Improving NLP Model Performance on Small Educational Data Sets Using Self-Augmentation

Keith Cochran[1,2][a], Clayton Cohn[1,2][b] and Peter Hastings[1,2][c]
[1]*DePaul University, Chicago IL 60604, U.S.A.*
[2]*Vanderbilt University, Nashville TN 37240, U.S.A.*

Keywords: Educational Texts, Natural Language Processing, BERT, Data Augmentation, Text Augmentation, Imbalanced Data Sets.

Abstract: Computer-supported education studies can perform two important roles. They can allow researchers to gather important data about student learning processes, and they can help students learn more efficiently and effectively by providing automatic immediate feedback on what the students have done so far. The evaluation of student work required for both of these roles can be relatively easy in domains like math, where there are clear right answers. When text is involved, however, automated evaluations become more difficult. Natural Language Processing (NLP) can provide quick evaluations of student texts. However, traditional neural network approaches require a large amount of data to train models with enough accuracy to be useful in analyzing student responses. Typically, educational studies collect data but often only in small amounts and with a narrow focus on a particular topic. *BERT-based* neural network models have revolutionized NLP because they are pre-trained on very large corpora, developing a robust, contextualized understanding of the language. Then they can be "fine-tuned" on a much smaller set of data for a particular task. However, these models still need a certain base level of training data to be reasonably accurate, and that base level can exceed that provided by educational applications, which might contain only a few dozen examples. In other areas of artificial intelligence, such as computer vision, model performance on small data sets has been improved by "data augmentation" — adding scaled and rotated versions of the original images to the training set. This has been attempted on textual data; however, augmenting text is much more difficult than simply scaling or rotating images. The newly generated sentences may not be semantically similar to the original sentence, resulting in an improperly trained model. In this paper, we examine a self-augmentation method that is straightforward and shows great improvements in performance with different *BERT-based* models in two different languages and on two different tasks which have small data sets. We also identify the limitations of the self-augmentation procedure.

1 INTRODUCTION

In educational contexts, researchers study student reasoning and learning processes. Some of this research is more basic, and focuses on testing hypotheses about what influences learning in various contexts. Other research is more applied, intending to produce interactive environments that can provide students with immediate feedback based on their progress and help make their learning more effective and efficient.

Some types of student data can be easily evaluated computationally with simple programs. However, when student data is textual, evaluation is more challenging. When done by hand, it can be quite time-consuming and subjective. This negatively affects basic research, requiring the researchers to spend more time scoring student answers and less time testing new hypotheses. For interactive learning environments, either classroom-based or computer-supported, manual grading of student work would normally be done after the students have left the classroom, delaying feedback and resulting in missed learning opportunities while the subject is still fresh in the students' minds. Real-time computational evaluations of texts could improve student learning opportunities and facilitate the testing of learning theories.

One avenue of text-based basic research is motivated by national and international literacy standards

[a] https://orcid.org/0000-0003-0227-7903
[b] https://orcid.org/0000-0003-0856-9587
[c] https://orcid.org/0000-0002-0183-001X

Cochran, K., Cohn, C. and Hastings, P.
Improving NLP Model Performance on Small Educational Data Sets Using Self-Augmentation.
DOI: 10.5220/0011857200003470
In *Proceedings of the 15th International Conference on Computer Supported Education (CSEDU 2023) - Volume 1*, pages 70-78
ISBN: 978-989-758-641-5; ISSN: 2184-5026

(Achieve, Inc, 2013; OECD, 2021), which state that students should learn to think critically about science-related texts. Students should be able to deeply understand scientific arguments that they have read, evaluate the arguments, and produce sound written summaries of their own. This connects with key societal issues beyond basic literacy, including bias, “fake news”, and civil responsibility.

Discourse psychology researchers have posited that reading is fundamentally influenced by the *goal* of the reading (Britt et al., 2017). One way of testing this theory is to ask participants to read short articles on a topic that contain varying viewpoints, then ask them to write what they think about that topic in different contexts. For example, they could be asked to do the writing in a lab or at home or to write “to friends” or as an academic submission. Researchers could measure the extent to which the participants personalize their writings, express opinions, and cite sources in different settings. Differences between the conditions can show how the participants approach the tasks. For example, prior research has shown that students express themselves more personally when they’re writing in a lab at a university vs. when they’re doing the “same task” from home. Of particular interest is the extent to which students consider different viewpoints in different contexts or whether they focus only on a subset that may agree closely with their prior biases.

An example of a text-based interactive research project is the SPICE curriculum for science learning curriculum (Zhang et al., 2020). Its goal is to strengthen students’ understanding of concepts and processes in earth sciences. In one example lesson, students are given a diagram that ostensibly comes from another student. They are asked to analyze the diagram, identifying and explaining any mistakes.

The students’ responses range from a few words to one or two sentences. Research based on analyzing these responses can shed light on the students’ depth of understanding of the relevant concepts and processes. At a deeper level, this research also shares the goal of shedding light on students’ abilities to perform causal reasoning in complex situations to inform educational advances. More immediately, automatic computational evaluation of the student’s answers can be used as formative feedback, helping them to improve their answers and learn more in the process.

Typically, student responses in both of these contexts will have a certain structure or intent, which aligns with a grading rubric and shows the extent to which the material has been understood. The preferred answer will contain certain concepts that can be simple in form, such as if the answer contains a certain keyword or not, or it can be complex, such as showing a causal reasoning chain of information that shows the student can put together complex concepts in the correct order. If the student response lacks that particular structure, feedback from the teacher, or a software system, could provide meaningful information on what the student is lacking so that they can improve their analysis of the given task on their next try. Natural Language Processing (NLP) techniques such as sentence classification show promise in analyzing both simple and complex causal reasoning to help facilitate quicker feedback on student performance (Hastings et al., 2014).

Traditional NLP techniques include neural networks that are created for a specific task, then trained on large amounts of data in order to make the neural model accurate enough to provide meaningful feedback. In the past, Recurrent Neural Networks have been shown to be able to identify causal reasoning chains, but they require a large amount of data to train the model (Hughes, 2019). The model must be trained using a certain quantity of examples of each type of student response the grading rubric identified. This training data is typically evaluated by a human, ensuring the answer falls into a certain category. It is then labeled to establish if the answer is correct or falls into one or more sub-categories of partially correct or incorrect responses. In order for the model to be accurate enough to be useful in analyzing student responses, there must be a large enough quantity of these hand-evaluated student responses to make the model accurate enough for the given task. Unfortunately, educational studies do not necessarily provide a large amount of data to be used for NLP training — they collect only enough data to aid the researchers in their evaluation of the systems they are creating. Furthermore, those data sets are often focused on narrow, customized tasks so the language in them might not correspond to general usage.

BERT-based models changed the landscape of NLP data requirements because they come pre-trained using very large data sets such as Wikipedia and the BooksCorpus (Devlin et al., 2018). This gives them robust knowledge of the words in a language and how they are used *in context*. These models are then “fine-tuned” on a specific task — that is, an additional layer is added to the network, and the weights in this layer are trained to perform the chosen task. This fine-tuning requires some labeled training data from the task, but a much smaller data set is required because of the prior language “knowledge” already present in the model due to the initial pre-training. What’s more, pre-trained *BERT-based* models have been created in many languages, allowing for a wide range of applications. However, although the amount of data needed for fine-tuning is much less, these models still need a certain base level quantity of hand-evaluated training data for a particular task to enable the model to perform accurate evaluations.

A traditional method used to improve model performance on insufficient data in other areas of artificial intelligence, such as computer vision, is to augment the data set (Shorten and Khoshgoftaar, 2019). In this setting, data augmentation is done by scaling, flipping, or otherwise modifying the existing images, producing similar images that can make the models more robust. Augmentation techniques have been attempted on text-based data (Chen et al., 2021). However, augmenting text data is much more difficult than simply scaling or rotating an image. When new textual data is created from existing data (i.e., student responses), the newly generated responses are not guaranteed to be semantically similar to the original responses.

Some current augmentation techniques make modifications from original responses by misspelling words that have the effect of injecting noise, or by replacing words in the response with another similar word (Wu et al., 2022). These techniques attempt to create new data with a low probability of changing semantic intent. However, the possibility still exists that the newly generated response has been altered so much that the *label* associated with it, i.e., the category it is intended to augment, is no longer applicable. Using data augmentation on text in this fashion does not guarantee that the original sentence's semantic intent remains intact. As more aggressive forms of data augmentation are employed, for example, using a text generator that creates a response from keywords, more care must be exercised to ensure that semantic intent has not been significantly altered. Semantic similarity between responses can be measured to some extent but is still an ongoing area of research (Chandrasekaran and Mago, 2021).

In this paper, we examine a "self-augmentation" method, similar to stratified sampling (Neyman, 1992), where items from the data set are replicated to allow balancing of the different classes and provide a comparison for other augmentation types (Cochran et al., 2022). Bootstrapping (Xu et al., 2002) is another technique, however, it requires human intervention to fill in areas where the existing data is missing, which is not practical in our application of these techniques. This technique increases the training data set using the same data set provided for a given problem many times. Similar techniques have been applied to computer vision with improved model performance (Seo et al., 2021). Examining this technique will identify levels of augmentation that are appropriate and provide a baseline measurement that can be used for comparison when additional augmentation techniques are studied in the future. Performance is evaluated at each new level of augmentation to determine if it is improving the model or degrading it.

2 BACKGROUND AND RESEARCH QUESTIONS

Transformer-based NLP architectures, such as BERT (Devlin et al., 2018), and GPT-3 (Brown et al., 2020), are now the industry standard for modeling many NLP tasks. *BERT-based* models are plentiful and have been trained using many different corpora. Some models are trained on general corpora, and some on specific types of texts, e.g., medical articles. Some models have been trained on texts from a single language, and others on multilingual corpora. When selecting a model, at a minimum, one must choose a model which includes the words from the language of the target task. For instance, if you are evaluating French data, the original BERT model, which was pre-trained solely on English texts, would be useless because it has no training on the French language. So a language-specific model or a multilingual model is required. For the evaluations in this paper, multilingual models were chosen to compare performance on two data sets that were in different languages, French and English. Using the same set of models removes one variable in the comparison of the models and how they perform with each data set.

When comparing models using different data sets, another variable that might affect performance is the specific data the models were pre-trained on. The data sets in this research focus on topics in STEM- and argumentation-related fields. However, the domain-specific subject matter often includes esoteric jargon not well-represented in the canonical corpora that the large transformer models are pre-trained on. This can reduce performance because the task-based vocabularies differ considerably from the data the model was initially pre-trained on (Cohn, 2020).

Many educational study responses by school-age children use informal syntax and are written more colloquially. This makes it difficult to match a *BERT-based* model pre-trained with highly scientific data with a student's STEM topic because the model will be typically trained on advanced texts related to that particular topic. For this reason, we chose to use less specific models that were more widely used for sentence classification as opposed to models which were pre-trained solely on scientific texts. Using more generic models will be less likely to have performance issues due to a mismatch between the subject on which the students are being evaluated and the one on which the model was pre-trained.

Data scarcity is often encountered when evaluating school essays because such studies are typically focused narrowly on one aspect of learning or evaluation. This makes it difficult to train neural network models to accurately differentiate the intent of the student responses. To help with this problem, researchers performing data classification problems often transform the problem into a binary selection problem (true/false, right/wrong, answer-A/answer-B) as opposed to a multi-class problem (one involving three or more classes, e.g., answer-A/answer-B/answer-C) which provides an increase in performance at the cost of a more accurate conclusion. In this study, we use data sets that have both binary and multi-class solutions to see if augmentation levels are affected by the type of classification encountered.

As mentioned above, one frequent solution to mitigate data scarcity issues is *data augmentation*. Data sets encountered in student essays are not only small but are often imbalanced as well. Imbalanced data is defined as the situation where one class of answers is dominant for a given question, such as the case where most students get that particular question right or most get it wrong. It is rare that all possible labels corresponding to the conclusion of the evaluation are equally represented across the data set. Often, there may be only a few or even no cases where an example of one answer type is given in the data. This makes it difficult to train a model to look for a specific label if there are only a handful of examples in the data — there is simply not enough data for the model to learn from. To overcome the data scarcity and imbalance issues, larger, more balanced data sets are needed to improve model performance on the original small and imbalanced data sets. Since such situations are frequent in educational contexts, a method for augmenting small text-based data sets is needed to artificially create a larger training data set from a smaller one.

Textual data augmentation methods include such techniques as adding noise in the form of substitution or deletion of words or characters (Wei and Zou, 2019), performing “back translation” where data is translated into another language and then translated back to the original language, producing alternate responses based on the original response, and using BERT’s masking feature to substitute words. But each of these runs the risk of generating an alternate version of the text that is different in meaning and would therefore have a different label. In order to reduce or remove the chances of the generated responses having a label mismatch, self-augmentation is exclusively explored in this paper to work toward establishing a baseline for augmentation so that more radical or aggressive types of augmentation can be compared to this baseline.

In this paper, we examine the benefits of textual self-augmentation for evaluating student responses in two different educational contexts in two different languages. The data sets are relatively small (less than 100 responses per question or concept) and exhibit varying degrees of imbalance. Recent research on text augmentation found balancing the data set imperative when augmenting small data sets as leaving the data set unbalanced caused instability in model performance. (Cochran et al., 2022). Since the ultimate goal is to provide appropriate, timely feedback to students, teachers, and researchers, accurate models are required so that the feedback can be trusted. Toward that goal, we formulate three research questions:

RQ 1. Is self-augmentation of a balanced data set sufficient to improve the classification of student answers?[1] Hypothesis **H1** is that the technique of increasing the amount of data on a balanced data set will improve classification accuracy by ensuring at least a minimal number of training examples for each label exist.

RQ 2. Does the *BERT-based* model variant affect performance when the data sets are augmented and balanced? Our hypothesis **H2** is that the BERT model used for sentence classification will have some performance differences, but will not significantly change in performance when using data augmentation.

RQ 3. Does self-augmentation have a limit where performance drops, or does it stabilize with additional augmentation? **H3** proposes that self-augmentation would benefit the model performance at first but then would tend to overfit because it is being constantly trained on the same information, and performance would degrade with additional augmentation beyond that point. Previous work has shown similar results.

3 DATA SETS

The two data sets analyzed in this paper were provided from different studies and are referred to herein as Task 1 and Task 2. Task 1 was done as part of the SPICE (Science Projects Integrating Computation

[1] How good is good enough? In this paper, we are arbitrarily choosing an F_1 score of 0.9 or above, under the assumption that that would allow a system to provide feedback that is *usually* correct. Appropriate language can be used to indicate some level of uncertainty in the evaluation, e.g., “It looks like . . . ”.

and Engineering). This data set contained responses to three questions. These were further divided into six different concepts because some of the questions requested multiple answers. The data originates from 6th-grade students and is in English and contained 95 responses by students in the United States. The average length of the student responses was about 15 words. This research curriculum focused on rain water runoff and measured if the student could formulate a mental model of how runoff worked based on the responses given. Each of the six concepts for this task has a binary label indicating if the concept was correct or incorrect in the response. The responses to the six concepts were imbalanced to varying degrees with the data label quantities shown in Table 1. The majority label quantity is shown in bold.

Table 1: Task 1 Student Response Data Split per Question.

Concept	Incorrect	Correct
1	10	**85**
2a	25	**70**
2b	**64**	31
3a	44	**51**
3b	**73**	22
3c	**57**	38

In Task 2, French university students were given a social psychology article describing links between personal aggression and the playing of violent video games. The participants were asked to read the article and then write a message to friends (in the "personal" condition) or colleagues (in the "academic" condition)[2] which summarized the connections made in the article. We evaluated one question of interest to the researchers here, which is whether or not the student's text expressed an opinion about the presence of a link between aggression and violent video games.

This data set contained 40 responses with an average length of about 90 words. The data label quantities are shown in Table 2. Since there was only one question, the possible outcomes are listed with their label, description, and quantities for each label in the data set. The majority label quantity, which is "No Opinion", is shown in bold.

Table 2: Task 2 Student Response Data Split for the One Question.

Label	Definition	Quantity
0	No Opinion	**32**
1	Link Exists	3
2	No Link Exists	2
4	Partial Link Exists	3

4 METHODS

4.1 Augmentation

The data sets were augmented by replicating the provided data to create an augmented data pool. Previous studies have indicated that a balanced data set is recommended prior to data augmentation (Cochran et al., 2022). Therefore, the data sets were first balanced and evaluated. In order to balance the data set, the lowest minority label quantity was used for each label in that particular data set. We define this as an augmentation amount of 0x. For example, in task 1, concept 1, there were 10 incorrect answers. To obtain the augmentation amount of 0x and maintain a balanced data set, the correct answers were reduced to match the level of the incorrect numbers. So, 75 of the 85 correct responses were removed so the data set only contained 10 correct and 10 incorrect student responses.

The amount of data used for augmentation above 0x was determined by measuring of the majority label quantity in each data set, and augmenting the minority labels enough to match the majority label. This created a 1x level of data augmentation. In a multi-class data set, each label that was not the majority label was augmented to equal the quantity of the majority label. Performance was measured as the augmentation level was increased up to 34x by adding more data from the augmented pool and balancing the training data set so that all labels had equal representation.

Augmented data was only used for training, not for testing the models. For each question or concept, a separate BERT model was fine-tuned for classification on the training data by adding a single feed-forward layer. We used the micro-F_1 metric as the performance measurement.

In all experiments, the models were trained and evaluated 10 times, with each training iteration using a different seed (a Fibonacci series starting at 5) for the random number generator, which partitions the training and testing instances. During training, the train/test split was 80/20. Batch size, learning rate, and other parameters were tuned to provide the best performance for the given data set but fit within the recommendations of the original BERT fine-tuning suggestions (Devlin et al., 2018).

4.2 Models

Since these two data sets are in different languages, a fair comparison between them was established by using multi-lingual *BERT-based* models. Four multi-lingual *BERT-based* models were chosen that can perform sentence classification. These four models were

[2] We do not further examine the differences between the conditions here.

fine-tuned separately for each data set using the methods described by (Devlin et al., 2018) in order to provide a comparison of the self-augmentation method. For Task 1, six separate models were created for each of the 6 concepts using binary classification: either right or wrong. For Task 2, a single multi-class model was trained. Since both tasks together needed seven models, and there were four base models used in testing, this resulted in 28 models being trained and tested.

The *BERT-based* models used in this research are all from the Hugging Face library[3], and are shown in Table 3. They were chosen because they could perform both French and English sentence classification. Each model was pre-trained for a specific purpose. The ambeRoad (model "A") is trained using the Microsoft MS Marco corpus. This training data set contains approximately 400M tuples of a query, relevant and non-relevant passages. Model A has been used in various works (Schumann et al., 2022; Litschko et al., 2022; Vikraman, 2022) and is owned by amberSearch (previously ambeRoad) — a private, German NLP enterprise. The cross-encoder (model "C") was pre-trained on the MMARCO corpus which is a machine-translated version of MS MARCO using Google Translate to translate it to 14 languages (Reimers and Gurevych, 2020). The Microsoft (model "M") generalizes deep self-attention distillation by using self-attention relation distillation for task-agnostic compression of pre-trained Transformers (Wang et al., 2020). Model M has also been used in various works (Verma et al., 2022; Nguyen et al., 2022). Finally, the nlptown (model "N") is a multilingual uncased model fine-tuned for sentiment analysis on product reviews in six languages: English, Dutch, German, French, Spanish, and Italian. It predicts the review's sentiment as a number of stars (between 1 and 5) and is intended for direct use as a sentiment analysis model for product reviews. Model N is owned by a private enterprise, NLP town, and is referenced in several works (Şaşmaz and Tek, 2021; Nugumanova et al., 2022; Casañ et al., 2022). Models were prioritized based on their prevalence in research and widespread adoption on HuggingFace.

4.3 Baseline Evaluation

For each concept or question, we created two different baseline models without augmentation. The *a priori* model simply chose the majority classification for each concept. In other words, this model is purely statistical — there is no machine learning involved. For our *unaugmented* baseline model, we applied BERT in a prototypical way, without data augmentation, but the data set was balanced in that each label had equal representation. We performed this by getting the count of the least represented label and reducing the quantity of the other labels in the data set to match that label, previously defined as 0x augmentation. This removed the possibility of data imbalance affecting the baseline measurement.

4.4 Augmentation Approach

The self-augmentation technique is an oversampling method using multiple copies of each instance in the data set for augmentation. To augment the data consistently, the majority label quantity in Table 4 and Table 5 became the **majority quantity of reference** for that particular data set. For example, in Table 4, concept C1 has a majority of the answers as correct (89%). Since responses were graded as either right or wrong, this means there were 85 correct answers and 10 incorrect answers. Therefore, the majority quantity of reference for concept C1 is 85. Augmenting the incorrect answers so that the quantity equals 85 (adding 75 incorrect responses) balanced the data set is referred to as 1x augmentation. We balanced each data set in this manner by augmenting the minority label(s) to equal the quantity of the majority label for 1x augmentation. The data sets maintained an equal balance across all possible values and were augmented in multiples of the majority quantity using multiples 3x, 5x, 8x, 13x, 21x, and 34x.

5 RESULTS

The results from Task 1 are presented in Table 4, and the results from Task 2 are in Table 5. Each row corresponds to a question or concept. The baseline results are shown in the middle columns. The maximum micro-F_1 for each model along with the data augmentation level where that maximum occurred is shown in the right columns. Bolded values show the maximum performance for each question/concept for each data set.

Figure 1 illustrates how the level of data augmentation affects performance for Task 1. Each line shows the performance with a different model. The x-axis shows the amount of augmentation applied from 0x to 34x. The models corresponding to the codes in the legend are listed in Table 3. As augmentation increases, there is a maximum performance achieved at different augmentation levels which depended on not only the concept but was also a function of the base model used. Each of the models appeared to peak

[3] https://huggingface.co/models

Table 3: *BERT-based* models used in this research.

Letter Designation	Hugging Face Model
C	cross-encoder/mmarco-mMiniLMv2-L12-H384-v1
M	microsoft/Multilingual-MiniLM-L12-H384
A	amberoad/bert-multilingual-passage-reranking-msmarco
N	nlptown/bert-base-multilingual-uncased-sentiment

Table 4: Task 1 Performance (micro-F_1) of Baseline vs All Augmented Models.

	%	Baseline		Max Performance		
Concept	Correct	*a priori*	Unaug.	F_1	Aug. Level	Model
C1	89	**0.940**	0.563	0.837	3x	N
C2a	73	0.850	0.795	**0.995**	13x	N
C2b	33	0.670	0.658	**0.921**	3x	N
C3a	54	0.700	0.779	**0.789**	3x	N
C3b	23	0.770	0.784	**0.968**	5x	N
C3c	40	0.600	0.784	**0.889**	13x	N

early, then drop off. For Task 1, the NLPTown model seemed to perform well across all questions and concepts whereas the Microsoft model performed poorly.

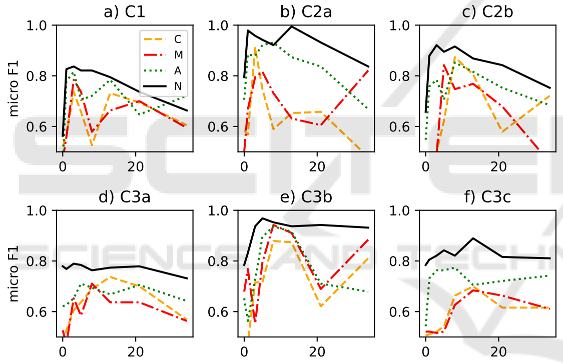

Figure 1: Task 1 Model Performance as a Function of Augmentation Level for All Models.

Figure 2 illustrates how the level of data augmentation affects performance for Task 2. In this task, where the student texts were considerably longer than in Task 1, (recall that the average length of student responses was 90 words for Task 2, but only 15 words for Task 1) the performance of most models continued to increase or only diminished slightly with increasing levels of augmentation. For Task 2, the Microsoft model appeared to outperform all other models, and the NLPTown model had the worst performance.

6 DISCUSSION

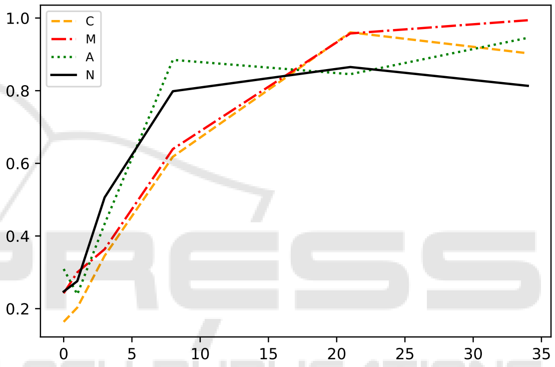

Figure 2: Task 2 Model Performance as a Function of Augmentation Level for All Models. The *x*-axis shows the amount of augmentation applied from 0x to 34x.

Recall **RQ 1** which asked whether the performance could be improved with an augmented and balanced data set. Table 4 and Table 5 show that data augmentation does improve classification performance in all cases except for on Task 1's Concept C1 which is so unbalanced that the best performer is the *a priori* baseline of simply guessing the student response was always correct. All the other results show improvement over every model's unaugmented baseline. Model performance where the micro-F_1 score is above 0.9 is ideal, but there are two instances where this level of performance could not be achieved. This indicates that self-augmentation may not be enough to fine-tune a model for all contexts, and additional augmentation techniques will be required in order to generalize a procedure for textual data augmentation. **H1** states that the effect of a balanced data set and larger amounts of data improve classification accuracy. Our results show that balancing and augmentation improves performance. In Task 1, Concept 1, there was such a high percentage of majority label quantities (>90%), that guessing the majority label every time outperformed any model used in our ex-

Table 5: Task 2 Performance (micro-F_1) Baseline vs All Augmented Models.

	%	Baseline		Max Performance		
Question	No Opinion	*a priori*	Unaug.	F_1	Aug. Level	Model
1	80	0.850	0.309	**0.994**	34x	M

periment. Therefore, **H1** was almost completely confirmed. The only exception was for concept 1, where the majority label represented over 90% of the given responses.

RQ 2 asks if the chosen BERT model affects performance when the data sets are augmented and balanced. **H2** predicted that the BERT model used would not significantly affect performance. Figures 1 and 2 reveal that performance does vary based on the BERT model. These empirical observations show that the N model (nlptown/bert-base-multilingual-uncased-sentiment) performed the best for Task 1, and the M Model (Microsoft) performed the best for Task 2. No clear winner was evident in which model to choose because the winning model in each study was the worst performer in the other study. Thus, **H2** was not supported, which means the model chosen for the particular task is important, and experimenting with more than one model is imperative for a given data set. Several factors were different across the two tasks. The languages were different, the length of the responses was different, and the age of the students was different in each data set, all potentially causing performance changes between the two tasks.

RQ 3 speculated that the level of augmentation might either drop off or stabilize model performance. Model performance did reach a maximum, then drop off with additional augmentation for Task 1. This shows that there is a limit to how much self-augmentation can be applied before the model levels off and begins to degrade in performance. This degradation in performance is likely due to the over-fitting of the model as the data used to train the model is being repeated, and the model is not generalizing, but specializing classification outcomes that are specific to the given data set. Task 2 had two models continue to increase even at the highest level of augmentation tested, but the results were approaching a micro-F_1 of 1.0, so additional augmentation could not possibly improve performance beyond that limit. For these results, **H3** was mostly supported with the exception of two models in Task 2 that continued to increase at maximum augmentation.

7 CONCLUSION

Using self-augmentation to increase the data quantity from student responses in two different studies, in two different languages (French and English), then training seven different models and measuring performance, we found that a balanced and augmented data set improved performance over an unaugmented data set. Binary classification peaked in performance with less augmentation but degraded as the model began to overfit. The highest level of augmentation for a binary classification task was 13x. Multi-class classification on the other hand was able to handle higher levels of self-augmentation, making all models tested stabilize and achieve high performance over the unaugmented model. Multi-class classification models in our experiment did not degrade significantly as we continued to add augmentation up to 34 times the initial majority label quantities.

8 FUTURE WORK

Since the ideal performance was not achieved by self-augmentation alone in all cases, further analysis will be performed using additional augmentation methods. This might allow sentence similarity variance so the models can tolerate additional augmentation before overfitting. Our hypothesis is that this will aid the model training so that it does not begin to degrade so quickly, and a micro-F_1 performance of greater than 0.9 can be consistently achieved. More experimentation will be performed comparing binary and multi-class classification to see if there are general guidelines that can be established for both types of classification. This may also determine if augmentation levels tolerated by a model are a function of the number of classes in the data set. A comparison of language-specific models versus multi-lingual models needs to be performed to determine if performance is affected by matching language type to the specific model or if a more generic model suffices.

REFERENCES

Achieve, Inc (2013). Next Generation Science Standards.

Britt, M. A., Rouet, J. F., and Durik, A. M. (2017). *Literacy Beyond Text Comprehension: A Theory of Purposeful Reading*. Routledge, New York.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Casañ, R. R., García-Vidal, E., Grimaldi, D., Carrasco-Farré, C., Vaquer-Estalrich, F., and Vila-Francés, J. (2022). Online polarization and cross-fertilization in multi-cleavage societies: the case of spain. *Social Network Analysis and Mining*, 12(1):1–17.

Chandrasekaran, D. and Mago, V. (2021). Evolution of semantic similarity—a survey. *ACM Computing Surveys (CSUR)*, 54(2):1–37.

Chen, J., Tam, D., Raffel, C., Bansal, M., and Yang, D. (2021). An empirical survey of data augmentation for limited data learning in nlp. *arXiv preprint arXiv:2106.07499*.

Cochran, K., Cohn, C., Hutchins, N., Biswas, G., and Hastings, P. (2022). Improving automated evaluation of formative assessments with text data augmentation. In *International Conference on Artificial Intelligence in Education*, pages 390–401. Springer.

Cohn, C. (2020). *BERT Efficacy on Scientific and Medical Datasets: A Systematic Literature Review*. DePaul University.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Hastings, P., Hughes, S., Britt, A., Blaum, D., and Wallace, P. (2014). Toward automatic inference of causal structure in student essays. In *International Conference on Intelligent Tutoring Systems*, pages 266–271. Springer.

Hughes, S. (2019). *Automatic Inference of Causal Reasoning Chains from Student Essays*. PhD thesis, DePaul University, Chicago.

Litschko, R., Vulić, I., Ponzetto, S. P., and Glavaš, G. (2022). On cross-lingual retrieval with multilingual text encoders. *Information Retrieval Journal*, 25(2):149–183.

Neyman, J. (1992). *On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection*. Springer.

Nguyen, K. H., Dinh, D. C., Le, H. T.-T., and Dinh, D. (2022). English-vietnamese cross-lingual semantic textual similarity using sentence transformer model. In *2022 14th International Conference on Knowledge and Systems Engineering (KSE)*, pages 1–5. IEEE.

Nugumanova, A., Baiburin, Y., and Alimzhanov, Y. (2022). Sentiment analysis of reviews in kazakh with transfer learning techniques. In *2022 International Conference on Smart Information Systems and Technologies (SIST)*, pages 1–6. IEEE.

OECD (2021). *21st-Century Readers*. PISA, OECD Publishing. https://www.oecd-ilibrary.org/content/publication/a83d84cb-en.

Reimers, N. and Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Şaşmaz, E. and Tek, F. B. (2021). Tweet sentiment analysis for cryptocurrencies. In *2021 6th International Conference on Computer Science and Engineering (UBMK)*, pages 613–618. IEEE.

Schumann, G., Meyer, K., and Gomez, J. M. (2022). Query-based retrieval of german regulatory documents for internal auditing purposes. In *2022 5th International Conference on Data Science and Information Technology (DSIT)*, pages 01–10. IEEE.

Seo, J.-W., Jung, H.-G., and Lee, S.-W. (2021). Self-augmentation: Generalizing deep networks to unseen classes for few-shot learning. *Neural Networks*, 138:140–149.

Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48.

Verma, A., Walbe, S., Wani, I., Wankhede, R., Thakare, R., and Patankar, S. (2022). Sentiment analysis using transformer based pre-trained models for the hindi language. In *2022 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, pages 1–6. IEEE.

Vikraman, L. N. (2022). Answer similarity grouping and diversification in question answering systems. *Doctoral Dissertation*.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., and Zhou, M. (2020). Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers.

Wei, J. and Zou, K. (2019). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Wu, L., Xie, P., Zhou, J., Zhang, M., Ma, C., Xu, G., and Zhang, M. (2022). Self-augmentation for named entity recognition with meta reweighting. *arXiv preprint arXiv:2204.11406*.

Xu, F., Kurz, D., Piskorski, J., and Schmeier, S. (2002). A domain adaptive approach to automatic acquisition of domain relevant terms and their relations with bootstrapping. In *LREC*.

Zhang, N., Biswas, G., McElhaney, K. W., Basu, S., McBride, E., and Chiu, J. L. (2020). Studying the interactions between science, engineering, and computational thinking in a learning-by-modeling environment. In *International Conference on Artificial Intelligence in Education*, pages 598–609. Springer.