Modeling review helpfulness with augmented transformer neural networks

1st Yunkai Xiao*§ and Tianle Wang†§ 2nd Xinhua Sun† 3rd Yicong Li‡ 4th Yang Song¶
5th Jialin Cui* 6th Qinjin Jia* 7th Chengyuan Liu* 8th Edward F. Gehringer*
**North Carolina State University*
*Raleigh, North Carolina, United States
*{yxiao28, jcui9, qjia3, cliu32, efg}@ncsu.edu
†*Beijing University of Posts and Telecommunications,* ‡*Tsinghua University*
†‡Beijing, China
†{louiswong.cs, sunxh0529}@bupt.edu.cn, ‡li-yc19@mails.tsinghua.edu.cn
¶*University of North Carolina at Wilmington*
¶Wilmington, North Carolina, United States
¶songy@uncw.edu

***Abstract*—The past two years have witnessed a dramatic change in the delivery of education, as most providers have pivoted to remote online learning. With the enrollment of some MOOCs platforms, for example, Coursera, going up by 444% between mid-March and mid-September 2020,[1] practical teaching on a large scale attracted significant attention from the public. For online educators, this was manifested as a significant increase in assessment workload. MOOC instructors have long used peer assessment to evaluate work submitted by online learners. Prior research has shown that when a student believes feedback is helpful, then the suggestion given by that feedback is more likely to be implemented. Researchers have studied machine learning models for detecting problem statements or suggestions in feedback. However, these models do not work as well in detecting secondary features like helpfulness. This paper introduces a new augmented model for detecting helpfulness and tests it against the original model that uses raw text, yielding a 7% increase in performance measured in F1 score.**

***Index Terms*—Machine learning algorithms, Classification algorithms, STEM, Educational technology, Engineering education**

I. Introduction

A. Peer assessment

The COVID pandemic has reshaped education to a new format; teaching and learning online has become a new norm for many people. As online courses, especially massive open online courses, or MOOCs, got practiced more often, online peer assessment has gained substantial popularity. Courses taught on such platforms have the characteristic of being highly interactive compared to traditional online lectures. Beyond lecture videos that everyone could access, students learn from frequent interactions with the platform, participating in online discussions, and submitting homework online. These homework assignments are not limited to auto-gradable formats like multiple-choice questions but can include free responses, creative essays, and coding projects. The nature of MOOCs creates a massive gap between the number of students taking courses and the number of staff members who can answer their questions and grade their homework. To meet this challenge, MOOC platforms let students review each other's work, not just for grading but also to provide formative feedback to help each other improve their skills. Learning from interactions, specifically reading and writing feedback, is an important supplement of reviewing videos, notes, and practice [1]–[3]. These activities, commonly categorized under social learning, involve changing participants' understanding of knowledge through social interactions such as the review activities mentioned above [4].

The feedback students give to each other is of varying quality. One student might spend an adequate amount of time reviewing his/her peers' work and give solid feedback, while others may only skim through the work or give vague and arbitrary comments on the artifact being reviewed. Though platforms have devised sophisticated algorithms [5], such as reputation systems [6] [7] to derive accurate quantitative feedback, such as grades, the quality of the textual feedback has been studied less thoroughly. In the following sections of this paper, the word "feedback", "review," and "comment", as well as "reviewee" and "receiver" are used interchangeably.

B. High quality feedback

Some researchers have qualitatively looked at the issue; Hansen and Liu [8] described how to teach students to give effective reviews. They mentioned that having clear "thesis statements," clearly indicating what problem needs to be addressed and how to resolve it, is key to an effective review. Nelson and Schunn [9] divided characteristics of effective peer reviews into two categories: cognitive and affective. Cognitive features include summarization, specificity, explanations, and scope, while "specificity" is a combination of problems, solutions, and localization. Affective features, also called affective language, include praise, mitigation compliments, and other

§Equal contribution

[1]Coursera 2020 Impact Report: https://about.coursera.org/press/wp-content/uploads/2020/09/Coursera-Impact-Report-2020.pdf

forms of mitigation. The purpose of their study was to test their proposed model, looking into how different types of feedback affect feedback implementation behavior.

We choose some features based on our specific task: problem statements, indications of a problem in the artifact; localization, the specific point in the text where there is a problem if one exists; suggestions of how to improve the work; positive tone, whether the sentiment of the review is generally positive; and praise, an indication of appreciation by the reviewer. These features are later referred to as "the five characteristics" in this paper. Nelson and Schunn's research [9] provides an overview of the qualities of "good reviews." However, their study employed people who have not taken the class as evaluate to measure the effectiveness of feedback, and did not taken into consideration of students' perspective on whether a review is helpful or not.

C. "Feeling helpful" and execution

Students often dislike or show resistance to reviews when they view them as unfair or unhelpful. Their perspective of whether a review is fair lies in their perception of whether a review is useful or positive [10] Whether a student would trust a peer review is influenced by several factors, such as their concerns about the validity, reliability, bias, and fairness of reviews they received. When different reviews are conflicting with each other, or reviews done clearly without diligence, students will find it challenging to implement revisions rendering that review unhelpful from their point of view [11]. The problem is, there has not been a way to automatically detect whether a student may think a review is helpful or not.

We would like to increase the number of reviews that students perceive as helpful, and such automatic detection could help achieve that goal—especially if reviewers are given feedback on helpfulness, and encouraged to improve their reviews before submitting them.

How do we construct such an automated way of "helpful detection"? Traditionally, detecting sentiments and characteristics could be achieved by building machine learning models. These models could intake raw text and classify whether a certain characteristic is there or not in a binary format [12]. If there exists an abundant amount of annotated data for training, such models trained in supervised means may perform well. However, the diverse nature of human opinion makes collecting the amount of data to create a general idea of "helpful" very hard. This paper explores a new possibility of augmenting a machine learning model with prior knowledge in achieving better performance when data is limited.

This paper is organized as follows. Section II discusses related research. Section III covers the data, methods used in this research, and results. Section IV presents conclusions and discussions. Finally, a discussion regarding things beyond this study and future work is included in Section V.

II. Literature review

A. Who is using peer assessment?

Expertiza is an online system that allows instructors to assign homework assignments to students and have them assess each other in a double-blind manner. When students receive peer feedback, whether in MOOCs or other online learning platforms such as Expertiza or SWoRD [13], they tend to reflect on their own work. Research has shown that given a clear set of rubrics, with four or more non-experts reviewing an artifact, the combination of their feedback may achieve similar accuracy as a subject expert [14]. Expertiza offers students to review each other's work on a given set of rubrics defined by instructors. This is done in multiple rounds so students could get the most benefits from formative feedback - all but the last round focus on improving the submissions. The final round requires students to evaluate works they have been reviewing. While doing the first few rounds, making helpful comments for student authors is crucial since reviewees rely on these review comments to make revisions and hopefully improve their artifact [15].

B. What kind of review is helpful from an educational perspective?

Prior research in this field indicates that comments containing suggestions could help students make more effective revisions to their work. In addition, Tseng and Tsai [15] have pointed out when reviews are done as soon as students submit their work, they could encourage students make improvements.

Other researchers mentioned that there are other features beyond suggestions that play an important role in helping students. Such features include whether comments contain praise and whether reviewers offer their opinions in a manner that avoids confrontation [9]. While researchers' opinions on such features are controversial, we would like to research their impact on review helpfulness [16].

C. Who benefits from peer feedback?

Reviews and review writing mentioned in the section above impacts not only the assessees but also the assessors. Assessors spend time and effort analyzing the work submitted and detect errors within the artifact before giving feedback. This is a meta-cognitive process that assessors need to understand and relate to their peers as well as their peer's work before making comments. Furthermore, to make their comments understandable to assessees, assessors need to use the right language and describe what they think with a lot more details.

Helpful reviews are crucial for assessees to understand pluses and minuses in their work; they also motivate assessors to think more deeply while giving feedback. Both of these activities improve students' learning outcomes.

Researchers have discovered that giving timely feedback to students is crucial for them to revise their work [17]. Reviewees produce better work when reviews are presented shortly after project submission, and reviewers write better reviews when quality of their review writing is prompt to them

as they are writing them. In fact, there have been researches done in this field and has shown positive results [18].

D. How to detect characteristics in review text?

Effective feedback, is beneficial to students, has been researched by many researchers. To automatically distinguish effective and ineffective feedback, some researchers have used machine learning to capture certain characteristics in text [12], [19]. As discussed above, there are characteristics well known by researchers that contribute to feedback effectiveness from an educational perspective. Zingle et al. have built rule-based models and classical and neural-network-based machine learning models to detect suggestions in reviews automatically. Xiao et al. have built machine learning models based on transformers to detect problem statements within reviews automatically. Other researchers, such as Xiong [20]–[23], and Sapna [24] have built models using similar techniques to classify characteristics in reviews.

We followed many of their approaches to build a baseline and compared them against the model we are going to propose.

E. Feature engineering

In Zingle's research, he compared the performance of classical machine learning models and neural-network-based machine learning models, each requiring a different way of creating text features.

For classical machine learning models such as decision tree and logistic regression, they relied on term frequency inverse document frequency or TF-IDF to create embedding from raw text. A drawback of this approach is that the embedding created by it depends on the size of vocabulary, thus making feature space very sparse. However, the result Zingle reported in his paper seems to be unaffected by this sparsity.

As for the neural network approaches, Zingle et al. [19] and Xiao et al. [12] both utilized Global Vectors for Word Representation, or GloVe, embedding.

In Xiao's research, a transformer-based classifier is built. He approached the embedding process with a sub-word tokenizer and a transformer-based embedding layer [25], specifically a Bidirectional Encoder Representations from Transformers based encoder.

F. Models

Models used in the papers mentioned above are used as baselines. Classical models include logistic regression, random forest, multiple flavors of naive Bayesian models, and support vector machines; neural network models covered are text-CNN, LSTM, and a transformer-based model.

In Zingle's research, classical models are built following similar approaches, with a data pipeline of data acquisition, preprocessing, encoding, model training, and evaluation. Left column of Figure 1 explains these standard procedures.

Both Xiao and Zingle's research explored the possibility of using neural networks on classifying characteristics such as problem statements and suggestions. These networks, namely text-CNN and various flavors of LSTM, took into consideration the sequence of text. Although the networks are structured differently internally, the workflows on training them are similar. See middle column of Figure 1 for reference.

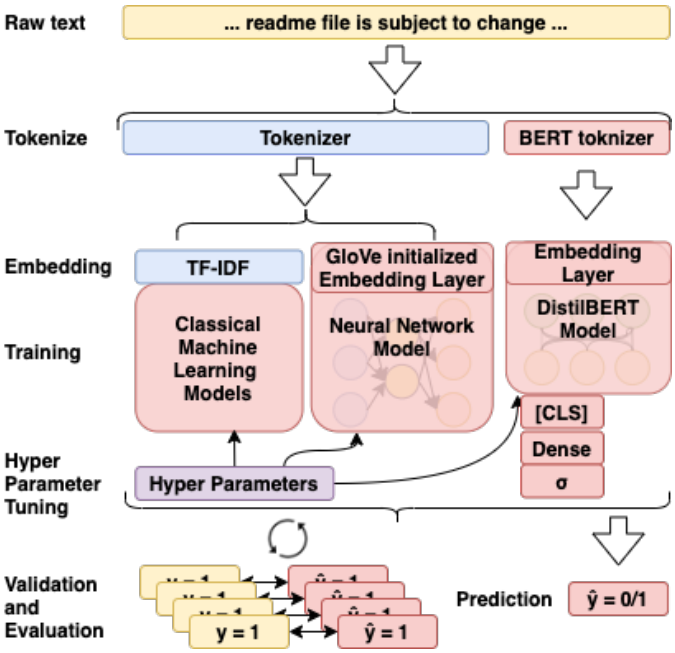

Fig. 1. Workflow for machine learning models

Finally, the last model experimented in Xiao's paper demonstrated the possibility of using transformers in this domain. In his paper, he described the application of a hierarchical neural network on classifying these characteristics [12]. There are several models in text processing developed in the industry after that paper was published [26]–[28]; recent research brought up a distilled version of the popular model, BERT, and named it distilBERT [29]. The distilled BERT model is a lightweight, easy-to-use model which could be finetuned and deployed with much less time [29]. This paper will use this model instead of the one used in Xiao's paper.

Many of the published distilBERT models have been trained on large datasets. A common practice is to use the model trained on the same dataset of the original BERT model, which are the BookCorpus [30] and English Wikipedia [2], to set up a baseline. The model comes with its own tokenizer that tokenizes and encodes raw text into subword tokens. During model training, the tokenizer is modified simultaneously to adopt new vocabulary, providing a new tokenizer when the model is trained. (as illustrated in the right column of Figure 1) When used as classifiers, the [CLS] token is connected by a fully connected layer to the activation function in order to produce a binary output. The authors and creator of BERT describe this token as an aggregate sequence representation specific for classification tasks [31].

The following section will discuss the experiment and means of automatically predicting whether students may think a review is helpful with different approaches. Then compare different approaches with our model.

[2]https://dumps.wikimedia.org/enwiki/

III. Experiment and result

A. Data

a) Nature of data: Datasets used in this experiment are collected from an upper-level-undergraduate/graduate-level computer science class. The class is project-based and peer-reviewed. Expertiza offers students an interface to submit their work and peer review. They would write two rounds of reviews (in English) to their peers for each project, within which there is one round of formative review and one round of summative review. Students will submit their work once, receive feedback from their peers, and revise before final submission. Upon submissions, they are given a few days to review each other. Within each round, a set of rubrics is given to students to guide them in assessing each other. Once feedback is received, students are presented with another interface; within it, they can view reviews they received and the numerical peer-reviewed grades associated with the textual review.

When the feedback period has passed, a new interface opens to students. Here they are asked to express their opinion and give feedback to feedback they have received. In addition, they are asked to label a set of predefined metrics. Such could include *"if they felt this comment is helpful to them," "if they felt the comment raised a valid problem," "if they felt the location of concerns mentioned was clearly expressed,"* and *"if the comment provided a suggestion to revise their work."*

In our experiment, students were incentivized through extra credit to label the reviews they received, with the presence of some of these features. Each semester the instructor selected a few such labeling metrics and set up the user interface for students to annotate these reviews.In this class, assignments are done in teams, which means that the team members are each asked to annotate the same set of review comments.

b) Data collection and preprocessing: Supervised machine learning models are trained with input-output pairs; in the case of binary classification of natural language, the input is often text, and the output being a 0 or 1 indicating whether the text carries desired characteristic.

In the case of peer review data mentioned above, students' peer reviews are treated as input and labels as output. Expertiza allows various input formats, including rich text, to provide a better visual experience and usability for reviewers and reviewers. While some of this information could potentially help examine what is stressed more by reviewers, the models we used in this experiment did not take that into account. Special characters, HTML annotations, and images are first removed from the dataset before next steps, while an inter-rater reliability score is calculated for objective labels. If a coding error has been noted in the review, it is reasonable to assume that problem statements addressing this type of issue are more objective than subjective; thus, all team members should agree on "yes, that review brought up a problem." The average inter-rater reliability across these objective labels (excluding "helpful" as it is more subjective than objective) is 0.88, measured in Cohen's Kappa [32]. Later, we decided to drop all controversial comments, where teammates cannot reach a consensus on objective labels, to ensure data quality. With that said, we did not consider reviews with conflicting opinions of the helpfulness as well; those are reviews being labeled as "helpful" and "not helpful" by different students. We believe that studying what most people concur could lead to findings commonly agreed by the general public.

Imbalance datasets, having too many entries of one class than the other, tends to bias machine learning models since such models are trained to minimize loss. There are several ways of removing bias: adding weights to one class, over-sample or augment the smaller class, downsample the class with more samples, etc... In this experiment, we choose to downsample in order to mitigate bias in our models.

Some of the attributes information for this dataset (post preprocessing and balancing) are following: There are 7,418 annotated reviews concerning the existence of suggestions, 18,392 for problem statements, 5,042 for localization, 2,664 for positive tones, 6,984 for praises, and 3,970 for helpfulness. All of which are balanced with a 1:1 on positive and negative cases. Table I shows a few sample data entries. For tables below, we use the following notation: C1 for Helpful, C2 for Positive tone, C3 for Praise, C4 for Problem statement, C5 for localization, and C6 for suggestion;

TABLE I
Sample data

Comment	C1	C2	C3	C4	C5	C6
The design appears to be simple for the most part. I just feel that explanations are a bit verbose and cloud the understanding of the reader. Also, the user roles seem to be complex and the roles have some complex flow of functionality.	1	1	1	1	1	0
Good test plan. They have considered most use cases and also covered edge cases.	1	1	1	0	0	0
The design appears to be simple for the most part. I just feel that explanations are a bit verbose and cloud the understanding of the reader. Also, the user roles seem to be complex and the roles have some complex flow of functionality.	1	0	0	0	1	1

B. Baseline models

As was brought up in the section above, we attempted to use the same approaches that other researchers have used to complete the task of analyzing the characteristics. These approaches, including classification on single characteristic with classical and neural network-based machine learning models, are used as a baseline to explore our first research question: **RQ1: Would these models work well to detect subjective opinions such as helpfulness?**

We then first re-established the workflows which follow past research cited. Traditional machine learning models are built using the same workflow as shown in Figure 1. The following

models are finetuned and tested in a 5-fold format; they are: logistic regression (LR), classical random forest (RF), random forest with ADA boosting (ADA), random forest with gradient boosting (GB), multinomial naive Bayes (MNB), complement naive Bayes (CNB), support vector machine (SVM), and support vector machine with stochastic gradient descent (SGD).

As for classifying more objective characteristics such as the existence of suggestion, problem identification, and so on, we have achieved similar results with Xiao, and Zingle [12], [19]. While for detecting the subjective characteristic of "helpfulness," things become a little unstable. The results are demonstrated in Table II. Our target (helpfulness) is bold-ed with best performing model for each characteristic underlined.

TABLE II
CLASSICAL MACHINE LEARNING MODEL F1 SCORE

Models	**C1**	C2	C3	C4	C5	C6
LR	**0.69**	0.83	0.79	0.88	0.81	0.86
RF	**0.67**	0.78	0.72	0.78	0.78	0.81
ADA	**0.68**	0.80	0.77	0.86	0.78	0.86
GB	**0.68**	0.79	0.77	0.85	0.77	0.85
MNB	**0.67**	0.78	0.77	0.84	0.80	0.81
CNB	**0.67**	0.78	0.77	0.84	0.80	0.81
SVM	**0.70**	0.83	0.79	0.89	0.81	0.87
SGD	**0.69**	0.82	0.79	0.89	0.80	0.88

Neural networks, being reported in Xiao, and Zingle's papers, did show improvements compared with classical models. With appropriate tuning, both text-CNN and LSTM models outperform average classical machine learning models, with the exception of predicting localization characteristic with LSTM (see Figure 2). The prediction accuracy for the helpfulness characteristic measured in F1 ranks the lowest.

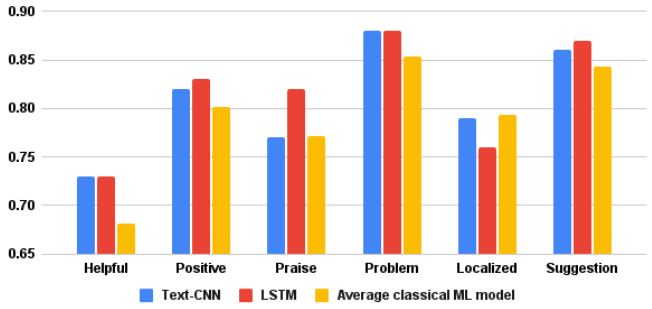

Fig. 2. Neural network & average classical machine learning model F1 score

Finally, the last model within baseline is the transformer-based distilBERT model. Xiao has reported in his result that models utilizing transformers tend to perform better than all other models. We have recreated his experiment and found that it is true for most cases, except for the helpful characteristic, as shown in Figure 3. In prior experiments, both Text-CNN and LSTM models could achieve a F1 score of 0.73 on classifying it, while the distilBERT model could only achieve 0.68.

Till here, **our first RQ is answered: Can we model the helpfulness characteristics in review comments with models used to classify other characteristic? Sure. Does it perform well on this rather subjective task? Not so.**

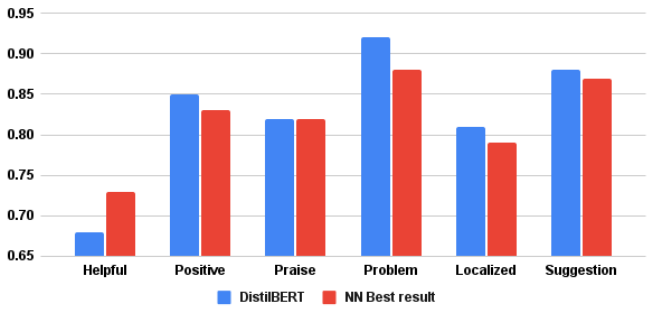

Fig. 3. DistilBERT & other neural network machine learning model F1 score

This finding initially attracted our attention. Why would models, being classical, neural network, or transformer machine learning models, that performs well on classification tasks for all other characteristics fail this single task? By examining the variables within our experiments, we soon found something obvious: the number of parameters and weights on each neuron are at different orders of magnitude. The number of parameters in text CNN rests around 4.5 million, and LSTM around 4.9 million; As for DistilBERT, that number jumped to 66 million. For characteristics that are precise, objective, and clear to define, fewer observations could provide models with enough information to understand their meaning. In comparison, relatively subjective characteristics would require a lot more data. The order of magnitude of our dataset size for helpful characteristic is far from the dataset used in pretraining of the published distilBERT model - being less than 1 MB versus over 6 GB.

This brings up the second research question: if machine learning needs more data to understand meanings that contribute to a characteristic, then can we, as human beings, explain what we understood to the model, thus assisting it in making better decisions? **RQ2: Can we use human knowledge to augment the model by guiding model construction?**

C. Human knowledge

Earlier pedagogical studies have noted that different characteristics within review comments could influence reviews' effectiveness. Some such research comes from field observations, and some of it from interviews and surveys [16]. Could these studies give insight on whether students would judge a review to be helpful, at least to a certain degree?

The characteristics mentioned here are the ones we experimented with, such as the presence of suggestions, localization, problem detection, praise-giving, and positive tone. We will refer to these as "the five characteristics."

At the outset, we sought to validate whether the five characteristics really have some connection with students' perception of helpfulness. First, a data exploration needs to be done. Since comments labeled "helpful" are not always labeled for the five

characteristics , classification is done first to generate labels for these characteristics.

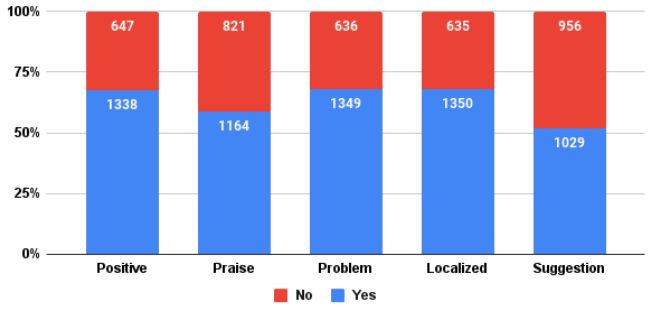

Fig. 4. Review characteristics when students think a review is helpful

Preliminary findings with the dataset have shown some interesting relationships between "helpful" and the other five characteristics predicted by models. What coexists with comments labeled as "helpful" does not seem to be very clear. A quick count on occurrences of characteristics coexisting with a positive "helpful" label (Figure 4) revealed the same.

On the other hand, when comments are labeled as "unhelpful," visualization becomes very intuitive. For example, Figure 5 shows that while praise and positive tone still often exist within unhelpful reviews, other features, such as suggestions, problems, and localization, appeared much less often.

Having a set of characteristics does not guarantee a review comment will be perceived as helpful, but lacking an essential characteristic certainly predicts a perception of "unhelpful."

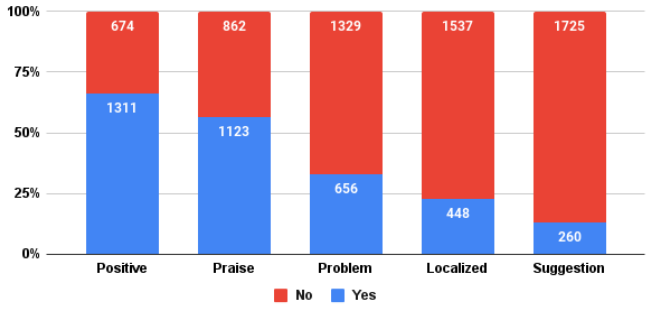

Fig. 5. Review characteristics when students think a review is not helpful

To measure the correlation between these five characteristics and the helpfulness characteristic, we calculated the Pearson correlation coefficient for each of them against students' perceived helpfulness. To obtain a complete view of the influences between these factors, correlation coefficients are calculated with combined characteristics and helpfulness.

Figure 6 shows the correlations of each characteristic and helpfulness on the diagonal, with the rest of the tiles showing correlations between two characteristics combined and helpfulness. It is clear that localization, suggestion, and problem statements are the three most influential factors in making a student think a review is helpful, while providing praise and positive tone shows very weak, but still positive, correlation.

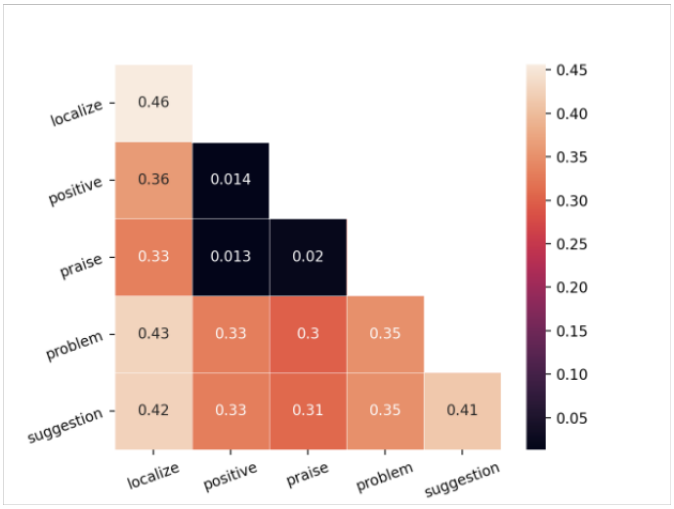

Fig. 6. Correlation of characteristics when students think a review is helpful

An example of what a student considered unhelpful despite all five characteristics is *"The test plan is exhaustive, but the plan is very abstract and can include some details here."* An example of a not helpful comment which does not contain any of the five characteristics is *"I find it a little confusing. But the diagrams help in understanding how the author is approaching the problem."* A comment perceived as helpful with all five characteristics is, *"The team has provided a great visualization to explain their approach and how the user would interact with the system once the changes are made. It would have been great if the team could have given a brief explanation of the flowchart."*, and there is not a single instance of a helpful review when none of the five characteristics exists. It is reasonable to say that these factors indeed influence students' decisions. Thus, knowing them, or having the transformer know them, will perhaps improve the accuracy of its predictions.

D. Augmenting the model with human knowledge

How do we make a model aware of human knowledge? We decided to directly augment the structure of a transformer to achieve this task. As shown in prior experiments, transformers can achieve better accuracy than other models over the five characteristics. Those transformers clearly have "understood" these characteristics to a certain extent. If we could extract or preserve the knowledge they have learned and apply it in training a new model, then we could say that human knowledge is supplied to augment that model.

The proposed model training pipeline is shown in Figure 7. We decided to adapt the trained transformer models described above into the input stream after freezing their weights. Previously, transformer models were trained on different datasets to achieve their individual goals, and such datasets contain

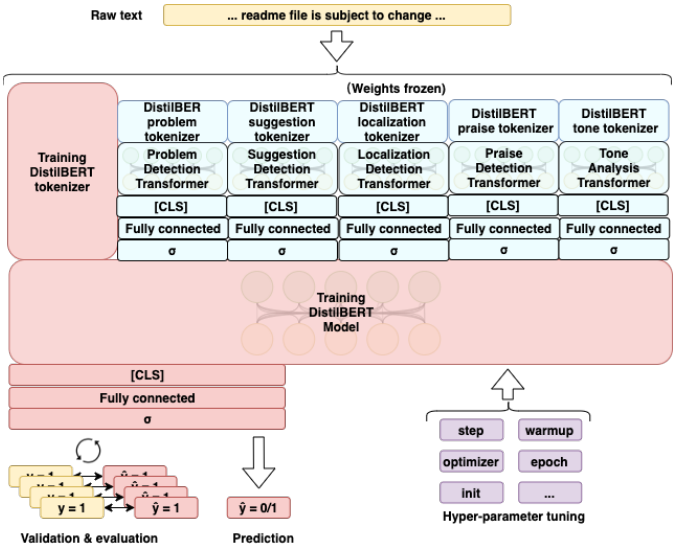

Fig. 7. Proposed augmented transformer model

different lists of vocabularies. When the weights of the transformers are frozen, their embedding layers become static and unable to update. Thus we preserved their own tokenizer so that tokens fed into the embedding layers can still be handled in the same way, with new tokens marked as [UNKNOWN]. The only tokenizer that is getting updated is the one fitted for the transformer model-under-training since the embedding layer of that model can still adopt new tokens in training.

Tokenized words and outputs from the frozen models are fed into the embedding layer of the transformer-model-under-training, so that decisions from models trained for classifying other characteristics are getting adapted in training (through the new embedding layer), which effectively expands the size and knowledge pool of the new model. The rest of the training process is like training any other transformers, and weights in the frozen models are excluded from updating, ensuring the total number of updateable parameters does not increase.

The result of this approach is promising. After augmentation the transformer model achieves an accuracy of 0.753 measured in F1 score. This result is so far the best compared with other models attempted as illustrated in Figure 8, and is much better than it was performing before augmentation.

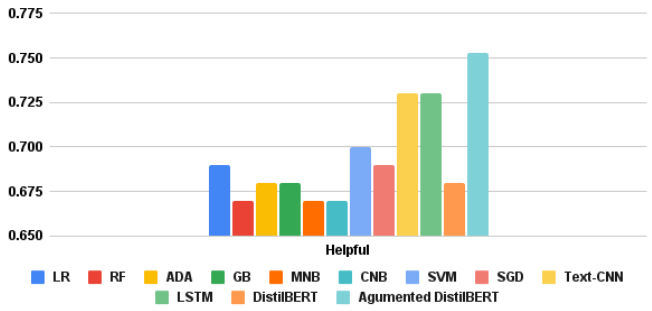

Fig. 8. F1 score on predicting helpful opinion across all models

At this point, **the second research question is answered. Using human knowledge to augment the model is indeed beneficial and will make the model perform better with limited training data.**

IV. Conclusion

This research has answered two questions. First, opinions that are more subjective can be modeled using machine learning in the same way that more objective opinions are, though only to a certain extent. The model's performance measured by F1 score will be far from those models that classify objective characteristics. A transformer outperforms other models while classifying more objective characteristics, but performs poorly on classifying the subjective characteristic of helpfulness. Second, augmenting the transformer with human knowledge does work; it improves its accuracy by a good 7%.

One way to include human knowledge in the model is to incorporate trained models into the structure. Our work does this by adding trained models with wights "frozen" into the input layer of the transformer model-to-be-trained.

Human knowledge may not be completely accurate. When predicted by trained models, our five characteristics showed different levels of correlation with the "helpfulness" characteristic. While other researchers have claimed that these characteristics are related and help to make reviews more effective, it may not always be true. While we agree that these characteristics make reviews helpful, some of them are scarcely as important as described in other researchers' work.

This is the first quantitative study in building machine learning models to study which characteristic makes a feedback comment helpful in peer assessment. Our augmented machine learning model does automatically detect subjective helpfulness. Moreover, when classifying helpfulness, it can achieve results that are similar to some other models' when they classify certain objective characteristics, e.g., LSTM in predicting characteristics regarding localization.

While providing the above findings, this study does have many limitations. First, human knowledge of possible characteristics that may contribute to helpful classification comes from limited research. We did not take into consideration research results that conflict with each other. Second, students' unanimous agreements on a characteristic may not be as reliable as we thought. The annotation process involves having reviewees from the same team label reviews they received. The whole team may be biased towards a review since it is their work getting judged, and they may hold negative feelings toward specific review comments.

V. Future Work

We have assumed that a large number of trainable parameters of the transformer cause poor performance. More experiments on transformers with lesser parameters could be done to validate them with the same dataset used in this study. We could also explore model behavior when we unfreeze the weights of models used in augmentation. Although this would effectively increase the number of the trainable parameters

to six times what it is right now, knowledge learned by those models will not be forgotten. We could experiment on whether this would affect the performance of the new model. Another way of reducing the number of trainable parameters yet being able to carry over knowledge learned from other characteristics is to use one model to train on different tasks — the five characteristics, then merge it into the transformer with or without freezing weights in controlled experiments. Jia has laid out a solution similar to this, while in our case, we cannot co-train models with the same approach since our data for modeling different characteristics do not overlap [33].

With the model established, it is possible to adapt it into the peer review system. One could then experiment with whether the characteristics the model picked up can really support better review writing and knowledge transfer through social learning. The system could prompt reviewers to modify their reviews when they are deemed to be less helpful. If reviewers disagree on what is helpful, we have a chance to calibrate our dataset, rethink our model building, or re-examine if studies in this domain have really derived the correct conclusion.

In conclusion, the models we set up, the technologies we used, and the analyses we carried out are just tools to achieve a greater goal—improving assessment and learning in online courses. We hope our research will help students learn better, both by being challenged in the reviews that they write, and through better insights gained from the reviews they receive.

REFERENCES

[1] J Lu and N Law. Online peer assessment: effects of cognitive and affective feedback. *Instructional Science*, 40(2):257–275, March 2012.

[2] Y H Cho and K Cho. Peer reviewers learn from giving comments. *Instructional Science*, 39(5):629–643, September 2011.

[3] C Evans. Making Sense of Assessment Feedback in Higher Education. *Review of Educational Research*, 83(1):70–120, March 2013.

[4] M S. Reed, A C. Evely, G Cundill, I Fazey, J Glass, A Laing, J Newig, B Parrish, C Prell, C Raymond, and L C. Stringer. What is Social Learning? *Ecology and Society*, 15(4), 2010.

[5] C Piech, J Huang, and Z Chen. Tuned Models of Peer Assessment in MOOCs. In *Proceedings of the 6th Educational Data Mining conference*, page 8, 2013.

[6] J Hamer, K T K Ma, and H H F Kwong. A Method of Automatic Grade Calibration in Peer Assessment. In *Proceedings of the 7th Seventh Australasian Computing Education conference*, page 6, January 2015.

[7] J Cho, K Kwon, and Y Park. Q-rater: A collaborative reputation system based on source credibility theory. *Expert Systems with Applications*, 36(2, Part 2):3751–3760, March 2009.

[8] J. G. Hansen. Guiding principles for effective peer response. *ELT Journal*, 59(1):31–38, January 2005.

[9] M M. Nelson and C D. Schunn. The nature of feedback: how different types of peer feedback affect writing performance. *Instructional Science*, 37(4):375–401, July 2009.

[10] J. Kaufman and C. Schunn. Students' perceptions about peer assessment for writing: their origin and impact on revision work. *Instructional Science*, 2011.

[11] R A Mulder, J M Pearce, and C Baik. Peer review in higher education: Student perceptions before and after participation. *Active Learning in Higher Education*, 15(2):157–171, July 2014. SAGE Publications.

[12] Y Xiao, G Zingle, Q Jia, H R. Shah, Y Zhang, T Li, M Karovaliya, W Zhao, Y Song, J Ji, A Balasubramaniam, H Patel, P Bhalasubbramanian, V Patel, and E F. Gehringer. Detecting Problem Statements in Peer Assessments. In *Proceedings of the 13th Educational Data Mining conference*, May 2020.

[13] K Cho and C D. Schunn. Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. *Computers & Education*, 48(3):409–426, April 2007.

[14] F Daniel, P Kucherbaev, C Cappiello, B Benatallah, and M Allahbakhsh. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys (CSUR)*, 51(1):1–40, 2018.

[15] S-C Tseng and C-C Tsai. On-line peer assessment and the role of the peer feedback: A study of high school computer course. *Computers & Education*, 49(4):1161–1174, December 2007.

[16] A N. Kluger and A DeNisi. The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2):254–284, March 1996. Publisher: American Psychological Association.

[17] E R Fyfe, J R de Leeuw, P F Carvalho, R L Goldstone, J Sherman, D Admiraal, L K Alford, A Bonner, C E Brassil, C A Brooks, et al. Manyclasses 1: Assessing the generalizable effect of immediate feedback versus delayed feedback across many college classes. *Advances in Methods and Practices in Psychological Science*, 4(3), 2021.

[18] L Ramachandran. *Automated Assessment of Reviews*. Ph.D., North Carolina State University, Raleigh, NC, 2013. ISBN: 9781303547966.

[19] G Zingle, B Radhakrishnan, Y Xiao, E Gehringer, Z Xiao, F Pramudianto, G Khurana, and A Arnav. *Detecting Suggestions in Peer Assessments*. International Educational Data Mining Society, July 2019. Publication Title: International Educational Data Mining Society.

[20] W Xiong and D Litman. Identifying Problem Localization in Peer-Review Feedback. In Vincent Aleven, Judy Kay, and Jack Mostow, editors, *Intelligent Tutoring Systems*, Lecture Notes in Computer Science, pages 429–431, Berlin, Heidelberg, 2010. Springer.

[21] H Nguyen, W Xiong, and D Litman. Classroom Evaluation of a Scaffolding Intervention for Improving Peer Review Localization. In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, A. Kobsa, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, D. Terzopoulos, D. Tygar, G. Weikum, S. Trausan-Matu, K.E. Boyer, M. Crosby, and K. Panourgia, editors, *Intelligent Tutoring Systems*, volume 8474, pages 272–282. Springer International Publishing, 2014.

[22] W Xiong and D Litman. Automatically Predicting Peer-Review Helpfulness. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 502–507, Portland, OR, June 2011. Assn. for Computational Linguistics.

[23] W Xiong, Diane L, and Christian S. Assessing Reviewers' Performance Based on Mining Problem Localization in Peer-Review Data. 2010.

[24] S Negi. Suggestion Mining from Opinionated Text. In *Proceedings of the ACL 2016 Student Research Workshop*, pages 119–125, 2016.

[25] H Xiao. bert-as-service. http://github.com/hanxiao/bert-as-service, 2018.

[26] M Lewis, Y Liu, N Goyal, M Ghazvininejad, A Mohamed, O Levy, V Stoyanov, and L Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, 2020. Association for Computational Linguistics.

[27] C Raffel, N Shazeer, A Roberts, K Lee, S Narang, M Matena, Y Zhou, W Li, and P J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[28] Z Zhang, X Han, Z Liu, X Jiang, M Sun, and Q Liu. ERNIE: Enhanced Language Representation with Informative Entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, July 2019.

[29] V Sanh, L Debut, J Chaumond, and T Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *EMCÂ²: 5th Edition Co-located with NeurIPS'19*, February 2020.

[30] Y Zhu, R Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015.

[31] J Devlin, M-W Chang, K Lee, and K Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, MN, June 2019. Assn. for Computational Linguistics.

[32] M. McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 2012.

[33] Q Jia, J Cui, Y Xiao, C Liu, P Rashid, and E Gehringer. ALL-IN-ONE: Multi-Task Learning BERT models for Evaluating Peer Assessments. In *Proc. 14th Educational Data Mining conference*, June 2021.