2024 IEEE International Conference on Knowledge Graph (ICKG) | 979-8-3315-0882-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICKG63256.2024.00032

Language Modeling for Sentence Level Assessment: A Case Study of First-Year English Composition

Rishabh Lingam†, Sipai Klein‡, Wendy W. Hinshaw‡, and Xingquan Zhu†
†*College of Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL-33431, USA*
‡*Dorothy F. Schmidt College of Arts and Letters, Florida Atlantic University, Boca Raton, FL-33431, USA*
{rlingam2023, kleins, whinshaw, xzhu3}@fau.edu

***Abstract*—Recent advancement in neural language models has witnessed many applications in writing instruction and assessment, including automated writing assessment in secondary and post-secondary education. Using language models for student writing assessment brings promises of increased accuracy, speed, and objectivity in evaluating student writing, benefiting writing students and teachers. Based on our case study using sentence-level prediction in Writing Across the Curriculum (WAC) assessment of College Writing 1 and 2 courses at the Florida Atlantic University (FAU), we argue in this paper that automatic writing assessment should be implemented at the sentence level. We collected end-of-semester argumentative essays, which were then systematically evaluated at the sentence- and the whole-document level. The study shows how neural language models can achieve a certain level of accuracy, but scoring inconsistency is a significant challenge to the learning model. We propose solutions to address such inconsistency, and also report important findings from the case study.**

***Index Terms*—Transformers, BERT, Automatic Essay Scoring, English Composition, Computational Linguistics.**

I. Introduction

Writing Across the Curriculum (WAC) programs are used to oversee curriculum and instruction in higher education institutions whose objective it is to employ writing as a central method for developing critical thinking skills, learning discipline-specific content, and understanding and building competences in the modes of inquiry and writing for various disciplines and professions [1]. As is often the case, WAC programs create assessment processes that reflect the assessment needs of institutions as well as state agencies that govern them [2]. Such writing instruction occurs both within English departments and in disciplines where writing is adopted as a central instructional strategy. For example, an undergraduate Engineering WAC-designated course may ask students to explain technical knowledge through writing assignments and to draft and revise reports for experiments.

At Florida Atlantic University, the WAC program annually assesses student writing from a sample of the writing-intensive courses in order to evaluate program effectiveness and determine whether courses meet state-mandated General Education Curriculum outcomes for written communication, as shown in Fig. 1. Student writing is evaluated based on an 11-category rubric, defined in Table I, that examines argument-driven student writing according to outcomes aligned with the Conference on College Composition and Communication's Principles for the Postsecondary Teaching of Writing position statement [10]. This annual WAC assessment provides departments with course-level data on student performance composing thesis-driven essays. Because writing assessments are time and labor intensive they can only occur annually, and so assessment results are provided long after writing assignments have been submitted, and ultimately rely on departments to interpret the assessment data and implement changes to curriculum or teaching practices in response. While this annual assessment process produces a tremendous body of data about how student writers perform according to its custom rubric, until now, the WAC program has not been able to utilize the data to provide feedback directly to students during their writing process. In other words, while departments benefit from data on student writing performance that they can interpret and apply at the level of instruction and course design, the student writers themselves lack input on WAC-specific writing criteria during their drafting process.

A. AES and AWE Literature Review

As artificial intelligence and literacy researcher McNamara [12] argues, AI research lacks large data sets by which to investigate key questions about the role of AI in education. This challenge complicates prior research on automated writing evaluations (AWE) and automated essay scoring (AES) systems which has demonstrated limitations in these systems' ability to provide detailed feedback and attend to higher-order writing concerns [13]. However, emerging research on AI-enabled writing tools suggests such tools can aid writing teachers and learners in metalinguistic awareness and development of writing skills ([13], [7], [17]). In addition, research into AI-driven assessment and intelligent tutoring has shown potential in AWE systems adopting NLP tools to improve students' persuasive writing skills [15]. Neural language modeling, such as BERT and ChatGPT, has the capacity to extend research into Automated Essay Scoring (AES) because it has the potential to identify linguistic patterns, be trained on human rating practices to generalize probabilistic assessments, and generate responses to prompt writing revision [18].

Prior AWE and AES research does not use argumentative, thesis-driven, reading-centered, long-form student essays as the corpus, leading to a substantial limitation in the training data; for example, AES scoring based on data sets of tenth grade ([14]) or seventh grade essays with an average length

979-8-3315-0882-1/24/$31.00 ©2024 IEEE
DOI 10.1109/ICKG63256.2024.00032

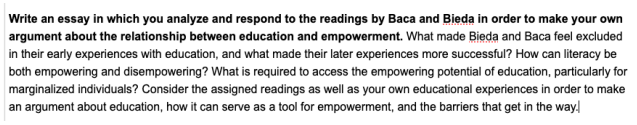

Fig. 1. Example of an ENC 1101 Essay Prompt

of approximately 250 words ([18]). This study proposes to use data from FAU WAC assessment of writing samples from a first-year composition course, ENC 1101: College Writing 1, as the basis for developing a neural language model for AWE that can accurately score student writing. Our objective is to develop a model that can assess and also provide feedback to students on their writing according to the categories of the learning model.

The above observations motivate the proposed research, which seeks to answer the following question: Can the elementary units of argumentation in thesis-driven, reading-centered, long-form essays taught in College Writing 1 and 2 be determined automatically? By using a real-world dataset as the case study material, our study essentially advances AI-integrated AWE systems for improving core writing skills in writing-intensive university courses.

II. WAC Data Collection

In this section, we will describe the data collected for this study, the instrument used to assess sample student papers, the WAC rating process, and the annotation process. In order to fulfill state-mandated General Education Curriculum outcomes for written communication, the WAC program at FAU provides a set of writing guidelines that can be adopted across curriculum of diverse disciplines, and assesses thesis-driven, research-based, near-end-of-term writing assignments. The WAC program defines "thesis-driven" as "papers with a thesis in which you build a case for a particular analysis, interpretation, or evaluation of data that leads to recommendations or specific conclusions," and defines "reading-based" as papers that "draw upon argument-driven articles or book chapters or in some cases works of literature. Typically, papers that are thesis-driven and reading-based are research projects of some kind'." [23]

A. Data Set: College Writing 1 and 2

While the WAC program assesses student writing from a variety of disciplines and course levels, which we intend to examine in future work, in this study we specifically examined thesis-driven, reading-based student writing in ENC 1101: College Writing 1 and ENC 1102: College Writing 2. This approach sought to focus on core writing competencies that form the basis of writing in the general education curriculum. In College Writing 1 and 2, students write analytical argument-based essays of 1000+ words in order to demonstrate skills in analytical thinking and reasoning in response to readings on contemporary topics in genres including memoir, creative nonfiction, editorials, and long-form journalism.

Student essays are evaluated for demonstration of rhetorical skills tied to course outcomes including argument and reasoning, evidence and support, organization, language and style, and meeting audience needs. Success in these assignments requires students to interpret assigned readings, research additional sources appropriate for their analytical purpose, and use textual evidence from the sources via paraphrasing and quotation to support their analytical argument. In such essays, students typically state their central argument in their introductory paragraph via a thesis statement, and also indicate the shape of their argument and the evidence they will use as part of their introduction and thesis. Arguments are analytical, not empirical, and so evidence will primarily come from the texts, but students may also include personal experience as part of their interpretation of textual evidence. Students may shift between registers of formality throughout their essay, introduce slang or writing from different languages or dialects, and alternate between first, second and third persons in order to persuade the reader of their analytical argument.

B. WAC Assessment Overview

In order to evaluate student writing from across the university and allow for different disciplinary conventions, WAC assessment employs a hybrid rubric to assess student writing across 11 categories. This approach combines analytical and holistic rubrics in place of a single holistic score per essay, which provides more detailed feedback on students' writing skills. As seen in Table I, the assessment evaluates core writing skills including opening strategy, argumentative features, organizational structure, concluding strategy, disciplinary concerns, grammar, and syntax. These core writing skills are divided into the following categories: thesis, organizational framework, reasoning, evidence, rhetorical structure, implication and consequences, academic tone, disciplinary conventions, clarity, style and, mechanics.

During assessment, each category is assigned an integer score on a 4-point continuum between 1 and 4, 4 – Extremely Effective, 3 – Effective, 2 – Adequate and , 1 – Inadequate. Each essay is scored by three raters using Agreement-Disagreement method. This is a simple percentage of agreement between the set of raters. The advantages of using this statistic is that it is physically easy to calculate and understand. The disadvantage is that raters could agree by chance or by factors that have little or nothing to do with the criterion measured. On a 4-point scale, such as the one used in this study, raters could agree 25% of the time by chance. Thus, this procedure tends to over-estimate Inter-Rater Reliability.

Final scores for each of the 11 categories are determined through a process of central tendency. What this means is that modal scores are used as the default. Median scores are used secondarily. For example, if all three raters award the same score (e.g., "3"), it is considered the final score for that paper, for that category. If only two raters award the same score, that score becomes the final score for that paper, for that category, regardless of the level of disagreement by the third rater. If all

TABLE I
ELEVEN CATEGORIES OF WAC ASSESSMENT

WAC Categories	Definitions
Thesis/purposes/argument	Persuasive purpose of the paper
Organizational Statement	A statement describing the building of the argument
Reasoning	Analysis of evidenciary materials and demonstrated comprehension of ideas
Evidence	Integration of data, quotations, visuals, and counterarguments
Rhetorical Structure	Sustained focus on argument's development and its progression
Implications and Consequences	Development of argument's conclusion
Academic Tone	Formality of specialized terms and concepts
Disciplinary Conventions	Discipline-specific formatting and citation
Clarity	Sentence-level comprehension and consistent usage of discipline-specific terminology
Style	Linguistic variation between sentences
Mechanics	Mechanical sentence level error patterns

three raters award different scores, the middle (median) rater's score is used as the final score for that paper, for that category.

C. WAC Data Annotation

In this study we focused on core writing skills specific to College Writing 1 and 2 courses. As with [16] where a large textual corpora is manually analyzed and annotated, we also annotated texts. Unlike [16], however, we did not adopt argumentative structures of premises and conclusions. Instead, we adopted annotation labels that aligned with elementary units identified in WAC and English Composition assessment rubrics. As noted in Table II, these elementary units aligned with writing genre common to English Composition courses and WAC assessment objectives.

Because argumentation is the primary skill taught in College Writing 1 and 2 courses, the annotation procedure adopted in our study identified six core writing skills in College Writing 1 and 2 essays that align with five WAC rubric categories: thesis, organizational framework, use of evidence, analysis of evidence, reasoning, evidence, and rhetorical structure. Table II provides descriptions of the six core writing skills annotated by the raters.

This annotation procedure enabled us to break the WAC category of rhetorical structure into two components: raters annotated 1) rhetorical focus to identify where a student paper returned the reader's attention to the argument's central idea, and 2) rhetorical progression where the paper identified the argument's organization. This is an especially appropriate distinction in writing intensive courses where students write thesis-driven and reading-based extended essays.

In order to develop a fixed model of argumentation that could be applied to our corpus, we adopted a monological model (or Dialectic Argumentation), whereby we sought to identify the elementary units. In other words, each rater who scored a student essay also annotated it by labelling every sentence into the six categories mentioned, as defined in Table II. The sentences which did not belong to any of the categories were labelled as "plain text".

Although, in few cases it so happened that the raters found that few sentences could not be labelled under a single category. A portion of the sentence, then, was annotated under one category and the remaining portion was labeled as another category. At times, some sentences contained three categories. We will discuss how we tackled this issue in the following subsection.

III. METHODOLOGY

Fig. 2 lists the proposed framework for consistency checking, filtering, model fine-tuning, and comparisons.

A. Data Preprocessing

As each essay is analyzed by more than one rater, it is natural to have contrasting opinions on sentence annotation. One sentence can be categorized under multiple categories by different raters. This problem became more severe when a sentence, in itself, contributed towards multiple categories (usually long sentences). We addressed this issue by breaking down the sentence into individual words and associated each word with its corresponding label. Then, we counted the number of words associated with each label and assigned that label to the sentence which had the maximum word count. In a similar fashion, after a label was assigned to a sentence, we compared the final label assigned to the sentence by each rater and selected the label that was the most frequent. A natural question one might ask is, "what if none of the raters agree upon a single category?" For cases like these, we proceeded with the annotation of the rater who showed the greatest consistency in their annotations.

B. Consistency Evaluation

Evaluating the consistency of the raters was an important step as erroneous annotations can cause improper training of the machine learning model and faulty predictions. It was

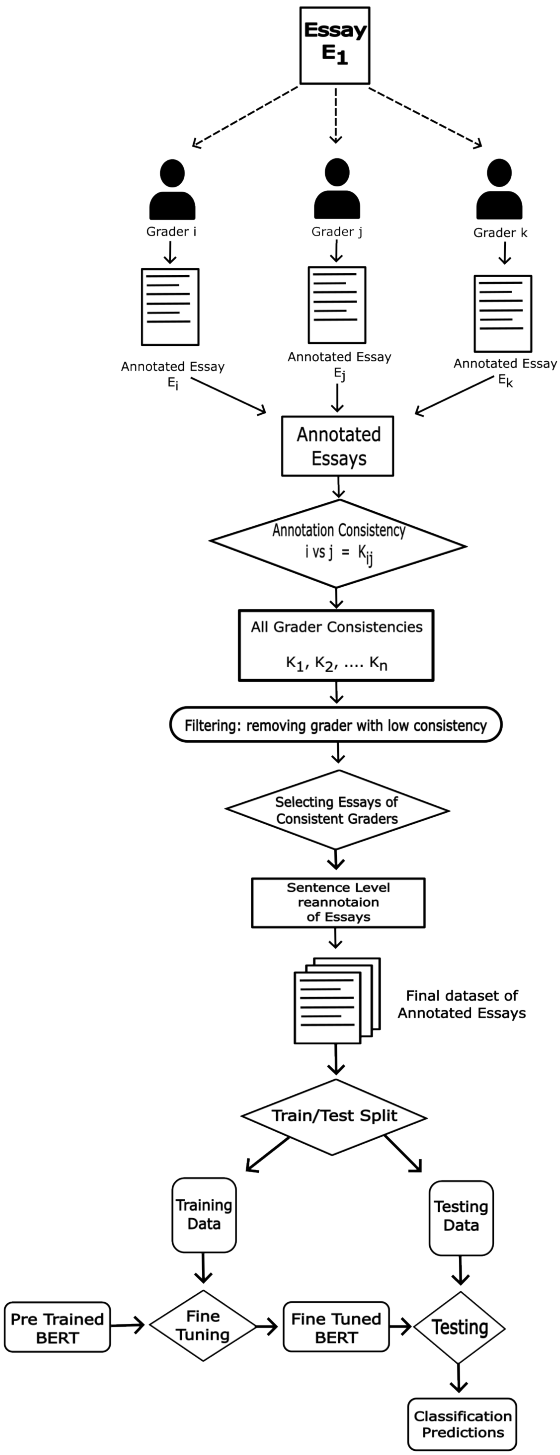

Fig. 2. The diagram of the essay grading consistency check and training/test splitting for validation and comparisons. From top to bottom, each essay is assessed by multiple raters, who annotate sentences based on their own perceptions. Annotation consistency intends to find raters highly inconsistent with others, with low consistency raters' annotations being removed. The final labels only considering consistent raters' annotations.

TABLE II
ANNOTATED CATEGORIES

Core Skills for College Writing 1 and 2	WAC Categories	Descriptions
Thesis	Thesis/purpose/argument	Persuasive purpose of the paper
Organizational Framework	Organizational Framework	A statement identifying the argument's oranization
Inclusion of Textual Evidence	Evidence	Paraphrasing and quotatations
Analysis of Evidence	Reasoning	Analysis of evidenciary materials, especially quotations
Rhetorical Focus	Rhetorical Structure	Statements identifying the argument's central idea
Rhetorical Progression	Rhetorical Structure	Statements identifying the argument's organization

crucial, then, to check the similarity between the annotations of raters and consider only the group of raters whose similarity reached a certain threshold. This process of measuring the consistency and agreement between two or more raters in their assessments, judgements or evaluation of any phenomenon or behavior is called Inter-Rater Reliability (IRR). And in the context of annotation, this is known as Inter-Annotation Agreement (IAA).

There are numerous metrics by which one can gauge the level of agreement between raters, such as Cohen's Kappa [8], Fleiss's Kappa [9] and Krippendorff's Alpha. Cohen's Kappa is used to compute the agreement between no more than two raters, while Fleiss's Kappa and Krippendorff's Alpha can be used to compute the agreement between multiple raters for a given set of annotation. We chose Cohen's Kappa as our consistency evaluation metric as we needed to measure the average degree of consistent annotation of each rater with every other rater as a tie-breaking criterion for ambiguous annotation rather than just the overall agreement of raters for every essay. It is given by,

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{1}$$

where p_o is the relative observed agreement among raters, i.e., the fraction of total number of sentences that same (or agreed) annotations by both raters.

$$p_o = \frac{\text{\# of agreement sentences}}{\text{Total number of sentences}} \tag{2}$$

and p_e is the hypothetical probability of chance agreement,

$$p_e = \sum_{i=1}^{K} \left(\frac{\mathrm{C}_1^i \times \mathrm{C}_2^i}{\text{Total number of sentences}^2} \right) \tag{3}$$

where K is the total number of categories (the labels of sentences in our case). Where $\mathrm{C}_1{}^{\mathrm{i}}$, $\mathrm{C}_2{}^{\mathrm{i}}$ are, for any given essay, the number of sentences annotated as label *i*.

To demonstrate the application of Cohen's Kappa for IAA, a brief example is shown below. Table III illustrates four sentences where two raters, G_1 and G_2, are asked to annotate the essay based on two categories - A and B. The annotations are shown in Table III for the four sentences.

TABLE III
AN EXAMPLE OF COHEN'S KAPPA SCORE CALCULATION

Sentences	Rater (G_1)	Rater (G_2)
Sentence 1	A	A
Sentence 2	B	B
Sentence 3	A	A
Sentence 4	A	B

Calculations for the level of agreement of the two raters, G_1 and G_2, using Cohen's Kappa metric employ the following steps: We initially calculate the relative observed agreement p_o between raters. Since both raters agreed on three sentences in total, sentence 1, 2 and, 3, the relative observed agreement is first set-up as follows:

$$p_o = \frac{3}{4}$$

The next step is to compute the probability of agreement by chance. According to equation 3 we proceed with the following:

$$p_e = \sum_{i \in [A,B]} \left[\frac{\mathrm{C}_1^i \times \mathrm{C}_2^i}{\text{Total number of sentences}^2} \right]$$

$$= \frac{1}{4^2} \times [(3 \times 2) + (1 \times 2)]$$

$$p_e = \frac{1}{2}$$

Now that we have p_o and p_e, the values are substituted in equation 1 to arrive at the Kappa value:

$$\kappa = \frac{\frac{3}{4} - \frac{1}{2}}{1 - \frac{1}{2}}$$

$$\kappa = \frac{1}{2}$$

Hence, in this brief illustration, the raters G_1 and G_2 have an agreement of 50%. In our study, this process was carried out for every pair of essay raters for every single essay.

Table IV shows the interpretation for kappa values introduced by Landis and Koch [11]. In our study, then, we adopted the approach discussed above to calculate the consistency of each rater in the following way:

1) For each essay, we calculated Cohen's Kappa Score for every pair of raters. Every score contributed to the consistency of both raters.
2) We then averaged the all the Cohen's Kappa values associated with each rater to compute their individual consistency scores.

TABLE IV
INTERPRETING KAPPA VALUES

Kappa Range	Interpretation
<0.00	No Agreement
0.00 - 0.20	Slight Agreement
0.21 - 0.40	Fair Agreement
0.41 - 0.60	Moderate Agreement
0.61 - 0.80	Substantial Agreement
0.81 - 1.00	Perfect Agreement

C. Sentence Level Assessment

Since the nature of data we used is nominal, we treated this as a classification problem. As we were using a dialectic model of argumentation, it was crucial to use a model that could capture the underlying patterns in a unit of text, i.e., an argument. Hence, we used a BERT model introduced by Devlin [4] which is built upon on a Transformer model introduced by Vaswani [5] in 2017. The Bidirectional Encoder Representations from Transformers (BERT) is a state-of-the-art NLP model introduced by Google in 2018, designed to understand the context of words bidirectionally, unlike previous transformer models that processed text in a single direction. Built on the encoder stack of the Transformer architecture, BERT uses self-attention mechanisms across multiple layers to capture complex relationships between words. Pre-trained on huge corpora using methods like Masked Language Modeling (MLM), where random words are masked and predicted, and Next Sentence Prediction (NSP), BERT is able to learn deep bidirectional representations. This pre-training enables BERT to be fine-tuned on various downstream tasks, such as question answering and sentiment analysis, with minimal task-specific data. The BERT model we used was pre-trained on the BookCorpus dataset (which is a collection of 11,038 unpublished books) and Wikipedia-English corpus. We then fine-tuned the model with our data of 5,064 labelled text sentences. We did 5-fold cross validation to evaluate the performance of our model on a split of 90% training data and 10% testing data for 10 epochs each.

After calculating the consistencies of all the raters, we attempted to identify raters with a low consistency scores. As mentioned earlier, this helped us improve the quality of data. After identifying a list of consistent raters, we created a data pool of essays from these raters. This pool consisted of multiple versions of every single essay, which we then analyzed in order to finalize the category label for each sentence. We resolved labeling discrepancies by assigning each sentence to the category for which there was the highest rater agreement. In cases where none of the raters agreed on any category, sentences were labeled according to the label assigned by the rater with the highest consistency value. This process created a new data pool of essays consisting of a single version of each essay and final annotations that were then used for fine-tuning evaluation.

We then used this refined data pool to conduct two experiments, the results of which are discussed in the following section.

IV. EXPERIMENTS AND RESULTS

Table V lists the dataset collected for the study, which consisted of sentences labeled in six categories (the table lists only sentences after low consistency raters were removed). The "plain text" category denotes sentences that were not labeled by raters. The dataset showed a clear imbalance, with "plain text" as the predominate category and "Organizational Framework" being the least represented category (which is 1/7 of the second dominant class "Evidence"). Due to this category distribution, we adopted precision, recall, and F-1 scores to assess the model performance.

TABLE V
A SUMMARY OF NUMBER OF SENTENCES ASSESSED BY RATERS AND THE RESPECTIVE ASSESSMENT CATEGORIES

Sentence Assessment Label Categories	# of Sentences
Plain text	2,590
Evidence	721
Reasoning	656
Rhetorical Structure (Focus)	563
Rhetorical Structure (Progression)	234
Thesis	192
Organizational Framework	108

A. Rater Consistency Evaluation Results

Figure 3 shows the consistency scores of all the raters involved in the scoring and annotation process. As Figure

V demonstrates, all the raters were in the range of Fair to Moderate Agreement, with the exception of one rater who was in the level of Slight Agreement. As explained above, these consistency scores were also used to filter out raters who showed poor annotation since low consistency score means their annotations were considerably and frequently different from other raters. By only considering annotations from raters who showed consistency, we increased the quality of data.

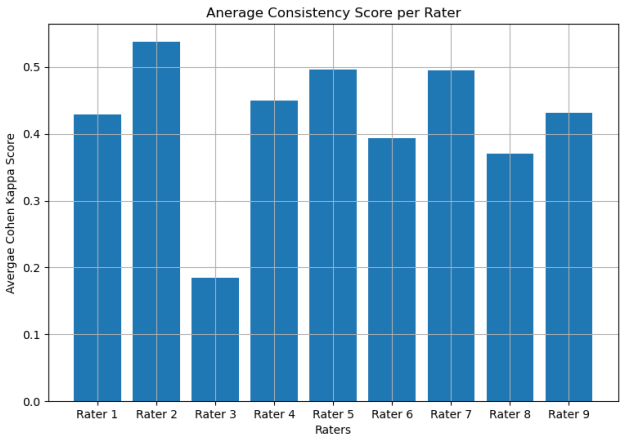

Fig. 3. Consistency score of Raters. x-axis denotes individual rater, and the y-axis denotes each rater's average Cohen Kappa score.

B. Sentence Assessment Results

To evaluate sentence-level assessment results, we designed two experiments in our study: (1) Aggregated sentence assessment, and (2) Isolated sentence assessment. The purpose of these experiments was to understand how the model performed in respect to training data and to fine-tune BERT for sentence assessment.

In the aggregated sentence assessment experiment, we modified the data such that consecutive sentences belonging to the same category within an essay were treated as a single data entry in the dataset. For each essay, all sentences belonging to one category were aggregated as one sentence to fine-tune BERT. In the isolated sentence assessment experiment, for each essay, we treated each sentence as a separate data entry to fine-tune BERT.

For both the experiments we used HuggingFace Transformer, specifically "bert-base-uncased" transformer. We ran the experiments in Google CoLab on the A1000 GPU.

Tables VI and VII summarize the results of the two experiments, showing the classification performance and overall performance metrics for the six labels. "Evidence" showed high precision in both experiments, with slightly better F-1 scores in isolated sentence assessment. "Organizational Framework" showed low recall and F-1 scores in both experiments, with a notable drop in isolated experiment's recall, which means that the model struggled to identify this class accurately. This drop can be considered because "Organizational Framework" only comprises 108 sentences in total, i.e., barely 2% of the entire dataset. The precision and recall of "Reasoning" were relatively consistent between the two experiments, indicating moderate model performance on this label. Both "Focus" and "Progression" indicated fluctuating precision and recall, with isolated experiment improving F-1 scores compared to aggregated experiment, primarily for "Progression." For "Thesis," while precision was high, recall and hence F-1 was notably low, indicating that the model could not identify all instances correctly. "Plain text," because of its high volume, showed high recall and precision in isolated experiment.

The results suggest that isolated sentences generally outperform aggregated sentences for sentence-level assessment. This increase in performance may be attributed to the increase in the total number of data points per category for fine-tuning the BERT.

V. Conclusion

In this paper, we explored the use of neural language models, such as BERT, for automatic sentence-level assessment of student writing.

The experiments conducted demonstrate that while these models show potential in evaluating sentence-level writing with reasonable accuracy, their performance heavily depends on the consistency of human annotations. The reliability of annotations plays an important role in the model's effectiveness, as shown in the metrics used for evaluation, including accuracy, precision, recall, F1-score, and Cohen's Kappa, which was employed to measure the Inter-Annotator Agreement.

The primary goal of developing an AI system for sentence-level assessment is to create an AES system that directly scores the student essays according to the WAC rubric (or other writing assessment criteria related to critical thinking) and that provides real-time feedback to the students. Such a system reduces the variability and subjectivity inherent in manual evaluations.

Our study shows the challenges posed by inconsistencies in human annotations, which affect the model's reliability and performance. These variations often result from employing different annotators over time, leading to inconsistent quality and ambiguity in the training data. To address this issue, it is essential to standardize the annotation process, ensuring that the data used to train the model reflects the writing categories being assessed. Properly annotated data enables the model to learn the differences between sentence categories and improve its ability to reflect human judgment. Establishing a standardized annotation system through an AI model reduces the need for continuous manual monitoring and the recalculation of annotation consistency every time raters change. An AI-based annotation system can provide a more stable and cost-effective solution for educational institutions.

Future work should address data imbalance, using approaches, such as data augmentation. Unlike computer vision, in Natural Language Processing, data augmentation should

TABLE VI
CLASSIFICATION REPORT

Experiment	Aggregated Sentence Assessment			Isolated Sentence Assessment		
Label	**Precision**	**Recall**	**F-1**	**Precision**	**Recall**	**F-1**
Evidence	0.94	0.82	0.88	0.92	0.87	0.89
Org. Framework	0.75	0.27	0.40	1.00	0.05	0.09
Reasoning	0.61	0.57	0.59	0.59	0.53	0.56
RS - Focus	0.41	0.40	0.40	0.36	0.58	0.45
RS - Progression	0.36	0.32	0.34	0.62	0.38	0.47
Thesis	0.83	0.26	0.40	1.00	0.24	0.38
Plain text	0.47	0.66	0.56	0.80	0.84	0.82

TABLE VII
PERFORMANCE METRICS

Experiment	Aggregated	Isolated
Accuracy	0.59	0.70
Mean Squared Error	4.40	3.32
Mean Absolute Error	1.20	0.91

be done carefully due to the grammatical structure of the text. Text Augmentation techniques like Synonym Replacement, Random Swapping, Random Deletion/Insertion, or back Translation can be used to balance the dataset by adding synthetic text sentences to class labels with low data count.

ACKNOWLEDGMENT

This research is partially sponsored by the U.S. National Science Foundation under grant Nos. IIS-2236579, IIS-2302786, the Florida Atlantic University Dorothy F. Schmidt College of Arts and Letters seed grant, and Florida Atlantic University's Writing Across the Curriculum Program. We would also like to thank Anudeep Reddy Raavi for his contributions in data cleansing, and implementing the fine-tuning mechanism.

REFERENCES

[1] Bazerman, Charles and Little, Joseph (2005). Reference guide to writing across the curriculum, *Parlor Press LLC*, 2005.

[2] Michelle Cox, Jeffrey R. Galin, Dan Melzer. (2018). *Sustainable WAC: A Whole Systems Approach to Launching and Developing Writing Across the Curriculum Programs,* National Council of Teachers of English.

[3] Van Eemeren, F. H., Grootendorst, R., Johnson, R. H., Plantin, C., & Willard, C. A. (2013). Fundamentals of argumentation theory: A handbook of historical backgrounds and contemporary developments. Routledge.

[4] Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[5] Vaswani, A. (2017). Attention is all you need. Advances in Neural Information Processing Systems.'

[6] Lindahl, A.,& Borin, L. (2024). Annotation for computational argumentation analysis: Issues and perspectives. Language and Linguistics Compass, 18(1), e12505.

[7] Ling, G., Elliot, N., Burstein, Jill., McCaffrey, D., MacArthur, C., Holtzman, S. (2021) Writing Motivation: A Validation Study of Self-Judgment and Performance. Assessing Writing, 48, 100509.

[8] Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and psychological measurement, 20(1), 37-46.

[9] Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. Psychological bulletin, 76(5), 378.

[10] Conference on College Composition and Communication Executive Committee. (2023). Principles for the Postsecondary teaching of writing. National Council of Teachers of English.

[11] Landis JRKoch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159174.

[12] McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. Written communication, 27(1), 57-86.

[13] Godwin-Jones, R. (2022). Partnering with AI: Intelligent writing assistance and instructed language learning.

[14] Roscoe, R. D., Allen, L. K., Weston, J. L., Crossley, S. A., & McNamara, D. S. (2014). The Writing Pal Intelligent Tutoring System: Usability Testing and Development. *Computers and Composition*, *34*(Complete), 39–59.

[15] Butterfuss, R., Roscoe, R. D., Allen, L. K., McCarthy, K. S., McNamara, D. S. (2022). Strategy uptake in writing pal: Adaptive feedback and instruction. Journal of Educational Computing Research, 60(3), 696-721.

[16] Palau, R. M., & Moens, M.-F. (2009). Argumentation mining: The detection, classification and structure of arguments in text. *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, 98–107.

[17] Su, Y., Lin, Y., Lai, C. (2023). Collaborating with ChatGPT in Argumentative Writing Classrooms. Assessing Writing, 57, 100752.

[18] Tang, X., Chen, H., Lin, D., Li, K. (2024). Harnessing LLMs for multi-dimensional writing assessment: Reliability and alignment with human judgments. Heliyon, 10(14).

[19] Ramesh, D., Sanampudi, S. K. (2022). An automated essay scoring systems: a systematic literature review. Artificial Intelligence Review, 55(3), 2495-2527.

[20] Chen, Y. Y., Liu, C. L., Lee, C. H., Chang, T. H. (2010). An unsupervised automated essay-scoring system. IEEE Intelligent systems, 25(5), 61-67.

[21] Kumar, V. S., Boulanger, D. (2021). Automated essay scoring and the deep learning black box: How are rubric scores determined?. International Journal of Artificial Intelligence in Education, 31, 538-584.

[22] Ludwig, S., Mayer, C., Hansen, C., Eilers, K., Brandt, S. (2021). Automated essay scoring using transformer models. Psych, 3(4), 897-915.

[23] Florida Atlantic University WAC Assessment Student Information. https://www.fau.edu/wac/assessment/students/

[24] Zhu, W., Sun, Y. (2020, October). Automated essay scoring system using multi-model machine learning. In CS IT Conference Proceedings (Vol. 10, No. 12). CS IT Conference Proceedings.

[25] Fernandez, N., Ghosh, A., Liu, N., Wang, Z., Choffin, B., Baraniuk, R., Lan, A. (2022, July). Automated scoring for reading comprehension via in-context bert tuning. In International Conference on Artificial Intelligence in Education (pp. 691-697). Cham: Springer International Publishing.