Evaluation with Language Models in Non-formal Education: Understanding Student's Persuasion Style in Competitive Debating

Ali Al-Zawqari[1], Mohamed Ahmed[2] and Gerd Vandersteen[1]

[1]*Department of Fundamental Electricity and Instrumentation, Vrije Universiteit Brussel, Elsene, 1050, Belgium*
[2]*QatarDebate Center, Doha, Qatar*

Abstract

This study investigates the application of language models in competitive Arabic debating, a non-formal educational activity aimed at enhancing critical thinking and argumentative skills. Given the extensive training and assessment typically required in competitive debating, our research proposes applying encoder-based language models to automate and refine the evaluation process of students' persuasive styles, thereby assisting in both assessment and training material design. Utilizing the Munazarat 1.0 corpus, which comprises approximately 50 hours of competitive Arabic debates, we employed three BERT-based models to classify students' persuasion into Aristotle's rhetorical categories: ethos, pathos, and logos. Our results indicate that language models can successfully identify different persuasion styles. This suggests a transformative potential for these models in debate training by providing automated, scalable, and detailed feedback. The study also identifies challenges in ensuring model fairness, with variations in performance across different demographic groups highlighting the need for further calibration to achieve equitable outcomes. The implications of these findings suggest that while language models hold promise in non-formal educational settings, their application must be carefully managed to avoid reinforcing existing biases.

Keywords

AI in education, competitive debating, language models, non-formal education

1. Introduction

Debating has a long history, and it was established as a teaching method by Protagoras in Athens between 481 and 411 B.C. [1]. As a pedagogical tool, debate aims to foster critical thinking, communication skills, and intellectual engagement among students. Students learn to research thoroughly, develop coherent arguments, and critically evaluate opposing viewpoints by participating in structured arguments [2]. This process enhances their ability to think logically and articulate ideas persuasively, which is crucial for academic and real-world settings. Moreover, debate encourages active learning and democratic engagement, preparing students for informed societal participation [3]. Despite this, debate participation in most high schools and universities is limited mainly to students on competitive debate teams [4]. Due to that, and like writing assignments, integrating debates into various subjects has been encouraged. These

EvalLAC'24: Workshop on Automatic Evaluation of Learning and Assessment Content, July 08, 2024, Recife, Brazil

aalzawqa@vub.be (A. Al-Zawqari); mahmed@qatardebate.org (M. Ahmed); gerd.vandersteen@vub.be (G. Vandersteen)

subjects include sociology, history, math, health, and marketing. Also, written debates have proven effective in online courses [5].

On the other hand, artificial intelligence in education (AIEd) has shown that through a statistical evaluation of rich data sources, it is possible to support teachers in gaining insights into student performance in physical, online, and hybrid environments [6]. Another aspect of supporting teachers with AIEd is evaluating learning and assessment content. This critical aspect traditionally relies on methods like Item Response Theory [7] and increasingly incorporates machine learning techniques for more dynamic assessments [8, 9, 10]. However, areas like the evaluation of lectures, whole courses, and curricula still predominantly require expert human intervention. Even components such as distractors in multiple-choice questions are mostly manually labeled due to the limitations of existing automated evaluation methods [11, 12, 13, 14]. With the rise of large language models in education, including both open models like Gemma and Llama 2, and closed systems like GPT-4, there is an urgent need to develop robust metrics to accurately assess the educational content they generate [15, 16, 17]. This is essential to measure the effectiveness of large language model applications in education and ensure that the generated content adheres to the learning objectives, maintains factual accuracy, and upholds equity, diversity, inclusion, and belonging (EDIB) standards. This problem is actively studied in other AIEd applications, such as predictive modeling [18, 19].

In this paper, we introduce our work on leveraging modern language models as a tool of AIEd to evaluate a non-formal educational activity, namely competitive debating. The work here is limited to competitive debating in the Arabic language. However, the same methodology can be applied to English since the rules of competitive debating are shared between the two languages. First, we briefly introduced related work in section 2. In section 3, we show the motivation behind integrating language models with competitive debating. In the same section, we explain the aim of this work and describe the initial output related to data creation/annotation and validation. In section 4, we present and discuss the early results. Finally, we summarize the current work and describe the ongoing and future research in section 5.

2. Related Work

Non-formal education is defined as any organized educational activity outside the formal system designed to meet specific learning objectives for identifiable audiences [20]. Non-formal education equips students with tools to enhance cognitive and creative skills [21]. Recognizing the importance of standardizing evaluation methods for non-formal education is crucial for its broader social acceptance [22]. Research has explored the impact of non-formal education through various studies, such as a survey assessing the value of non-formal education programs in providing life skills and employment opportunities to 994 graduates [23], and an evaluation involving non-test techniques like assessment rubrics, observations, interviews, and documentation studies across 24 non-formal educational institutions [24].

Incorporating debates within non-formal educational settings has proven effective in fostering critical and higher-order thinking skills [25]. Debates enhance learning outcomes and student satisfaction, particularly in online education settings [26]. Also, debating shows that it promotes a deeper understanding of course concepts, improves critical thinking, and increases collaboration among students [27]. The use of formal debates can also stimulate student interest and enhance conceptual understanding in college classrooms [28].

The exploration of language models in educational contexts, particularly following the release of ChatGPT in 2022, has highlighted their potential in various educational domains. For instance, employing language models like ChatGPT in healthcare education has been beneficial, aiding students and researchers in tasks such as literature reviews, dataset analysis, and enhancing critical thinking skills [29]. Furthermore, these models have shown promise in boosting engagement in language learning by teaching advanced skills like inference, interpretation, and critical analysis [30].

Despite the growing body of research on language models in education, studies examining their effects and applications in non-formal education settings remain lacking. Here, we explore the use of encoder-based language models in evaluating and aiding competitive debating as a non-formal education activity.

3. Methodology

3.1. Integrating Language Models with Competitive Arabic Debating

Competitive debating is an intensive oral discourse activity under strict rules and guidelines. Students from universities and schools across various regions engage in local and international Arabic debate tournaments, which follow a modified version of the World Schools Debating Championship format [1]. Each debate involves a motion, with two teams debating as either the proposition or opposition. Each team comprises three speakers, each given 6 to 7 minutes to present their arguments. An adjudication panel then determines the winning team based on their persuasive abilities and effectiveness in argumentation and refutation. These debates are rich in argumentative content because the motion is crafted to address a specific issue, the speaking time is limited, and the evaluation criteria emphasize argumentative and persuasive skills. As mentioned above, competitive debating is designed to help students develop skills related to critical thinking, the ability to analyze the opposition's point of view, and engaging in democratic discourse. Students usually go through training sessions before entering competitive debating competitions.

Here, we propose to utilize the latest advancements in language models to assist teachers in analyzing students' debates. In addition, language models can be used to design the training materials based on the results received from the adjudicating panel. In Figure 1, we show the proposed framework for integrating language models with competitive debates. We create a dataset based on the recorded debates via Speech-to-text models. This dataset was built via human-validated machine transcription. Once we have a debate transcripted, a language model can help in three different stages that serve the final goal of evaluating and creating exercise materials for the students: 1) Extract the argumentative elements from the student's speech, 2) map the student's argumentative elements to one of Aristotle three persuasion modes, and 3) evaluating the teaching materials and adapting the training based on the argumentative elements, debating style, and the debate results. At this stage of work, we explore the usage of

[1] https://www.wsdcdebating.org/

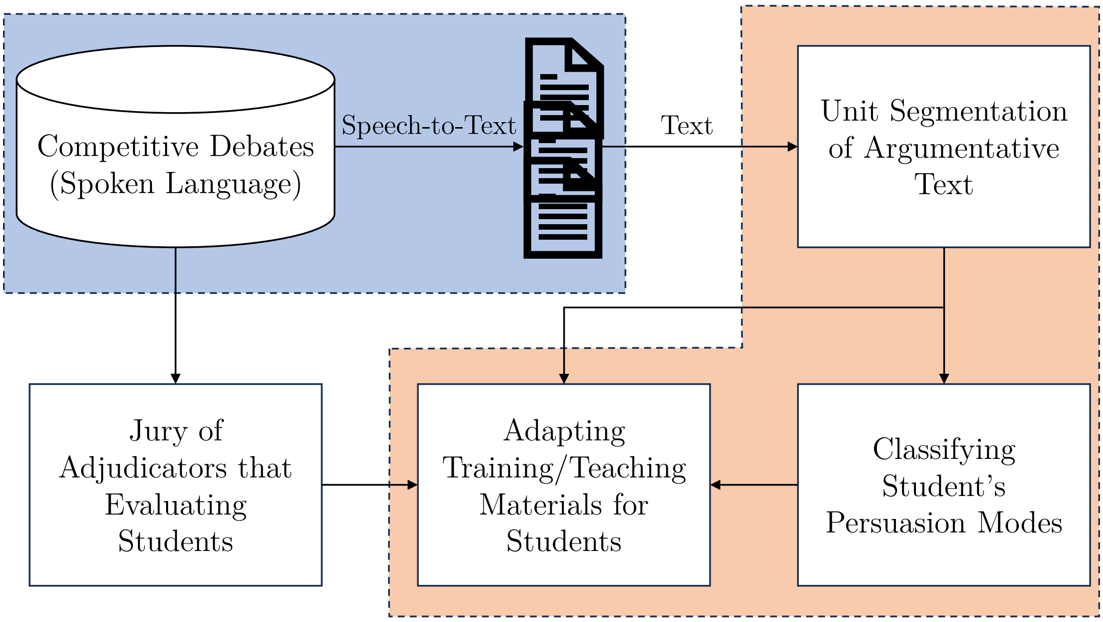
Competitive Debates (Spoken Language)
Speech-to-Text
Text
Unit Segmentation of Argumentative Text
Jury of Adjudicators that Evaluating Students
Adapting Training/Teaching Materials for Students
Classifying Student's Persuasion Modes

Figure 1: Integrating language models with the evaluation of competitive debating. 🟦 denotes the current dataset that was created with human-validated transcriptions, 🟧 denotes the stages where language models are used in evaluating and designing teaching materials

language models in the second stage, which is in classifying students' persuasion modes.

3.2. Dataset Building and Annotation

Munazarat 1.0[2] is a corpus of competitive Arabic debates comprising approximately 50 hours of transcribed debates that the QatarDebate Center hosted in several tournaments. The corpus is created using 73 debates pre-recorded and already published online by the QatarDebate. In the first stage of transcription, a multi-lingual Arabic/English speech-to-text tool called Fenek [31] was used to transcribe the videos. Then, each debate went through three stages of manual review, which three different reviewers performed to correct any mistakes from the tool in the transcription process. The Arabic debating corpus is described in detail with the meta-data in [32].

A hybrid annotation model was proposed in [33]. The model combines Aristotle's three appeals of logos, ethos, and pathos with Toulmin's model of argument structure analysis, in addition to some added labels inspired by the unique nature of competitive debates. Forty debates from the Munazarat 1.0 corpus were selected and annotated. The annotation process was separated into two stages: 1) annotating each debate twice by two different annotators and then undergoing a thorough review by one of the authors to standardize and enhance the annotation process, and 2) annotating the later 20 debates once and reviewing the annotation results by an expert. A detailed description of the annotation models is presented in [33].

[2] https://github.com/moh72y/Munazarat1.0/

4. Initial Experiments

4.1. Classification of Student's Persuasion Style

As mentioned in subsection 3.1, the focus at this stage is on the classification of students' persuasion style in a non-formal education setting. In the initial phase of our study, the debate transcripts were segmented into six parts, aligning with the three debater positions (first, second, and third) on both the proposition and opposition sides. Each segment is labeled with the debater's position and a unique identifier for each debate. We then grouped Toulmin's argumentative elements into broader syllogistic categories: Ethos, Pathos, and Logos, thus simplifying our dataset into three main persuasion styles. This dataset contains 240 speeches and approximately 204k words. The dataset is split into two sets: a training set (including validation data) containing 34 debates and a testing set comprising 6 debates. No debate was represented in both sets to prevent data leakage. Additionally, the testing set was curated to maintain a balance across various demographics like gender and debate topics to mitigate potential biases.

Given the small size of the annotated dataset, we select a baseline and three BERT-based language models for training and evaluation: ULMFiT [34] as the baseline model, CAMeL-BERT [35] pre-trained on diverse Arabic datasets, ARBERTv2 [36] developed from the largest Modern Standard Arabic (MSA) corpus, and DeBERTa [37] trained only on English data. The baseline model involved training an AWD-LSTM [38] on Arabic Wikipedia and then fine-tuning it using texts from the training set of the Munazarat 1.0 corpus. For the baseline model, we tokenize the speeches using the ARBERTv2 tokenizer since it has the largest vocabulary of MSA. For all the other models, we use each language model tokenizer. We add a classification head to the encoder-based model for fine-tuning and then use the training data to train the new classification head.

Our performance evaluation centered on the language models' capability to understand students' persuasion styles. Therefore, the F1-score is selected as the evaluation metric:

$$\text{F1-Score} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

$$\text{Macro F1-Score} = \frac{\sum_{i=1}^{n} \text{F1-Score}_i}{n} \tag{1}$$

where TP is the true positives, FP false positives, FN false negatives, and n the number of classes. Additionally, we assessed model fairness, which is crucial for ensuring adherence to EDIB standards in educational settings. This include tests for demographic parity and equal opportunity [39, 40], defined respectively as:

$$\mathbb{E}[f(X) \mid Z = z] = \mathbb{E}[f(X)] \quad \forall z \tag{2}$$

$$\mathbb{E}[f(X) \mid Z = z, Y = 1] = \mathbb{E}[f(X) \mid Y = 1] \quad \forall z. \tag{3}$$

Demographic parity requires the classifier's output $f(X)$ to be independent of specific attributes Z (like gender). Equal opportunity demands $f(X)$ and Z be conditionally independent given the positive class $Y = 1$.

4.2. Numerical Results

The evaluation results of the language models based on F1-score and fairness metrics (demographic parity & equal opportunity) are shown in Figure 2 and Figure 3. The results show that CAMelBERT MSA models outperform all the other models with around 90% F1 score in both validation and testing sets, which shows the capabilities of language models in categorizing the persuasion style of the student. This shows promising potential in using those models in debate training around the world, especially in areas with fewer resources in terms of capable debate judges and teachers.

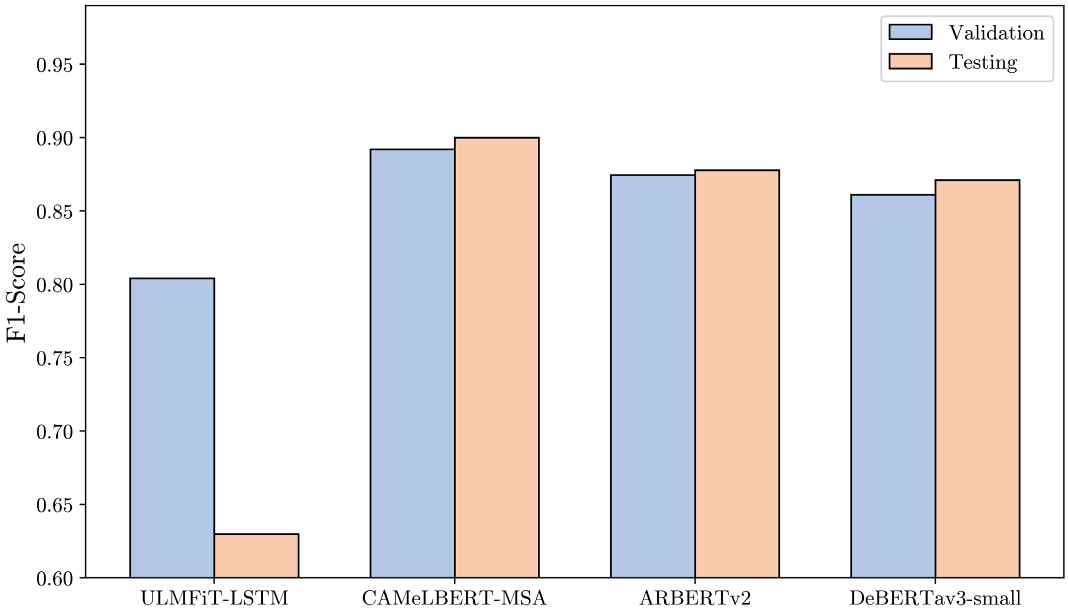

Figure 2: Evaluation of the language models on validation and testing sets.

In evaluating the fairness of the language models we use in this study, all models showed differences between male and female debaters, with values ranging between 3.5% to 5.4% for demographic parity and 2.2% to 4.6% for the equal opportunity measure. It is worth mentioning that the current dataset is less representative of the entire corpus as it only focuses on university-level students and native speakers. We expect the numbers to get worse with the introduction of school students and non-native speakers which highlights the importance of further investigating the fairness of language model-based classifier a critical issue. More extensive experiments with other pre-trained language models on different data are detailed in [41].

5. Conclusion and Future Work

This study explored the integration of language models into non-formal educational settings, particularly in the context of competitive debates. The primary focus was assessing how these

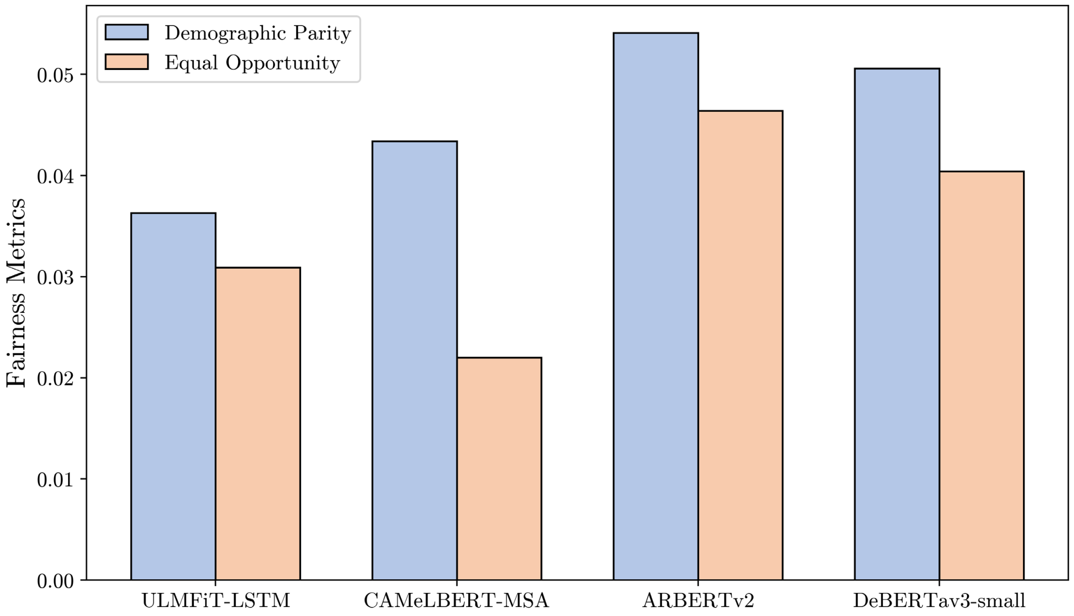

Figure 3: Evaluation of models' fairness using demographic parity & equal opportunity.

models can classify and analyze students' persuasive styles, a key skill in debate education. The results demonstrate that language models are capable of identifying different persuasion strategies employed by students. However, they also highlighted issues related to fairness that need addressing, especially when these models are intended to be used at scale. These findings suggest that while language models can be a valuable tool for analyzing debate performances, their current application is limited by concerns over equitable outcomes.

In future work, the scope of this research will be expanded at three main levels. Firstly, we will aim to enhance the granularity of analysis by using Human-LLM Collaborative Annotation to annotate debates into distinct argumentative units. This refinement will allow for a more detailed examination of argument structures and the effectiveness of various persuasive elements, thereby increasing the sophistication of the classification process. Secondly, the study will include diverse participants, not limited to university-level students, but to include the school-level debaters. In addition, non-native Arabic speakers will be another added sample to the data. This expansion is important to ensure the models are robust and equitable across a wider spectrum of linguistic and educational backgrounds, which will help in creating a more inclusive analysis of debating skills. Lastly, there will be a focus on linking debate outcomes (results obtained from the panel of adjudicators) with debating styles (persuasion modes of students) through language models. This approach involves developing algorithms that can assess the style and quality of debates and suggest specific educational content tailored to enhance the debaters' skills based on their performance metrics. This targeted feedback mechanism could help in improving debate training, making it more personalized and effective.

Acknowledgements

This work was made possible by the QD Fellowship award [QDRF-2022-01-005] from the QatarDebate Center. In addition, the work was supported in part by the Vrije Universiteit Brussel (VUB-SRP19 and SRP78).

References

[1] H. W. Combs, S. G. Bourne, The renaissance of educational debate: Results of a five-year study of the use of debate in business education., Journal on Excellence in College Teaching 5 (1994) 57–67.

[2] J. S. Huryn, Debating as a teaching technique, Teaching Sociology 14 (1986) 266–269.

[3] R. Kennedy, In-class debates: Fertile ground for active learning and the cultivation of critical thinking and oral communication skills., International Journal of Teaching & Learning in Higher Education 19 (2007).

[4] J. Bellon, A research-based justification for debate across the curriculum, Argumentation and Advocacy 36 (2000) 161–175.

[5] K. Jugdev, C. Markowski, T. Mengel, Using the debate as a teaching tool in the online classroom, Online Classroom (2004) 4–7.

[6] R. Luckin, W. Holmes, M. Griffiths, L. B. Forcier, Intelligence unleashed: An argument for ai in education (2016).

[7] R. K. Hambleton, H. Swaminathan, Item response theory: Principles and applications, Springer Science & Business Media, 2013.

[8] S. AlKhuzaey, F. Grasso, T. R. Payne, V. Tamma, Text-based question difficulty prediction: A systematic review of automatic approaches, International Journal of Artificial Intelligence in Education (2023) 1–53.

[9] L. Benedetto, A quantitative study of nlp approaches to question difficulty estimation, in: International Conference on Artificial Intelligence in Education, Springer, 2023, pp. 428–434.

[10] L. Benedetto, P. Cremonesi, A. Caines, P. Buttery, A. Cappelli, A. Giussani, R. Turrin, A survey on recent approaches to question difficulty estimation from text, ACM Computing Surveys 55 (2023) 1–37.

[11] D. J. Chamberlain, R. Jeter, Creating diagnostic assessments: Automated distractor generation with integrity, Journal of Assessment in Higher Education 1 (2020) 30–49.

[12] R. Rodriguez-Torrealba, E. Garcia-Lopez, A. Garcia-Cabot, End-to-end generation of multiple-choice questions using text-to-text transfer transformer models, Expert Systems with Applications 208 (2022) 118258.

[13] B. Ghanem, A. Fyshe, Disto: Evaluating textual distractors for multi-choice questions using negative sampling based approach, arXiv preprint arXiv:2304.04881 (2023).

[14] S. K. Bitew, J. Deleu, C. Develder, T. Demeester, Distractor generation for multiple-choice questions with predictive prompting and large language models, arXiv preprint arXiv:2307.16338 (2023).

[15] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser,

G. Groh, S. Günnemann, E. Hüllermeier, et al., Chatgpt for good? on opportunities and challenges of large language models for education, Learning and individual differences 103 (2023) 102274.

[16] A. Caines, L. Benedetto, S. Taslimipoor, C. Davis, Y. Gao, O. Andersen, Z. Yuan, M. Elliott, R. Moore, C. Bryant, et al., On the application of large language models for language teaching and assessment technology, arXiv preprint arXiv:2307.08393 (2023).

[17] J. Jeon, S. Lee, Large language models in education: A focus on the complementary relationship between human teachers and chatgpt, Education and Information Technologies 28 (2023) 15873–15892.

[18] V. Bayer, M. Hlosta, M. Fernandez, Learning analytics and fairness: do existing algorithms serve everyone equally?, in: Artificial Intelligence in Education: 22nd International Conference, AIED 2021, Utrecht, The Netherlands, June 14–18, 2021, Proceedings, Part II, Springer, 2021, pp. 71–75.

[19] A. Al-Zawqari, G. Vandersteen, Fairness in predictive learning analytics: A case study in online stem education, in: 2023 IEEE Frontiers in Education Conference (FIE), IEEE, 2023, pp. 1–5.

[20] P. H. Coombs, et al., New paths to learning for rural children and youth: Nonformal education for rural development. (1973).

[21] I. Ivanova, Non-formal education: Investing in human capital, Russian Education & Society 58 (2016) 718–731.

[22] A. Karasavvoglou, P. Polychronidou, G. Tsirigotis, L. Tsourgiannis, Non formal and formal learning and the role of higher education institutions, in: 2011 Proceedings of the 22nd EAEEIE Annual Conference (EAEEIE), IEEE, 2011, pp. 1–4.

[23] D. E. Egbezor, B. Okanezi, Non-formal education as a tool to human resource development: An assessment, International journal of scientific research in education 1 (2008) 26–40.

[24] P. Rahabav, T. R. Souisa, Evaluation of non-formal education management in maluku province, indonesia., International Journal of Evaluation and Research in Education 10 (2021) 1395–1408.

[25] C.-H. Yang, E. Rusli, Using debate as a pedagogical tool in enhancing pre-service teachers' learning and critical thinking., Journal of International Education Research 8 (2012) 135.

[26] C. Park, C. Kier, K. Jugdev, Debate as a teaching strategy in online education: A case study, Canadian Journal of Learning and Technology/La revue canadienne de l'apprentissage et de la technologie 37 (2011).

[27] E. T. Mitchell, Using debate in an online asynchronous social policy course., Online learning 23 (2019) 21–33.

[28] R. A. Mercadante Jr, Formal debate as a pedagogical tool in the college classroom. (1988).

[29] M. Sallam, Chatgpt utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns, in: Healthcare, volume 11, MDPI, 2023, p. 887.

[30] V. A. Hamaniuk, The potential of large language models in language education, Educational Dimension 5 (2021) 208–210.

[31] S. Khurana, A. Ali, Qcri advanced transcription system (qats) for the arabic multi-dialect broadcast media recognition: Mgb-2 challenge, in: 2016 IEEE Spoken Language Technology Workshop (SLT), IEEE, 2016, pp. 292–298.

[32] M. M. Khader, A. G. Al-Sharafi, H. Sioufy, W. Zaghouani, A. Al-Zawqari, Munazarat 1.0: A corpus of Arabic competitive debates, in: The 6th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation (OSACT 2024)@ LREC 2024, European Language Resources Association (ELRA), accepted 2024.

[33] A. G. Al-Sharafi, M. M. Khader, M. Ahmed, M. H. Al-Sioufy, W. Zaghouani, A. Al-Zawqari, A hybrid annotation model for Arabic argumentative debate corpus, in: The Eighth International Conference on Arabic Language Processing, ICALP 2023, Rabat, Morocco, April 19–20, 2024, Springer, accepted 2024.

[34] J. Howard, S. Ruder, Universal language model fine-tuning for text classification, arXiv preprint arXiv:1801.06146 (2018).

[35] G. Inoue, B. Alhafni, N. Baimukan, H. Bouamor, N. Habash, The interplay of variant, size, and task type in Arabic pre-trained language models, in: Proceedings of the Sixth Arabic Natural Language Processing Workshop, Association for Computational Linguistics, Kyiv, Ukraine (Online), 2021.

[36] A. Elmadany, E. M. B. Nagoudi, M. Abdul-Mageed, Orca: A challenging benchmark for arabic language understanding, arXiv preprint arXiv:2212.10758 (2022).

[37] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, in: The Eleventh International Conference on Learning Representations (ICLR), 2023.

[38] S. Merity, N. S. Keskar, R. Socher, Regularizing and optimizing lstm language models, arXiv preprint arXiv:1708.02182 (2017).

[39] M. B. Zafar, I. Valera, M. G. Rogriguez, K. P. Gummadi, Fairness constraints: Mechanisms for fair classification, in: Artificial intelligence and statistics, PMLR, 2017, pp. 962–970.

[40] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, H. Wallach, A reductions approach to fair classification, in: International Conference on Machine Learning, PMLR, 2018, pp. 60–69.

[41] A. Al-Zawqari, A. G. Al-Sharafi, M. Ahmed, M. M. Khader, G. Vandersteen, Classifying persuasion modes in arabic debates: A preliminary language model-based analysis, in: The Eighth International Conference on Arabic Language Processing, ICALP 2023, Rabat, Morocco, April 19–20, 2024, Springer, accepted 2024.