From Simple Detection to Quality-Aware Prediction: Exploring Argument Complexity with Machine Learning

Anvi Alex Eponon, Muhammad Tayyab Zamir, Lemlem Kawo Eyob, Luis Israel Ramos Perez, Ildar Bartyrshin, Grigori Sidorov, Olga Koleniskova, Alexander Gelbukh

Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC), Mexico City, Mexico
epononanvialex@gmail.com, tayyab.awan8001@gmail.com, lemlemeyob19@gmail.com, lramos2020@cic.ipn.mx, kolesolga@gmail.com, sidorov@cic.ipn.mx, batyr1@cic.ipn.mx, gelbukh@cic.ipn.mx

Abstract. Argument mining is a critical area within artificial intelligence with significant implications for the future of machine learning models. It is widely believed that advances in argument mining will enhance the ability of models to construct more effective arguments in diverse contexts, including educational and political settings. However, existing research predominantly focuses on identifying argument structures without sufficiently considering the nuanced quality dimensions inherent within them. This study addresses this gap by conducting several experiments. Firstly, it evaluates the performance of traditional machine learning models in detecting arguments. Subsequently, the research investigates how selected quality dimensions impact the performances of argument prediction. The methodology leverages BM25 features with a Random Forest model, achieving notable results with an F1-score of 0.88 and a Spearman's correlation coefficient of 0.73. These outcomes surpass those of previous models such as IBM's 2019 Arg-ranker and base-Arg-ranker, which utilized Bert embeddings and achieved Spearman's scores of 0.41 and 0.42 respectively.

Keywords: Argument mining, Machine learning, Deep learning, Argument quality assessment

1 Introduction

An argument in general is a combination of sentences or paragraphs that tries to convey a reason or many reasons to specific conclusions. By so doing, an argument can be seen as a 'system of reasoning' for providing or arriving at a particular state, being logical, dialectical, rhetorical, true, false, good, or bad. In Linguistics and computer science, Argument mining becomes crucial for machines to understand the real reasoning behind the human language. The works on arguments theoretically and philosophically speaking have been significantly influenced by Aristotle [26].

In almost every setting in the society, arguments are used. In political debates, online discussions via social media, educational settings online product reviews, or even in written books either scientific or fictional, arguments are presented. So Argument mining becomes more and more important in the field of artificial intelligence specifically in the area of Natural language understanding as it helps spread light on how humans reason to communicate effectively through language. Thus several studies have been done in artificial intelligence concerning this task.

It is worth noting that even though many research studies have been conducted on argument mining, which is the field of artificial intelligence aiming at automatically detecting and extracting argumentative structures and their relations from text, several challenges still need to be uncovered. One of them is the identification and the impact of dimension qualities of arguments in various settings or domains of applications. A good explanation is that the Logic dimension is the one which is the most used in scientific settings such as mathematics where coherence is more important than finding truth while the Dialectical dimension which appears mostly in social avenues deals with finding the truth, what should be acceptable, agreed on or not, etc. It appears then that studying the argument dimensions and their qualities through Natural Language Processing (NLP) techniques will enhance the understanding of argumentation in general for machines but also for specific cases, which could impact positively the way machines model the reasoning behind the human language.

The current study tries to bring answers to these two specific questions:

- Can traditional machine learning (ML) models identify arguments using state-of-the-art Natural Language Processing techniques such as Best Match 25 (BM25)?
- Given the specific quality dimensions from the dataset, can they enhance the performance of Deep Learning models in discriminating arguments compared to traditional ML? (This question addresses both the evaluation of the dataset and the identification of the structural complexity of arguments.)

2 Literature review

Argument mining can be defined as the action of identifying and extracting the structure of an argument in natural language and the inferences and reasons behind it. This way, knowing argumentative structures, an understanding is built not only from where people stand but also the reasons they have for doing so. This is useful in several contexts, ranging from the prediction of financial markets to public relations [10], also it has been applied in political debates, online discussions, and customer reviews [17,18].

Argument mining has been a major topic in the Natural language processing literature [14]. In several domains, Argument mining has been explored. For example, argumentation in learning has been found to have the effect of enhancing argumentation skills among students, and computational models of argumentation have been synthesized to enhance this process [12].

One aspect of argument mining is the use and identification of dimensions present in arguments that make them strong or not for a particular purpose. A lot of dimension qualities have been studied and designed, but one of the most recurrent in the literature are rhetorical, logical, and dialectical quality [15]. Analyzing arguments is considered essential for understanding public discourse and enhancing critical thinking skills [1]. Different techniques and models are employed for argument mining and argument quality analysis.

In fact, in [15], twelve qualities have been identified related to the Logic dimensions of arguments but only three relate more with Logic which are Cogency, Fallaciousness, and Strength [19,20,21], only one to Rhetoric, which is Effectiveness [22] and finally three to Dialectic, which are Convincingness, Reasonableness, and Global sufficiency [22,23,24]. A total of 25 qualities have been identified in this research related to only three dimensions.

For instance, in the research [2], authors discuss three methods for extracting the argumentative structure from a piece of natural language text. The first method uses discourse indicators to determine argumentative relationships between nearby propositions in a text. The second method uses topic changes to classify argument components and identify their relationships with supervised machine learning. The last method is concerned with the capability of combining all these individual techniques to enhance argument structure identification.

In this paper, the authors report the first complete work on computational argumentation quality in natural language. They summarize the broad range of existing theories and approaches for considering the logical, rhetorical, and dialectical quality aspects, out of which taxonomy is developed systematically. It also contributes 320 argumentation cases that have been annotated for all of the 15 dimensions, for instance, Cogency, Local relevance, Local sufficiency, Well Formedness, Effectiveness, Arrangement, appropriateness of style, Convincingness, Global acceptability, Reasonableness. The research findings provide the basis for comparison for research on computational approaches to argument quality assessment [11].

Another study explores current NLP feedback systems by categorizing each into four important dimensions of feedback: The four major areas for improvement are richness, visualization, interactivity, and Personalization. Each of the dimensions is also reviewed in terms of its drawbacks, and recommendations for feeding and explanation are given with the aim of developing users' critical thinking capabilities [12].

In the work of [3], argument relevance is analyzed based on user perception. This paper attempts to make the first study on this dimension to establish the foundation for the future advancement of the technology the authors reviewed over 300,000 arguments using four retrieval models across forty topics on twenty controversial issues, considering both biased and neutral perspectives.

However, few works in NLP have been done on the importance of dimension qualities cited earlier in the prediction of arguments on different settings or domains. Most of them focused on the overall argument detection itself, or its structures.

In the study [4], the authors state that BERT outperforms most baselines for modeling causational hierarchies in typical argument structures within online discourse. This model generates embeddings, which are then processed through a transformer encoder layer to identify edges between them.

Another study proposes the creation of a written corpus for argumentative reasoning, analyzed with advanced argumentation techniques, and marked up using an open, reusable language. It highlights how this resource can be used in linguistic, computational, and philosophical research and also discusses its role in initiating a program for automatic detection of argumentative structure [5].

The advancement of artificial intelligence also benefits argument mining with the use of deep learning models and large language Models (LLMs).

The work [6] involves using LLMs as argument quality annotators and evaluating the agreement between LLMs, human experts, and novices based on argument quality dimensions. LLMs show moderate agreement with experts and improve inter-annotator consistency, proving valuable for automated argument quality assessment of large datasets.

[7] carried out a review of the literature on argument quality and suggests using instruction-following LLMs for assessment, stresses systematic training with argumentation theories and examples, and discusses practical implementation, including benefits and moral considerations.

In [8] the researchers describe the first dialogue conference competition for recognizing argumentation analysis of Russian language texts. It included a stance detection task and argument classification with a dataset of 9,550 comments gathered from various social media platforms regarding COVID-19 topics. The presented NLI-BERT-TargetMask obtained F1-scores of 0. 6968 and 0.7404 for stance detection and argument classification in particular.

The research [9] proposes (What Is Being Argued?) WIBA is a new framework to address what is being argued in a range of settings. Their approach identifies the existence of the argument, its topic, and its stance, using the fine-tuning of LLMs. They get an F1 of between 79%-86%, the method of identifying topics gets an average of 71% similarity, and the Stance Classification method gets 71%-78% F1. The authors concluded that WIBA facilitates analysis of the arguments in large contexts and across the domains of linguistics, communication, social, and computer sciences.

Finally, in the work of [13], an end-to-end approach is proposed for jointly predicting all predicates, argument spans, and the relations between them. The model independently determines what relationship, if any, exists between every possible word-span pair and learns contextualized span representations that offer rich, shared input features for each decision.

In paper [14], the data used are the discourse of students and annotations that were obtained from the Kaggle platform. They use DeBERTa for predicting effective arguments. The lowest of the metric is achieved by the Deberta-large which owns 0.619 among these models, which is 0.007, 0.114, and 0.030 lower than BERT, and RoBERTa respectively.

As observed, the literature on argument mining has less focus on the importance of dimension qualities of arguments and how they impact the strengths of arguments in different settings. The objectives of such previous research were to theoretically identify the quality dimensions without evaluating the impact of their presence in identifying arguments. The current study aims to introduce a series of studies that aim to present with NLP techniques the performances on the identification of argument dimension qualities and their impact on arguments. This ablation approach mainly missing in previous studies focusing on identifying dimension qualities brings insights into the importance of the quality dimensions features in argument mining.

3 Methodology

3.1 Assumptions and task objectives

The methodology designed for the current study aligns with the objectives and assumptions made. To conduct the experiments, some assumptions were made regarding the nature of an argument and how it could be identified. Throughout the experiment, we assumed that a good argument has a well-defined structure (either inductive or deductive related in part, to a syntactic nature) but also aligns with the understanding of the target audience (contextual nature). Not only that, we defined a "good" argument as a sentence or a paragraph that contains a conclusion or an opinion related to illustrations (personal or general) and is possibly supported by regulation facts.

This definition does not attribute any Truthfulness to arguments, in other sense that in our study we don't assess an argument as being "good" because it is accepted as "true" but rather if the statement given earlier has a coherence between the conclusion, illustrations, and regulations. This assumption is supported by the fact that any argument can be "strong" or "good" without being necessarily "true" as the notion of "Truth" can be ambiguous. An example can be observed in Legal statements such as:

"The defendant should be acquitted because there is no conclusive evidence linking them to the crime.". - Refers to the legal term Acquittal (US Legal Terms Glossary)

The argument stated is strong with respect to the legal context, and guilt in this context, must be proven beyond reasonable doubt. However, The truth of the actual involvement of the defendant in the crime remains uncertain; the argument is based on the current legal standard rather than an objective truth about the guilt of the defendant or innocence where this standard can change from one culture to another.

Then, the study is divided into two tasks. The first one is to implement and evaluate the performances of Machine learning models in predicting arguments on several topics either in political debates or online review quality assessments

with a leading question: Can Machine learning models identify easily and effectively arguments from statements that are not (considered from the annotators point of view) using Natural Language Processing techniques?

The second task has the purpose of evaluating deep learning models limited to Bilstm and CNN models but this time with combinations of features related to dimension qualities of arguments to analyze if the selected qualities in the datasets help identify arguments or not or if their presence or not impact the strength of the argument.

3.2 Dataset and Exploratory data analysis (EDA)

Dataset

To conduct the overall study, two datasets were used. One from the IBM Debater datasets [14] was made specifically for Argument quality. This dataset contains more than 34,000 samples of arguments focusing on identifying the better ones. And the second dataset is from Gaqcorpus [15].

In the current experiment, we used a portion of the IBM Debater datasets containing a bit more than 23,000 samples. The dataset was presented during the EMNLP conference led in 2019 which contains 5 times more samples than the UKPRank dataset [14].

The whole Gaqcorpus dataset containing 6,424 samples was used in the study. This dataset [15] fills the gap by bringing a large-scale (more than 5000 arguments) English multi-domain corpus (Debate forums, Community Question Answering, Reviews) annotated with a theory-based Argument-quality score.

EDA (GaqCorpus)

The GaqCorpus dataset introduces a textual English corpus of arguments in several domains such as debate forums, review forums, or community QA forums. The dataset is comprised of 6,424 premises and conclusions associated with quality features such as degree of:

- Logic;
- Dialectic;
- Rethoric;
- Relevance.

The dataset contains 4,873 considered arguments and only 513 considered non-arguments. Which creates a fair imbalance dataset for the binary prediction of arguments.
Out of the arguments, 3,442 relevant arguments have been identified (threshold put on relevance ≥ 3 out of 4). The relevant arguments are dispatched as follows (Fig.2 and Fig.1):

- 3,388 logical arguments;

– 3,122 dialectical arguments;
– 3,212 rhetorical arguments.

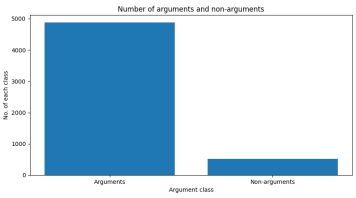


Fig. 1: Argument vs non-argument proportions

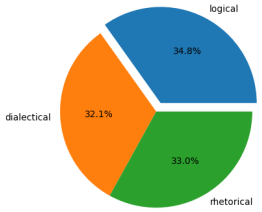


Fig. 2: Proportion of dimensions

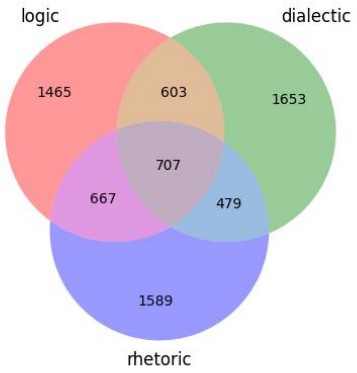


Fig. 3: Overlap between dimensions

3.3 Task 1: Binary Argument prediction from premises only

Task 1 is about finding if traditional machine learning models can effectively predict arguments given a premise. For this end, we used a portion of the IBM Debater dataset which contains +23,000 samples of pairs of arguments which is approximately 67.94% of the full dataset. The final dataset after preprocessing contains two columns. The premise and the label. To handle this experiment, two separate experiments have been done. The first one involves training traditional machine learning models with TF-IDF features, and the second one on BM25 features without making any preprocessing on the textual data with

the assumption that the original structure of the text is crucial to identifying arguments.

Feature extraction phase

During this phase, only two techniques were selected. The first one commonly used is the Term Frequency Inverse Document Frequency (TF-IDF). Even though in the current datasets, premises are small in size compared to essays or political speeches, applying TF-IDF to the datasets, will help the models identify patterns or important words related to the identification of arguments.

The second feature extraction technique used in this part of the experiment is the Best Matching 25 (BM25) which has been proven excellent in the literature as a ranking function [16].

Model selection and experimental phase

1- Model selection

For this task, four models were used which are:

- Random Forest;
- Logistic Regression;
- Support Vector Machine (SVM);
- Naive Bayes.

Random Forest has been proven efficient in several tasks of classification in Machine learning. Due to its capacity to learn implicit features from different sub-trees and also its robustness to overfitting, Random Forest has been chosen.

Due to the binary nature of the task, logistic regression has been chosen. Its efficiency over large datasets and capacity to differentiate between two classes with a sigmoid function make it suitable for our experiments.

Support Vector Machine algorithm has been chosen in this task first, for its performance to overfitting like Random Forest, but also for its ability to handle non-linearity over the features that happen on complex textual datasets.

Lastly, Naive Bayse has been selected for the experiment due to its ability to handle many features or large vocabulary.

2- Experimental phase

1- Feature extraction hyperparameters

Concerning the feature extractions used, all the models used TF-IDF features on n-grams varying from 1 to 3. Additionally, after experiments, we set the number of estimators for the random forest model at 200 and set the kernel parameter

of the SVM model to linear. Below is the resume of the best parameters used for each model:

Models	**Best Hyperparameters**
Logistic Regression	Only TF-IDF (1-3 grams)
Naive Bayes	Only TF-IDF (1-3 grams)
SVM	Classifier C: 1, Kernel: Linear
Random Forest	n_estimators: 200

Table 1: Best hyperparameters for different models

Training Phase

The models selected were implemented from the sci-kit-learn library. Most of the parameters have been left by default except the ones mentioned earlier to better measure the performance of the models.

3.4 Task 2: Binary predictions from quality dimensions and premises

Task 2 has been conducted by using solely the dataset provided by Gaqcorpus [15]. This dataset contains +6,000 samples of arguments scored based on their dimension qualities. This task also tries to understand if deep learning models perform better on argument detection but also if the presence of the selected dimension qualities impacts this detection. To answer these questions, the study has been divided into two experiments too each related to one model, a Bidirectional Long-Short Term-Memory (BiLSTM) model and a Convolutional Neural Network (CNN).

Preprocessing phase

The preprocessing phase of the experiments made with the BiLSTM models involves a few transformation steps of the texts such as:

- Remove of English stop words;
- Lemmatization of tokens;
- Part of Speech Tagging.

However, only the removal of stopwords has been done on the second BiLSTM model in order to see the impact of syntactic processing on the performance of the models.

Concerning the CNN models, the first one has been trained on the quality dimensions of the arguments while the second has been trained solely with the premises to detect whether the qualities give an advantage for the predictions of arguments.

Feature extraction phase We used one deep word embedding model for all the models selected which is SentenceBert [25]. The deep word embedding model used is from the sentenceBert library.

Model selection and experimental phase

1- Model selection

For this task, two models were used which are:

- BiLSTM;
- CNN.

Bidirectional Long-Short Term-Memory(LSTM) is well suited for the second task due first to its capability to handle sequential input data. It has a longer memory dependency compared to the model used in the previous task but also with its ability to have a combination of context from both directions of the sequence, it can capture more detailed features related to context.

On the other hand, CNN also proved to be efficient in Hierarchical feature learning where in the context of argument mining this capability is crucial.

2- Experimental phase

1- Feature extraction hyperparameters

Below, are the best hyperparameters used for the training of the models:

Models	**Epochs**	**k-folds**	**Batch Size**	**Learning Rate**
BiLSTM 1	10	5	64	0.001
BiLSTM 2	10	0	64	0.001
CNN Numeric	10	0	64	0.001
CNN textual	10	5	64	0.001

Table 2: Experimental setup for different models

4 Results

4.1 Task 1 results

The research has been conducted through several experiments. The first Task, which is comprised of baseline experiments and improved baseline considered four traditional machine learning models as mentioned in the methodology. The baseline experiments have been conducted on the 23,000 samples while the baseline improved using BM25 used half of the dataset, which is approximately 11,000 samples. Below are the results concerning the baseline experiments:

Model	F1	Spearmanr
Logistic Reg	0.7741	0.5452
Naive Bayes	0.7775	0.5474
SVM	0.7749	0.5464
Random Forest	**0.7706**	**0.5391**

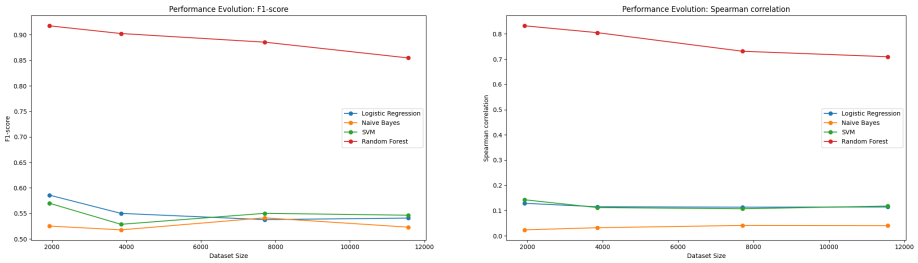
Table 3: Model Performance Comparison on 23,000 samples

The best results of the second part of Task 1 also called improved baseline are recorded in the table below:

	F1	Spearmanr
Logistic Reg	0.5378	0.1135
Naive Bayes	0.5413	0.0412
SVM	0.5501	0.1072
IBM Arg-ranker		0.42
Random Forest	**0.8855**	**0.731**

Table 4: Performance Comparison - 7,709 samples

The figure below shows the evolution of the models according to the increase in the sample size.



(a) F1-score Evolution (b) Spearman Correlation Evolution

Fig. 4: Performance Evolution using BM25

4.2 Task 2 results

Task 2 considered all the 23,000 samples of arguments to train two Deep learning models (CNN, BiLSTM). The table below displays the results of the models:

Model	F1	Spearmanr
BILSTM_1	**0.8812**	**0.7876**
BILSTM_2	0.8812	0.7876
CNN_numeric	0.8812	0.7816
CNN_textual	**0.8815**	**0.7881**

Table 5: Model Performance Comparison of Deep Learning Models

A last ablation experiment has been conducted on the BiLSTM model in order to uncover the real impacts of the dimensions selected in the datasets. Below are presented the results at the last fold of the training using either one of the dimensions for predictions:

Feature Excluded	Accuracy	F1-score
cogency_(logic)	0.7875	0.8811
effectiveness_(rhetoric)	**0.7876**	**0.8812**
reasonableness_(dialectic)	0.7876	0.8812

Table 6: Ablation experiments on dimension qualities

5 Discussions

As mentioned earlier in the current study, the objectives are to discover if traditional machine learning models could effectively predict arguments given

premises and finally evaluate deep learning models on predicting arguments based on several quality dimensions. In the first task, specifically in the baseline experiments, all traditional models perform just above luck concerning the Spearmanr score and at more than 0.70 f1 scores, beating the IBM Arg-ranker-base model with a minimum Spearmanr score of 0.53.

These results, at the initial step of the experiment, might be due to the size of the dataset, where 23,000 samples were used compared to approximately 6,000 samples in the case of the IBM Arg-ranker-base model. However, this initial baseline does not make use of any contextual embedding such as Bert, rather relies on bag-of-word techniques specifically TF-IDF.

The second experiment in task 1 which is the baseline improved, makes use of the BM25 ranking function in order to predict arguments. Due, to the heavy computational resources needed, only half of the dataset was used, thus approximately 11,000 samples.

Except for the Random Forest model, all the models perform at luck gradually decreasing in performance with the increase of the samples. This highlights the complexity of the features embedded in the premises since all the models were trained on their best parameters. The performance that can be compared with the IBM Arg-Ranker obtained at samples equal to 7,709 where the Random Forest performs at 0.88 of f1-score with 0.73 for the Spearmanr score, which performs better than the IBM Arg-Ranker-based but also the IBM Arg-ranker which were trained on vanilla Bert and finetuned Bert embeddings. This performance of the Random Forest model might be due to its capability to detect and associate specific structures to arguments from the subtrees.

On the other hand, the experiments on the deep learning models show that deep learning models are much more stable in predicting arguments either from premises only or with quality dimensions. This is observed by the constant f1 score turning around 88% at each epoch and fold experiment in the ablation experiment. However, from those same experiments, the presence of the quality dimensions does not influence significantly the identification of arguments. This could be due to the fact that most of the dimensions scores overlap as seen in the EDA study, which might indicate a limit in the annotation process.

Conclusion

The task of argument mining even though presenting several interests in the literature remains a task with several challenges in NLP. Constructing models that leverage understanding of the human language to generate correct arguments can influence several sectors of society.

In the present work, experiments showed that the identification of arguments can be effectively done by traditional models with the correct feature extractions such as BM25 ranking functions. However, if deep learning models such as BiLSTM and CNN can be more stable and capture more complex hidden features the question of which quality dimensions impact this prediction is still unanswered.

The development of dimensions and quality dimensions in argument has been a serious topic since the Ancient Greeks. Finding an automatic approach to learning the correct argument construction will push forward the performances of future models. Finally, in our study, Random Forest performs the best at 0.88 for the f1-score and 0.73 for the Spearmanr score with approximately 7,000 samples of arguments which creates a new baseline surpassing the baseline proposed by IBM Args-ranker which lies at 0.42 for the Spearmanr score.

Limit of the study and future work

The current study presents itself as an introduction to a series of experiments to be conducted in argument mining specifically in the modeling of dimension qualities. Thus it has been done with a lot of limitations. The first and main one is the experiments focused on the influence of dimension qualities on the predictions of arguments regardless of the specific domain in which the arguments have been constructed. Given that different fields have their own rules and methods for making arguments, the current approach struggles to distinguish between these different characteristics specific to each field. For example, the way arguments are constructed in Political debate and online reviews are not the same. From this perspective, how does understanding the unique qualities of each field impact not only the prediction of the argument but also the understanding of the domain? Alternatively, how can knowledge of the domain help in predicting the quality dimensions present in a given argument? Those are questions the current research is not answering.

The second limit lies in the Gaqcorpus dataset itself. The annotating procedure ended up with qualities that significantly overlap (Fig.3). This brings a lot of ambiguities in differentiating the three dimension qualities. A better approach to annotating such information might need to be addressed. Also, using only two datasets, may not fully represent the diversity of argument structures and qualities across different domains. Thus a construction of a diverse argument dataset covering different languages can be an adequate future avenue. Finally, the development of NLP techniques such as tokens or sets of n-tokens specifically targeting dimension qualities in an argument could present several advantages in effectively detecting the dimension qualities.

Acknowledgments

The work was done with partial support from the Mexican Government through the grant A1- S-47854 of CONACYT, Mexico, grants 20241816, 20241819, and 20240951 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico, and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

References

1. Yang, J., Du, X., Hung, J.L. and Tu, C.H., 2022. Analyzing online discussion data for understanding the student's critical thinking. *Data Technologies and Applications*, 56(2), pp.303-326.
2. Lawrence, J. and Reed, C., 2015, June. Combining argument mining techniques. In *Proceedings of the 2nd Workshop on Argumentation Mining* (pp. 127-136).
3. Potthast, M., Gienapp, L., Euchner, F., Heilenkötter, N., Weidmann, N., Wachsmuth, H., Stein, B. and Hagen, M., 2019, July. Argument search: Assessing argument relevance. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1117-1120).
4. Srivastava, P., Bhatnagar, P. and Goel, A., 2022, December. Argument mining using BERT and self-attention based embeddings. In *2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)* (pp. 1536-1540). IEEE.
5. Reed, C., Mochales Palau, R., Rowe, G. and Moens, M.F., 2008, January. Language resources for studying argument. In *Proceedings of the 6th conference on language resources and evaluation-LREC 2008* (pp. 2613-2618). ELRA.
6. Mirzakhmedova, N., Gohsen, M., Chang, C.H. and Stein, B., 2024. Are Large Language Models Reliable Argument Quality Annotators?. arXiv preprint arXiv:2404.09696.
7. Wachsmuth, H., Lapesa, G., Cabrio, E., Lauscher, A., Park, J., Vecchi, E.M., Villata, S. and Ziegenbein, T., 2024. Argument Quality Assessment in the Age of Instruction-Following Large Language Models. arXiv preprint arXiv:2403.16084.
8. Kotelnikov, E., Loukachevitch, N., Nikishina, I. and Panchenko, A., 2022. RuArg-2022: Argument mining evaluation. arXiv preprint arXiv:2206.09249.
9. Irani, A., Park, J.Y., Esterling, K. and Faloutsos, M., 2024. WIBA: What Is Being Argued? A Comprehensive Approach to Argument Mining. arXiv preprint arXiv:2405.00828.
10. Guerraoui, C., Reisert, P., Inoue, N., Mim, F.S., Singh, K., Choi, J., Robbani, I., Naito, S., Wang, W. and Inui, K., 2023, December. Teach Me How to Argue: A Survey on NLP Feedback Systems in Argumentation. In *Proceedings of the 10th Workshop on Argument Mining* (pp. 19-34).
11. Wachsmuth, H., Naderi, N., Hou, Y., Bilu, Y., Prabhakaran, V., Thijm, T.A., Hirst, G. and Stein, B., 2017, April. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers* (pp. 176-187).
12. He, L., Lee, K., Levy, O. and Zettlemoyer, L., 2018. Jointly predicting predicates and arguments in neural semantic role labeling. arXiv preprint arXiv:1805.04787.
13. Tang, T., 2022, September. Predicting Effective Arguments with A Natural Language Processing Model. In *2022 2nd International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI)* (pp. 436-439). IEEE.
14. Toledo, A., Gretz, S., Cohen-Karlik, E., Friedman, R., Venezian, E., Lahav, D., Jacovi, M., Aharonov, R., & Slonim, N., 2019. Automatic Argument Quality Assessment - New Datasets and Methods. arXiv preprint arXiv:1909.01007.
15. Lauscher, A., Lapesa, G., Cabrio, E., Lauscher, A., Park, J., Vecchi, E.M., Villata, S. and Ziegenbein, T., 2020. Rhetoric, Logic, and Dialectic: Advancing Theory-

based Argument Quality Assessment in Natural Language Processing. In *COLING 2020*.
16. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M., & Gatford, M., 1994. Okapi at TREC-3. *Text Retrieval Conference*.
17. Chakrabarty, T., Hidey, C., Muresan, S., McKeown, K., & Hwang, A., 2019. AMPERSAND: Argument Mining for PERSuAsive oNline Discussions. arXiv preprint arXiv:2004.14677.
18. McCloskey, B.J., LaCasse, P.M. & Cox, B.A., 2024. Natural language processing analysis of online reviews for small business: extracting insight from small corpora. *Ann Oper Res*. https://doi.org/10.1007/s10479-023-05816-2
19. Johnson, R.H., & Blair, J.A., 1983. Logical self-defense. http://ci.nii.ac.jp/ncid/BA26931398
20. Damer, T.E., 1980. Attacking Faulty Reasoning: A Practical Guide to Fallacy-Free Arguments. http://ci.nii.ac.jp/ncid/BA80518138
21. Govier, T., 1985. A practical study of argument. http://ci.nii.ac.jp/ncid/BB16391451?l=en
22. Aikin, S.F., 2008. Perelmanian Universal Audience and the Epistemic Aspirations of Argument. *Philosophy & Rhetoric*, 41(3), 238-259.
23. Andone, C., 2005. A Systematic theory of argumentation: the Pragma-Dialectical Approach. *Journal of Pragmatics*, 37(4), 577–583. https://doi.org/10.1016/j.pragma.2004.07.003
24. Cohen, J., 2001. Defining identification: A theoretical look at the identification of audiences with media characters. *Mass Communication & Society*, 4(3), 245–264. https://doi.org/10.1207/S153278C25MCS0403_01
25. Reimers, N., & Gurevych, I., 2020. Making Monolingual Sentence Embeddings Multilingual Using Knowledge Distillation. arXiv preprint arXiv:2004.09813.
26. Bench-Capon, T.J., & Dunne, P.E., 2007. Argumentation in artificial intelligence. *Artif. Intell.*, 171, 619-641.