Learnersourcing Quality Assessment of Explanations for *Peer Instruction*

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 an instance of "learnersourcing" higher quality explanations that can be shown to future students. 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 study compares the performance of state-of-the-art methods across different reference datasets and validation schemes, and our findings lays the groundwork for future research along this novel line of inquiry. We demonstrate that when we extend the task of argument quality assessment along the dimensions of *convincingness*, from curated datasets to data from a real learning environments, new challenges arise, and simpler vector space models perform as well as a state-of-the-art neural approach.

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][15], 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 [23][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 [22], 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? [28]

This challenge is a hallmark of learning environments that leverage *learnersourcing*[27].

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 a state-of-the-art neural approach, 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 [27], 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 [15] is a similarly built system, but with an added 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 [28] system prompts students to generate an 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 [23] 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 [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. Moderating the quality of these explanations is important work that can begin with some unsupervised clustering [10], but identifying the best explanations that promote learning is at the heart of this research.

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 [16] and are at the heart of the IBM Project Debater initiative.

Work in the area of automatic evaluation of argument quality finds its roots in detecting evidence in legal texts [18], 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 earlier efforts included work on automatically scoring of persuasive essays [21] and modelling persuasiveness in online debate forums [25]. 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 [12], the authors propose a feature-rich support vector machine, as well as an end-to-end neural approach based on pre-trained GloVe vectors[20] and a bidirectional Long-Short-Term Memory network for the pairwise classification task. This is extended in [11], where the authors build a Siamese network architecture, where each leg is a BiLSTM, taking as input the pair of explanations as GloVe embeddings, 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 [26] release a dataset of argument pairs and show

that the pre-trained weights can be fine tuned to 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 *convincingness*, 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 argument pair from each of the datasets.

3.1 UKP & IBM

UKPConvArgStrict[12], 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, collected from online debate portals. 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 was released by the research group associated with IBM Project Debater, **IBMArgQ-9.1kPairs**[26], 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 [11]. 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. We filter the dataset to only include topics that have at least 50 associated argument pairs.

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

(a) " A pair of arguments from the UKP dataset, for the prompt topic: "uniform". Argument a2 is labelled as more convincing.

a1	a2
I take the view that, school uniform is very comfortable. Because there is the gap between the rich and poor, school uniform is efficient in many ways. If they wore to plain clothes every day, they concerned about clothes by brand and quantity of clothes. Every teenager is sensible so the poor students can feel inferior. Although school uniform is very expensive , it is cheap better than plain clothes. Also they feel sense of kinship and sense of belonging. In my case, school uniform is convenient. I don't have to worry about my clothes during my student days.	I think it is bad to wear school uniform because it makes you look unatrel and you cannot express yourself enough so band school uniform OK

(b) " A pair of arguments from the IBM_ArgQ dataset, for the prompt topic: "privacy". Argument a1 is labelled as more convincing.

a1	a2
if a company is not willing to openly say what they are going to do with my data, they shouldn't be allowed to do it.	if you are against information privacy laws, then you should not object to having a publicly accessible microphone in your home that others can use to listen to your private conversations.

(c) " A pair of arguments from the dalite dataset, for the prompt topic: "C16 Q12 Two Positive Charges Electric Field Lines Magnitudes.csv". Argument a1 is labelled as more convincing.

a1	a2
At B, the electric field vectors cancel (E=0). C is further away than A and is therefore weaker.	A is closest, B experiences the least since it is directly in the middle, and C the least since it is most far away.

Table 1: Examples of argument pairs from each dataset

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 the most convincing. We construct our **dalite** dataset by keeping only the observations where students chose a peer's explanation as more convincing than their own, on the review step.

To ensure internal reliability, we only keep argument explanations that were also 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. In **IBM_ArgQ** and **UKP**, annotators are presented pairs of arguments that are always for the same stance, in order to limit bias due to their opinion on the topic when evaluating which argument is more convincing (this is also true of many of the pairs in **IBM_Evi**).

In a *Peer Instruction* learning environment, other pairings are possible, and pedagogically relevant. In **dalite**, the majority of students keep the same answer choice on the review step, 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 that models must learn, and 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"*)? (We leave this for future work).

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.

Thirdly, each argument pair is made up of the one written by the current learner, while the other is an alternative generated by a peer that the current learner chose as more convincing. In this respect, our the **dalite** dataset is different than the reference datasets, since it is the same person, the current learner, who is author and annotator. In the refernce datasets, data was always independantly annotated using crowdsourcing platforms.

4 Methodology

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 [13]. 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) [14], but also more recently in evaluating the journalistic quality of newspaper and magazine articles [17]. The study accompanying the release of **UKP** first proposed this method for predicting which argument is more convincing [12].

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 vector space models for text classification is the exclusion of words that are "out-of-vocabulary" in the test set when compared to the training data. This is addressed in experiments for this task that employ language models that offer pre-trained word-embeddings that have already learned from massive corpora of text [12][11]. In our `ArgGloVe` model, we encode each token of each argument using 300-dimensional GloVe vectors [20], and represent each argument as the average of its token-level 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 [26], 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 for the task of sequence classification 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 the boundary between arguments `A` and `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 environments built upon learnersourcing of student explanations. This will help automatically moderate the

[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

quality of content in the database, so as to be able to be present student explanations that are *convincing* to future students as they engage with their own reasoning when they answer the same 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.

However standard practice in the argument mining community is to employ "cross-topic" validation for this task. For each fold, the arguments for one topic (or question item, in the case of **dalite**) are held out from model training. Evaluating model performance on a yet unseen topic is a stronger estimate for how a model will perform when new question items are introduced into the content base. We evaluate our models using both validation schemes.

4.5 State of the Art

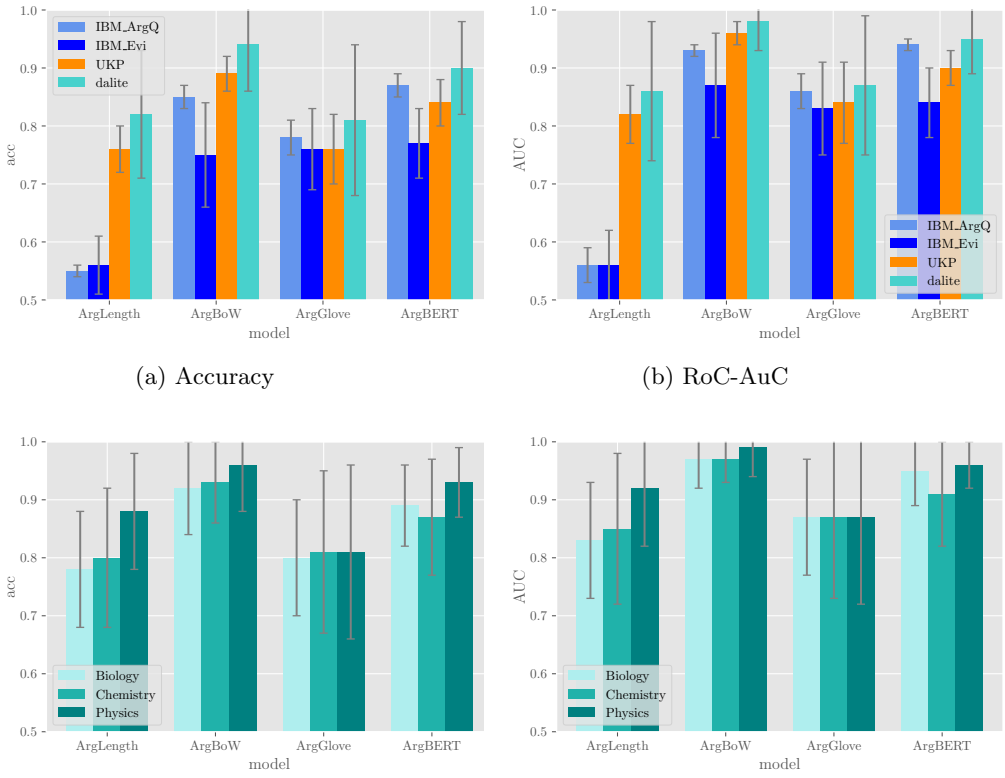
In Table 4, we denote, to the best of our knowledge, the state-of-the-art performance for each dataset on the task of pairwise classification for *convincingness*. The state-of-the-art performance for pairwise classification accuracy on the **IBM_Evi** dataset employs Siamese architechture[11], where 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.

Dataset	Acc	AUC	model
UKP	0.83	0.89	ArgBERT[26]
IBM_ArgQ	0.80	0.86	ArgBERT[26]
IBM_Evi	0.73	-	EviConvNet[11]

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 & Discussion

The performance of different models across the different datasets are presented in figures 1 and 2, using either 5-fold cross-validation, or cross-topic validation, respectively.



(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

.

The first point we remark is that the performance of `ArgBERT`, which is the state-of-the-art for our reference datasets, also performs relatively well on **dalite** data. This, in part, provides some support for the premise that student explanations, and the peer-vote-data we extract from a *peer instruction* learning environment, can actually be modelled via a pairwise argument ranking methodology.

We propose that this validates our line of inquiry, and lays the foundation for future research in this area.

Second, we see that the baseline performance set by `ArgLength` is not uniform across the datasets, which is due in most part to how carefully curated the datasets are to begin with. **IBM_ArgQ** and **IBM_Evi** are datasets that were released, in part, for the purpose of promoting research on automatic assessment of argument quality, and are meant to serve as a benchmark that cannot simply be learned with word counts alone. As can be seen in Table 2, $\overline{\Delta wc}$ is greatest **UKP** and **dalite**, as they are less carefully curated, and hence there are many more argument pairs that have a larger relative difference in length. This is particularly relevant in the context of learning environments based on *peer instruction*: depending on the importance placed on the the review step, students may, or may not, truly engage with their peers' explanations, reverting to simply choosing explanations based solely on how many words they have to read (or not). This sets the bar very high for other more sophisticated approaches.

Third, when evaluating performance using 5-fold cross-validation, as shown in figure 1, the relatively simple `ArgBoW` performs at least as well, if not better, than all other models, including the state-of-the-art `ArgBERT`. Since this is not the case in figure 2, where each individual topic is completely held out from training, we surmise that `ArgBoW` models can learn the necessary requisite vocabulary to explain what makes an argument more *convincing* and perform well under conditions where at least some data is available for a topic. This effect is pronounced for **dalite**, where the vocabulary is relatively constrained, as seen in Table 2, where the ratio of N_{vocab} to N_{args} is lowest for **dalite**.

This result is not without precedent: in a study on the 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 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[17]. While `ArgBoW` will suffer from the *cold-start* problem when new question items are added to the set of learning activities, as no student explanations are yet available, and the vocabulary is still unknown, this may be remedied by the addition of calculated features to the model.

Fourth, we observe the performance of models across different disciplines in **dalite**. Results seems to indicate that argument pairs from items in **dalite:Physics** are easiest to classify. This effect is even more important when we take into account that we have the least data from this discipline (N_{pairs} in Table 2). This may be in part due to the impact of upstream tokenizers, which fail to adequately parse arguments that have a lot of chemical equations, or molecular formulae. Most of the items in **dalite:Physics** are conceptual in nature, and the arguments contain fewer non-word tokens.

Finally, we highlight the variance in performance across models, datasets, and validation schemes. We posit that the task of pairwise classification of argument

quality, along the dimension of *convincingness*, is more challenging when the data is generated by students as part of a learning activity, than with data collected from crowd-sourcing annotation platforms and online debate portals. This effect is expected to be more pronounced in explanations for items from STEM education, as the difference between correct and incorrect answer choices will be more prevalent than in learning contexts focused on humanistic disciplines. When a student is comparing their explanation, which may be for a incorrect answer choice, with that for an explanation of a correct answer choice, we refer to this as "wr". This would be the equivalent of argument pairs which contained arguments of opposite stance, which is only true in **IBM_Evi**. Of note is the relative stable performance of `ArgGlove` across datasets. This may further indicate that there is promise in vector space approaches, as the semantic information captured in well-trained word embeddings can be leveraged to address the challenge when there is large variance in words used to express arguments (most pronounced in **UKP**).

6 Future Work

The performance of vectors space models in this study indicates that more work should be done in now expanding the representation of text arguments using calculated features. This approach has been the best performing in other studies: [17] and [19] and both 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 and teaching contexts.

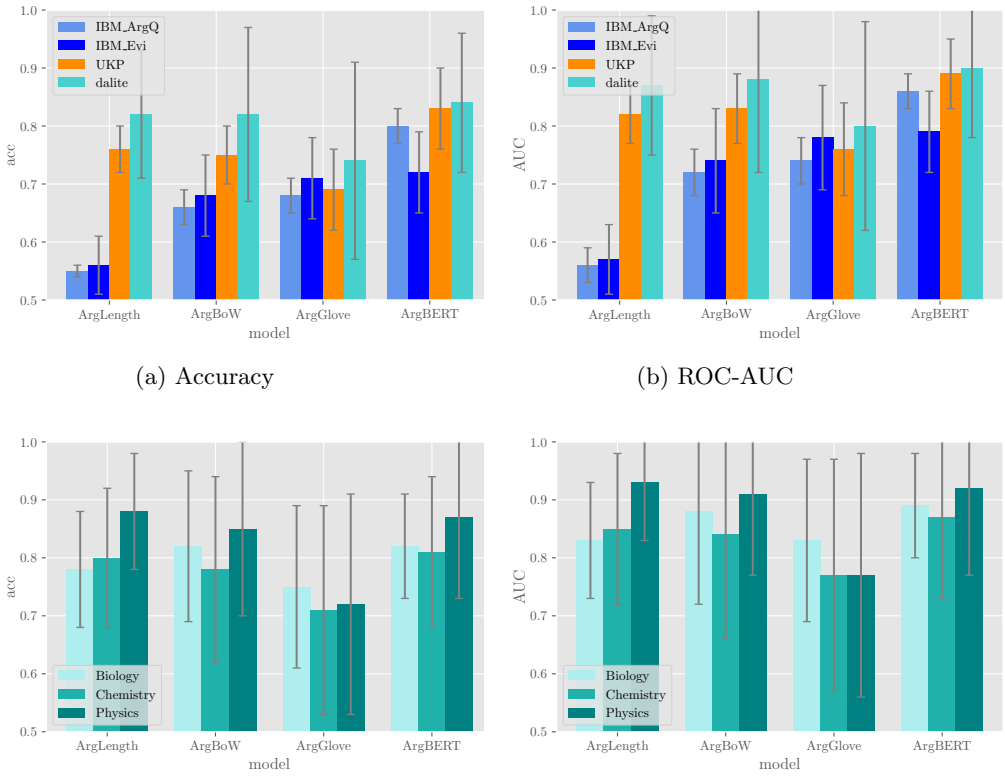
A second important line of future work lies in inferring a global ranking of argument quality, using pairwise preference data. In this study we do not ever infer which are, overall, the most convincing student explanations for any given item. There is a rich body of research from the information retrieval community [5] that can be leveraged here. Work on deriving point wise scores for argument pairs is proposed as a Gaussian Process Preference Learning task by [24]. Seeing the lack of pointwise labels for overall convincingness, [26] 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.

7 Acknowledgements

NSERC Discovery grant

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)



(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

.

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. Gagnon, V., Labrie, A., Desmarais, M., Bhatnagar, S.: Filtering non-relevant short answers in peer learning applications. In: Proc. Conference on Educational Data Mining (EDM) (2019)
11. 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/
12. 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
13. Herbrich, R., Graepel, T., Obermayer, K.: Support vector learning for ordinal regression (1999), publisher: IET
14. 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)
15. Khosravi, H., Kitto, K., Williams, J.J.: Ripple: A crowdsourced adaptive platform for recommendation of learning activities. arXiv preprint arXiv:1910.05522 (2019)

16. Lippi, M., Torroni, P.: Argumentation mining: State of the art and emerging trends. ACM Transactions on Internet Technology (TOIT) **16**(2), 10 (2016)
17. 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)
18. 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)
19. Nguyen, D., Doğruöz, A.S., Rosé, C.P., de Jong, F.: Computational Sociolinguistics: A Survey. arXiv preprint arXiv:1508.07544 (2015)
20. Pennington, J., Socher, R., Manning, C.D.: Glove: Global Vectors for Word Representation. In: EMNLP. vol. 14, pp. 1532–1543 (2014)
21. Persing, I., Ng, V.: End-to-End Argumentation Mining in Student Essays. In: HLT-NAACL. pp. 1384–1394 (2016)
22. 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
23. 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)
24. 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
25. Tan, C., Niculae, V., Danescu-Niculescu-Mizil, C., Lee, L.: Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith 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
26. 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
27. 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)
28. 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