Abstract—Transfer learning is proving quite useful in Natural Language Processing. One of the most important problems in Natural Language Processing is Automated essay scoring, which remains partially unsolved especially when we are dealing with single language model capable to evaluate essays of multiple topics. In this work we examine the effectiveness of transformer models like BERT, RoBERTa, etc using data augmentation technique to build a model capable of evaluating essays of multiple topics having little data.

Index Terms—Transfer Learning, Transformers, Automated Scoring

I. INTRODUCTION

During COVID-19 pandemic, online schooling system became necessary. From elementary schools to colleges, almost all educational institutions have adopted the online education system. The majority of automated evaluations are accessible for multiple-choice questions, but evaluating short and essay-type responses remains unresolved since, unlike multiple-choice questions, there is no one correct answer for these kinds of questions. It is an essential education-related application that employs NLP and machine learning methodologies. It is difficult to evaluate essays using basic computer languages and methods such as pattern matching and language processing.

Automated Essay Scoring (AES) is the method of automatically grading short answer and essay questions without human intervention. Initiated by Page’s [1966] groundbreaking work on the Project Essay Grader system, this area of study has seen continuous activity ever since. The bulk of AES research has been on holistic scoring, which provides a quantitative summary of an essay’s quality in a single number. At least two factors contribute to this concentration of effort. To begin with, learning-based holistic scoring systems may make use of publicly accessible corpora that have been manually annotated with holistic scores. Second, there is a market for holistic scoring algorithms because they may streamline the arduous process of manually evaluating the millions of essays for tests like GRE, IELTS, SAT.

Past research on automated essay grading has included training models for essays for which training data is available and those models are topic specific. This model is trained on all the topics thus could be used for assessment of essays of all those topics without training model specific for each topic. This would be useful in the scenario where we did not have enough data to train a model that is specific to a particular topic, but we still needed to evaluate essays on that topic. Therefore, in order to assess them, We utilize a model that has been trained on essays on a variety of topics and a tiny amount of data on the topics for which we need to develop a model, which will then be fine-tuned using the limited data available on the topics being assessed.

This paper is organized as follows: In Section II, we explore pertinent prior research on automated essay scoring; in Section III, we cover experimental setup; and in Section IV, we describe our methodology for augmenting essay data. In Section V, we give the results and analysis of the automated essay grading model. Section VI comprises of conclusion and future work for Automated Essay Scoring.

II. RELATED WORKS

The Project Essay Grader (PEG) research on Automatic essay scoring was initiated by [1]. Shermis (2001) [2] improved the PEG system by including grammatical characteristics in the evaluation. Around the turn of the century, the vast majority of essay scoring systems relied on conventional techniques, such as latent semantic analysis by Foltz (1999) [3], pattern matching, and statistical analysis, such as Bayesian Essay Test Scoring System by [4]. In order to determine an essay’s grade, these systems use natural language processing (NLP) techniques that emphasize grammar and content.

Multiple studies studied AES systems, from the earliest to the most recent. Blood (2011) [6] reviewed the PEG literature from 1984 to 2010, it has discussed just broad features of AES systems, such as ethical considerations and system performance. However, they have not addressed the implementation aspect, nor has a comparison research been conducted, nor have the real problems of AES systems been highlighted.

Automatic scoring systems developed after 2014, such as those by [5] and others, utilized deep learning techniques to generate syntactic and semantic traits, producing better results compared to prior systems. Burrows (2015) [7] examined various facets, such as datasets, NLP techniques, model construction, model grading, model evaluation, and model efficacy. Ke (2019) [8], Hussein (2019) [9] and Klebanov (2020) [10] provided an overview of the AES system.
Ramesh (2019) [19] offered us a comprehensive summary of all accessible datasets and the machine learning algorithms used to grade essays. Recently Park (2022) [23] tried using GAN’s for evaluation of essays. Recent study by Elijah (2020) [20] and Wang (2022) [25], it showed us that using BERT based models it certainly improves the accuracy of the Automated essay scoring which motivated us to use Deep Learning models for our task.

Jong (2022) [24] demonstrated the effectiveness of data augmentation techniques on automated essay scoring. Ludwig (2021) [21], Ormerod (2021) [22] and Sethi (2022) [26] examined the transformer based models on automated essay scoring, these recent findings prompted us to test our data augmentation strategy on transformer-based models for automated essay grading.

III. EXPERIMENTAL SETUP

We used Automated Student Assessment Prize (ASAP1) dataset for experimentation purposes. This corpus was released as part of a Kaggle competition in 2012. This corpus is used for comprehensive scoring of essays and consists of around 2000 human-evaluated essays from eight different topics. Each topic had a different scoring method, so, we normalized each essay score from 0 to 10 so that we could train all the data together.

We run our experiments using the BERT, RoBERTa, ALBERT, DistilBERT, XLM-RoBERTa2, implementation available in the Simple Transformers library3. We run our experiment on Google Colab. For baseline trials, we consider the essays which are not supplemented with our data augmentation technique and trained on models based on Transformers.

During training, we additionally fine-tuned our parameters by adjusting the parameters like learning rate, weight decay rate, etc. For the purpose of evaluating our model, we will use the accuracy score measure given by the Scikit library4. Since we are approaching our issue as a multi-label classification, we used accuracy metric for evaluations.

IV. METHODOLOGY

A. Large Pre-trained models

The models which we used for training are based on the Transformers(Fig. 1) architecture introduced by [11] which uses an encoder-decoder based architecture.

Given below is the brief description of each of the model which we are using for training purposes:

1) BERT: The Bidirectional Encoder Representations (BERT) introduced by [12] is a model in which every output element is linked to every input element and the weights between them are dynamically calculated based on their relationship. It is pre-trained on two tasks: Masked Language Modeling(MLM) and Next Sentence Prediction(NSP).

2) RoBERTa: Robustly optimized BERT Pre-training Approach (RoBERTa) introduced by [13] builds upon BERT’s MLM task, in which the system learns to anticipate purposely masked bits of text inside unannotated language samples. As a result, RoBERTa is able to get better results than BERT on the masked language modeling task.

3) ALBERT: ALBERT, introduced by [14] stands for “A Lite BERT”, an encoder-decoder based model with self-attention at the encoder and attention to encoder outputs at the decoder end. It is based on techniques like parameter sharing, embedding factorization, and sentence order prediction(SOP).

4) DistilBERT: DistilBERT, introduced by [15] aims to optimize the training by reducing the size of BERT and increasing the speed of BERT—all while trying to retain as much performance as possible. Specifically, DistilBERT is smaller than the original BERT-base model, is faster than it, and retains its functionality.

5) XLM-RoBERTa: XLM-RoBERTa, Unsupervised Cross-lingual Representation Learning at Scale, introduced by [16] is a scaled cross-lingual sentence encoder. It is trained on 2.5 GB of data extracted from Common Crawl more than 100 languages. XLM-RoBERTa achieves state-of-the-art results on multiple cross-lingual benchmarks.

B. Data Augmentation

Researchers have attempted to use several RNNs and LSTMs as training models for automated essay scoring. However, the fundamental disadvantage of such models is that they are topic-specific, and we want to construct an automated essay scoring system that can perform well not just on subjects for which we have an abundance of data but also on subjects for which we have a limited amount of data. Now, we want to augment the essay so that it can accurately assess a essay on a different topic for which we have a very small amount of data.

When training a model, we add each essay with its topic at certain intervals. We are including essay topics after an interval

---

1https://www.kaggle.com/c/asap-aes
2we used base models for our experiment
3https://simpletransformers.ai/
4https://scikit-learn.org/stable/index.html
because essays are lengthy, and if we train only on essays and exclude its topic, the trained model will become topic-specific. In order to construct a more robust model, we append the essay’s subject to each essay so that the training model will learn the relationship between the essay’s topic and the essay itself. This is done so that it accurately evaluates essays on a different topic, which are used majorly for fine-tuning since for low-resource topics, it becomes extremely difficult to build topic-specific models for its evaluation.

However, it is not possible to add a complete topic to an essay because sometimes essay topics are quite large and include that would reduce accuracy. Hence, we summarize the topics of essays using the summarization pipeline provided by [17] implementation of BART, which was introduced by [18]. Now, the second issue is, after how many lines should we put the subject for optimum precision? We conducted extensive data trials by inserting them at different places.

After a comprehensive investigation of several transformer-based models and essay subject insertions, we determined that the tenth place is optimal for inserting the topic. It implies that after every tenth line, a summary of the subject is added to the essay. It may be due to the fact that the model struggles with lengthy text classification; thus, we keep this in mind by inserting topics at regular intervals. Fig. 2 shows the complete data augmentation procedure of essay. Now that the summary of the topic has been inserted into the essay, the data is modified for training transformer models. The next section provides discusses the results of our experiment.

V. RESULTS AND ANALYSIS

The ASAP1 dataset contains around 17K essays on eight topics. We are using that data for pre-training by augmenting those essays using our technique. For fine tuning For research and testing purposes, we used this dataset. It consists of 1241 essays on four topics. We fine-tuned and tested our models on each topic individually. We used around two-thirds of the above mentioned dataset for training along with ASAP1 and fine-tuning, and the remaining one-third for testing our model.

We followed a very simple yet state-of-the-art modeling technique for multi-label classification using transformer models. We bucketed scores into each interval class, resulting in 11 buckets. These 11 classes correspond to a score from 0 to 10. Our methodology assigns each essay to a particular category. If an essay is categorized by my model as being in Bucket 6, then it receives a score of 5.

We approached this in the same manner as sentiment analysis, which utilizes classification algorithms and yields extremely positive results. As shown in Fig. 3, our data augmentation training method improves accuracy of our model. It is quite evident from that BERT and RoBERTa outperform other models, although by a small margin. It might be due to the fact they are trained for more number of parameters than rest of the models. The analysis of these findings demonstrates that utilizing Transformer-based models are more efficient than traditional RNN based models like LSTM.

When used on top of pre-trained classification models, our data augmentation technique significantly improves the performance of automated essay scoring. The accuracy of all those pre-trained models increases after applying our augmentation technique. We believe this performance is because after mixing a summary of the topic with each essay, it encourages the internal representation of each essay to align with the topic, so that when we test it on a essay with a different topic after fine tuning, it checks for the alignment between the topic and the essay as a result of training and fine-tuning and grades it accordingly.

VI. CONCLUSION AND FUTURE WORK

In this paper, we automated essay scoring using transformer based models and an augmentation approach. The conclusion
are as follows:

- Pre-trained transformer-based models BERT, RoBERTa, ALBERT, DistilBERT, and XLM-RoBERTa are very efficient for Automated Essay Scoring.
- Data augmentation approaches further enhance its performance for analyzing long texts like essays to attain higher accuracy score.

In the future, we will come up with a better way to include elements that are relevant to the topic instead of just a summary, so that training with this data will lead to a more accurate model for our task.

REFERENCES

[1] Ajay HB, Tillett PI, Page EB (1973) Analysis of essays by computer (AEC-II) (No. 8-0102). Washington, DC: U.S. Department of Health, Education, and Welfare, Office of Education, National Center for Educational Research and Development.

[2] Shermis MD, Mzumara HR, Olson J, Harrington S (2001) On-line grading of student essays: PEG goes on the World Wide Web. Assess Eval High Educ 26(3):247–259.

[3] Foltz PW, Laham D, Landauer TK (1999) The Intelligent Essay Assessor: Applications to Educational Technology. Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 1, 2, http://imej.wfu.edu/articles/1999/2/04/index.asp.

[4] Rudner, L. M., Liang, T. (2002). Automated essay scoring using Bayes’ theorem. The Journal of Technology, Learning and Assessment, 1(2).

[5] Dong F, Zhang Y, Yang J (2017) Attention-based recurrent convolutional neural network for automatic essay scoring. In: Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017) p 153–162.

[6] Blood, I. (2011). Automated essay scoring: a literature review. Studies in Applied Linguistics and TESOL, 11(2).

[7] Burrows S, Gurevych I, Stein B (2015) The eras and trends of automatic short answer grading. Int J Artif Intell Educ 25:60–117. https://doi.org/10.1007/s40593-014-0026-8.

[8] Ke Z, Ng V (2019) “Automated essay scoring: a survey of the state of the art.” IJCAI.

[9] Hussein, M. A., Hassan, H., Nassef, M. (2019). Automated language essay scoring systems: A literature review. PeerJ Computer Science, 5, e208.

[10] Klebanov, B. B., Madnani, N. (2020, July). Automated evaluation of writing–50 years and counting. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 7796–7810).

[11] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Minnesota. Association for Computational Linguistics.

[13] Liu, Yinhan, Ott, Myle, Goyal, Naman, Du, Jingfei, Joshi, Mandar, Chen, Danqi, Levy, Omer, Lewis, Mike, Zettlemoyer, Luke and Stoy-
anov, Veselin RoBERTa: A Robustly Optimized BERT Pretraining Approach. (2019). cite arxiv:1907.11692.

[14] Lan, Zhenzhong, Chen, Mingda, Goodman, Sebastian, Gimpel, Kevin, Sharma, Piyush and Soricut, Radu. "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations." Paper presented at the meeting of the ICLR, 2020.

[15] Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." arXiv preprint arXiv:1910.01108 (2019).

[16] Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. "Unsupervised cross-lingual representation learning at scale." arXiv preprint arXiv:1911.02116 (2019).

[17] Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac et al. "Huggingface’s transformers: State-of-the-art natural language processing." arXiv preprint arXiv:1910.03771 (2019)

[18] Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension." arXiv preprint arXiv:1910.13461 (2019).

[19] Ramesh, Dadi, and Suresh Kumar Sanampudi. "An automated essay scoring systems: a systematic literature review." Artificial Intelligence Review (2021): 1-33.

[20] Elijah Mayfield and Alan W Black. 2020. Should You Fine-Tune BERT for Automated Essay Scoring?. In Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 151–162, Seattle, WA, USA - Online. Association for Computational Linguistics.

[21] Ludwig, Sabrina, et al. "Automated essay scoring using transformer models." Psych 3.4 (2021): 897-915.

[22] Ormerod, Christopher M., Akanksha Malhotra, and Amir Jafari. "Automated essay scoring using efficient transformer-based language models." arXiv preprint arXiv:2102.13136 (2021).

[23] Park, Yo-Han, et al. "EssayGAN: Essay Data Augmentation Based on Generative Adversarial Networks for Automated Essay Scoring." Applied Sciences 12.12 (2022): 5803.

[24] Jong, You-Jin, Yong-Jin Kim, and Ok-Chol Ri. "Improving Performance of Automated Essay Scoring by using back-translation essays and adjusted scores." Mathematical Problems in Engineering 2022 (2022).

[25] Wang, Yongjie, et al. "On the Use of BERT for Automated Essay Scoring: Joint Learning of Multi-Scale Essay Representation." arXiv preprint arXiv:2205.03835 (2022).

[26] Sethi, Angad, and Kavinder Singh. "Natural Language Processing based Automated Essay Scoring with Parameter-Efficient Transformer Approach." 2022 6th International Conference on Computing Methodologies and Communication (ICCMC). IEEE, 2022.