove: Global Vectors for Word Representation. In: EMNLP. vol. 14, pp. 1532–1543 (2014)
20. Persing, I., Ng, V.: End-to-End Argumentation Mining in Student Essays. In: HLT-NAACL. pp. 1384–1394 (2016)
21. Pollitt, A.: The method of Adaptive Comparative Judgement. Assessment in Education: Principles, Policy & Practice 19(3), 281–300 (2012). https://doi.org/10.1080/0969594X.2012.665354, https://doi.org/10.1080/0969594X.2012.665354, publisher: Routledge _eprint: https://doi.org/10.1080/0969594X.2012.665354
22. Potter, T., Englund, L., Charbonneau, J., MacLean, M.T., Newell, J., Roll, I., others: ComPAIR: A new online tool using adaptive comparative judgement to support learning with peer feedback. Teaching & Learning Inquiry 5(2), 89–113 (2017)
23. Simpson, E., Gurevych, I.: Finding Convincing Arguments Using Scalable Bayesian Preference Learning. Transactions of the Association for Computational Linguistics 6, 357–371 (2018), https://www.aclweb.org/anthology/Q18-1026
24. Tan, C., Niculae, V., Danescu-Niculescu-Mizil, C., Lee, L.: Winning Arguments: Interaction Dynamics and Persuasion Strategies in Goodfaith Online Discussions. arXiv:1602.01103 [physics] pp. 613–624 (2016). https://doi.org/10.1145/2872427.2883081, http://arxiv.org/abs/1602.01103, arXiv: 1602.01103
25. Toledo, A., Gretz, S., Cohen-Karlik, E., Friedman, R., Venezian, E., Lahav, D., Jacovi, M., Aharonov, R., Slonim, N.: Automatic Argument Quality Assessment-New Datasets and Methods. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 5629–5639 (2019), https://www.aclweb.org/anthology/D19-1564.pdf

26. Weir, S., Kim, J., Gajos, K.Z., Miller, R.C.: Learnersourcing subgoal labels for how-to videos. In: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing. pp. 405–416. ACM (2015)
27. Williams, J.J., Kim, J., Rafferty, A., Maldonado, S., Gajos, K.Z., Lasecki, W.S., Heffernan, N.: AXIS: Generating Explanations at Scale with Learnersourcing and Machine Learning. In: Proceedings of the Third (2016) ACM Conference on Learning @ Scale - L@S '16. pp. 379–388. ACM Press, Edinburgh, Scotland, UK (2016). https://doi.org/10.1145/2876034.2876042