Argumentation Quality Assessment: an Argument Mining Approach

Santiago Marro, Elena Cabrio, Serena Villata

ARGUMENTATION QUALITY ASSESSMENT: AN ARGUMENT MINING APPROACH

SANTIAGO MARRO

Université CôteD'Azur, CNRS, Inria, I3S

smarro@unice.fr

ELENA CABRIO

Université CôteD'Azur, CNRS, Inria, I3S

SERENA VILLATA

Université CôteD'Azur, CNRS, Inria, I3S

Abstract

Argumentation is used by people both internally, by evaluating arguments and counterarguments to make a decision, and externally, e.g., by exchanging arguments to reach an agreement or to promote a position. A major component of the argumentation process concerns the assessment of a set of arguments and of their conclusions in order to establish their justification status, and therefore compute their acceptability degree. The assessment of the justification status of the statements supported by arguments allows the agent to decide what to believe and what to do. Argumentation semantics provide formal criteria to determine which sets of arguments (i.e., extensions) can be regarded as collectively acceptable (Baroni, Caminada, and Giacomin 2011). However, the assessment of the arguments acceptability is only a (basic) part of the complex assessment tasks required in argumentative processes in many everyday life applications, e.g., in medicine and education.

Assessing argumentation is a crucial issue in the context of artificial argumentation, encompassing various aspects such as identifying real natural language arguments and their relations in text, computing the justification status of abstract arguments, and gradually evaluating arguments. While some approaches have tackled the automatic assessment of natural language arguments (Wachsmuth et al. 2017, 2020), this issue remains largely unresolved.

In this paper, we address this open issue and we answer the following research question: what are the basic quality dimensions to characterize natural language argumentation and how to automatically assess them?

More precisely, we propose an Argument Mining (AM) approach to identify and classify natural language arguments along with quality dimensions.

In this work, we decide to characterize argument quality along with three quality dimensions for natural language argumentation, i.e., cogency, rhetoric, and reasonableness. The assessment of cogency involves determining the acceptability and sufficiency of the premises that support an argument's conclusion, while rhetoric identifies the use of rhetorical strategies such as ethos, logos, and pathos in the argument's conclusion. Additionally, reasonableness rates whether the argument effectively rebuts counterarguments, assessing the dialectical quality dimension of the argumentation.

Our interest focuses on the education scenario, where students are asked to interact with our AM system to assess the quality of their persuasive essays with respect to these three quality dimensions. To train our AM model, we annotated an existing dataset of 402 student persuasive essays (Stab and Gurevych 2017) with these quality dimensions.

We then propose a new deep learning AM method based on a transformer architecture, exploiting the structure of the argumentation graph through graph embeddings. Our approach automates the evaluation process proposed by Stapleton and Wu (2015) in social science by utilizing a scoring rubric for persuasive writing that combines the assessment of argumentative structural elements and reasoning quality. The obtained results are satisfactory and outperform standard baselines and similar approaches in the literature.

1 Introduction

A major component of the argumentation process concerns the assessment of a set of arguments and of their conclusions to establish their justification status, and therefore compute their acceptability degree (Baroni et al., 2011). Both qualitative and quantitative approaches have been proposed in the literature to assess the acceptance of an argument. However, the assessment of the arguments acceptability is only a (basic) part of the complex assessment tasks required in argumentative processes in many everyday life applications and contexts, e.g., in medicine and education.

The issue of assessing an argumentation is particularly critical when considering the different aspects of artificial argumentation, from the identification of real natural language arguments and their relations in text, to the computation of the justification status of abstract arguments (Baroni et al., 2011), to the gradual assessment of arguments (Amgoud et al., 2022) based, e.g., on the trustworthiness of the argument proponents (da Costa Pereira et al., 2011) or on the value promoted by the argument (Bench-Capon, 2003). Despite some approaches addressing the automatic assessment of natural language arguments (Wachsmuth et al., 2017), this issue remains largely unexplored and unsolved. In this paper, we address this open issue, and we answer the following research questions: *(i)* what are the basic quality dimensions to characterize natural language argumentation? and *(ii)* how to automatically assess these quality dimensions on natural language argumentative text?

More specifically, we propose an argument mining (Cabrio and Villata, 2018; Lawrence and Reed, 2020; Lauscher et al., 2022) approach to identify and classify natural language arguments along with quality dimensions. We first define and annotate three prominent quality dimensions for natural language argumentation, i.e., *cogency*, *rhetoric,* and *reasonableness*, on an existing dataset of student persuasive essays (Stab and Gurevych, 2017). We then train a neural network model classifier empowered with properties from the argument graphs to address the task. Our core contribution is twofold:

- We enrich a linguistic resource of persuasive essays (1908 arguments) with a new annotation layer, i.e., the quality dimensions of *cogency*, *rhetoric,* and *reasonableness*.
- We propose a new model architecture, exploiting the structure of the argument graph through graph embeddings. To the best of our knowledge, this is the first method that combines the graph structure

of the argumentation with the textual content to assess the argumentation quality.

This paper is motivated by the lack of natural language argumentation resources annotated with quality dimensions, and the need for effective methods to address this task. Our contribution offers a novel resource and method to advance the field.

2 Related Work

Recent approaches in Argument(ation) Mining (AM) tackle specific argument qualities features, such as argument relevancy (Wachsmuth, 2017), convincing arguments (Habernal, 2016) and overall argument quality (Toledo, 2019).

Defining the characteristics of a good and successful argument is a hard task. First, we must address the several text rating procedures proposed in the literature. Different factors, such as the aim of the assessment, the freedom given to the raters, and the number of texts to be analyzed should be considered when evaluating the quality of argumentative texts. Following (Coertjens et al., 2017), rating procedures can be classified in two dimensions: Holistic vs. analytic and absolute vs. comparative. Holistic rating entails evaluating texts as a complete entity, while analytic rating involves assessing multiple text features. In absolute ratings, each text is assessed based on a predefined criteria or description, while in comparative ratings, texts are compared to each other to determine their score. In this study, our objective is to assess the quality of argumentative texts in persuasive essays through the application of a consistent and absolute analytic rating system. The aim is to ensure that the evaluation is based solely on the essay's content, rather than the subjective bias of the evaluator, and that the assessment results are consistent across all raters.

A commonly used rating method is rubrics. In an analytic rubric, text features are predetermined, but the weight assigned to each feature may not be predetermined. (Coertjens et al. 2017) found that evaluators may assign different weights to predetermined text features, potentially leading to variations in assessments of a single text among evaluators.

To tackle this issue, (Stapleton and Wu, 2015) describe the weight of the separate text features in a rubric as fixed. In this rubric, the authors stated that a strong

argumentative text is composed of two important elements. *(i)* an argumentative text must be constructed considering all elements contributing to a good *quality of argumentation* and *(ii)* attention must be paid to the *quality of the content* of the text.

Different approaches have been proposed to assess both points from a logical, rhetorical, and dialectical point of view. Wachsmuth (2017), derive a taxonomy of argumentation quality that systematically decomposes quality assessment based on the interactions of 15 widely accepted quality dimensions. The three main characteristics are *Cogency*, *Effectiveness* and *Reasonableness*. As a follow up, Wachsmuth (2020), investigate how effectively each dimension can be automatically assessed, modelling features such as content, style, length, and subjectivity. This text-only assessment yields moderate learning success for most of the evaluated dimensions.

Following the work by (Stapleton and Wu. 2015) and (Wachsmuth et al. 2017), we argue that it is important to evaluate both the quality of argumentation and the quality of the content to provide a complete assessment of the argumentative texts. We, therefore, advance the state of the art of natural language argument quality assessment by investigating three main quality properties of *persuasive essays* (i.e., cogency, reasonableness, and rhetorical strategy) using the rubric provided by (Stapleton and Wu. 2015). Additionally, we propose a novel approach to evaluate argument reasonableness by integrating cogency properties with the argumentation graph structure.

3 Quality dimensions of persuasive essays

To annotate the quality dimensions on persuasive essays, we rely on the corpus built by (Stab and Gurevych, 2017), containing 402 persuasive essays annotated with the argument components (i.e., evidence, claims and major claims) and relations (i.e., support or attack). We add a new annotation layer by manually labelling for each argument in the essays the following three quality attributes: *cogency*, *reasonableness,* and *argumentation rhetoric*.

3.1 Annotation guidelines

Given that our goal is to assess persuasive essays written by students, we rely on an absolute analytic quality evaluation process proposed by (Stapleton and Wu. 2015). The authors propose a scoring rubric for persuasive writing that integrates the assessment of both argumentative structural elements and reasoning quality. This rubric contemplates several characteristics of the standard definition of Cogency and Reasonableness, such as Relevancy, Acceptability, and Soundness as well as the presence of counterarguments and rebuttals. Tables I, II and III show the analytic scoring rubrics proposed by (Stapleton and Wu. 2015). A scale of 0, 10, 15, 20, 25 is given to assess the Cogency and Reasonableness of a given argument.

Definition 1. Cogency *An argument should be seen as cogent if it has individually acceptable premises that are relevant to the argument's conclusion and that are sufficient to draw the conclusion* (Wachsmuth et al., 2017).

Table I. Analytic Scoring Rubric for assessing Cogency (Stapleton and Wu, 2015).

Score 25	***Score 20***	***Score 15***	***Score 10***	***Score 0***
a. Provides multiple reasons for the claim(s), and b. All reasons are sound/acceptable and free of irrelevancies.	a. Provides multiple reasons for the claim(s), and b. Most reasons are sound/acceptable and free of irrelevancies, but one or two are weak.	a. Provides one to two reasons for the claim(s), and b. Some reasons are sound/acceptable, but some are weak or irrelevant.	a. Provides only one reason for the claim(s), or b. The reason provided is weak or irrelevant.	a. No reasons are provided for the claim(s); or b. None of the reasons are relevant to/support the claim(s).

Following Table I, we define the *acceptable* premises as the ones that are worthy of being believed, and the *relevant* one as those that contribute to the acceptance or rejection of the argument's conclusion. These criteria are considered in point (b) (Table I) whilst the structural information about the argument graph is addressed in point (a). Examples 1, 2, 3 show the cogency annotation on three different persuasive essays from (Stab and Gurevych, 2017).

Example 1. We should attach more importance to cooperation during primary education. [**Through cooperation, children can learn about interpersonal skills which are significant in the future life of all students**]$_1$.[*What we acquired from team work is not only how to achieve the same goal with others but more*

importantly, how to get along with others]$_1$. [*During the process of cooperation, children can learn about how to listen to opinions of others, how to communicate with others, how to think comprehensively, and even how to compromise with other team members when conflicts occurred*]$_2$. [*All of these skills help them to get on well with other people and will benefit them for the whole life*]$_3$.

Example 2. Animals should live in natural habitats instead of zoos. [**it is our responsibility to create a natural and safe environment for animals to live in**]$_1$, [*Given the fact that human beings are responsible for the heavy pollution and severe damage to the natural habitats of many wild animals*].$_1$[it is the right of wild species to live in a environment away from human beings].$_2$

Example 3. Television devastate families ties. [**Most of people do not have a plan for make a limitation or schedule for watching television**]$_1$.

The first sentence represents the major claim, while the claim to be assessed is in bold and the premises supporting it are in italics. Example 1 is annotated with cogency score 25, given that the author presents multiple premises which are acceptable and relevant to draw a conclusion. Example 2 shows a cogency score of 15, given that the author presents two premises that are relevant to the topic but not sufficient to draw the conclusion. Finally, Example 3 is annotated with a cogency score 0, given that the author does not presents a premise to support the claim.

Table II. Rubric for Reasonableness Counterargument (Stapleton and Wu, 2015).

Score 25	***Score 20***	***Score 15***	***Score 10***	***Score 0***
a. Provides multiple reasons for the counterargument claim(s), and b. All reasons for the alternative view(s) are sound/acceptable and free of irrelevancies.	a. Provides multiple reasons for the counterargument claim(s), and b. Most reasons for the alternative view(s) are sound/acceptable and free of irrelevancies, but one or two are weak.	a. Provides one to two reasons for the counterargument claim(s), and b. Some reasons for the alternative view(s) are sound/acceptable, but some are weak or irrelevant.	a. Provides only one reason for the counterargument claim(s), or b. The reason for the alternative view is weak or irrelevant.	a. No reasons are provided for the counterargument claim(s); or b. None of the reasons are relevant to/support the counterargument claim(s)/alternative views.

Definition 2. Reasonableness *An argumentation should be seen as reasonable if it contributes to the resolution of the given issue in a sufficient way that is acceptable to the target audience* (Wachsmuth et al., 2017).

The Analytic Scoring Rubric for Reasonableness (Table II and Table III) integrates these concepts and follows the idea of evaluating the argumentation graph with a focus on the counterarguments and their respective rebuttals. (Stapleton and Wu. 2015) separates the evaluation of Reasonableness for two different argumentative components, the counterarguments, and the rebuttals.

Table III. Analytic Scoring Rubric for assessing Reasonableness (Stapleton and Wu, 2015).

Score 25	***Score 20***	***Score 15***
a. Refutes/points out the weakness of all the counterarguments, and b. All rebuttals are sound/acceptable, and c. The reasoning quality of all the rebuttals are stronger than that of the counterarguments.	a. Refutes/points out the weakness of all the counterarguments, and b. Most rebuttals are sound/acceptable, but one or two are weak. c. The reasoning quality of most rebuttals are stronger than that of the counterarguments, while one or two are equal to that of the counterarguments.	a. Refutes/points out the weakness of all the counterarguments, and b. Some rebuttals are sound/acceptable, but some are weak c. The reasoning quality of some rebuttals are stronger than that of the counterarguments, while some are weaker to that of the counterarguments.

Score 10	***Score 0***
a. Refutes/points out the weakness of some counterarguments, or b. Few of the rebuttals are sound/acceptable; most of them are weak, or c. The reasoning quality of most rebuttals are weaker than that of the counterarguments.	a. No rebuttals are provided; or b. None of the rebuttals can refute the counterargument.

The rubric score given to evaluate the reasonableness of the *counterarguments* (Table II) stipulates an analysis on the cogency of the counterargument, providing the same definitions given in Table I. Similarly to *cogency* and *reasonableness counterargument*, an evaluation on the soundness and acceptability of the text is required for the evaluation of the *rebuttals* (Table III). However, it differs from the others when evaluating if *(i)* the rebuttal refutes the counterarguments and *(ii)* does so with a stronger reasoning quality.

To evaluate the Reasonableness of an argument, one must analyze its counterarguments and related rebuttals. In Figure 1, the reasonableness quality dimension is assessed following Tables II and III. Counterargument Claim E receives a Score 0 for Reasonableness Counterargument as no supporting reasons or premises are provided. In contrast, for the rebuttal, Claim F provides a sound premise and stronger reasoning quality than the counterargument, resulting in a Score 25 for Reasonableness Rebuttal.

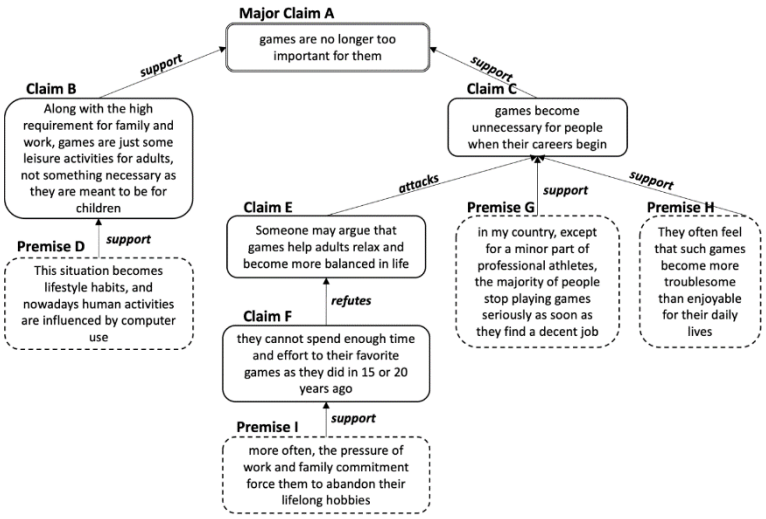

Figure 1. Example of an argument graph of a persuasive essay (Stab and Gurevych, 2017).

Argumentation Rhetoric. Annotators were asked to evaluate at the argument level which rhetoric strategy the argument is following among *ethos*, *logos*, and *pathos* (Aristotle, 2004). Examples 6, 7, and 8 show the rhetorical strategy annotation on three different arguments from (Stab and Gurevych, 2017).

Example 4. The advanced medical care brings with it more benefits than disadvantages. [**The main advantage of high-tech medical care is that people are better taken care so that they have a good health**]$_1$. [*Healthy workers can create more productivity*]$_1$ [*They can contribute effectively to the development of the economy*]$_2$. [*They do not have to spend more time in health checking or treatment*]$_3$. [*this saves an amount of time as well as cost*]$_4$.

Example 5. People should sometimes do things that they do not enjoy. [**In personal live, we have some responsibilities towards to other people, there is nobody who likes all of these responsibilities**]$_1$. [*Housework is very difficult for me, although my*

husband helps me some of them, but it is my responsibility]$_1$. [*I really don't like any of them, however I should do*]$_2$, [*most people's lives are filled with tasks that they don't enjoy doing*]$_3$.

Example 6. Following celebrities can be dangerous for the youth. [**This has an overall effect on personality and future of an individual, following celebrities blindly affects the health of adolescents.**]$_1$ [*Many young people indulge themselves in drugs and start smoking at an early age*]$_1$. [*In a survey carried out in a university, it was asked to students that why did they start smoking, then around forty percent of individuals answered that they wanted to look like their favorite screen actor while smoking cigarettes*]$_2$ [*Imitating celebrities has a negative influence on health of young individuals*]$_3$.

In Example 4 the claim (in bold) appeals to emotions *Pathos* when the author describes how "people are better taken care" in the premises 1 and 3 (in italic). In Example 5 the authors employ *Ethos*, we can notice that the author refers to personal experiences in premises 1 and 2. Example 6 employs *Logos*, the author refers to a formal study, in premise 2, to support its claim.

3.2 Inter-annotator agreement

Before starting the annotation process, three expert annotators carried out a training phase, during which they studied the guidelines and discussed about the ambiguities between the scores for the definitions of Cogency and Reasonableness, amongst others. Then, the annotators were presented with an argument from a persuasive essay and its full argument graph, and they had to annotate the argument quality following the rubric scores.

To ensure the reliability of the annotation task, the inter-annotator agreement (IAA) was calculated on a set of 33 essays, resulting in a Fleiss' kappa of 0.68 for Cogency, 0.78 for Reasonableness Counterargument, 0.84 for Reasonableness Rebuttal, and 0.85 for Argumentation Rhetoric. Despite this substantial agreement, the annotators encountered difficulty in selecting precise scores, such as 25 or 20. To address potential subjectivity issues in the manual annotation, we opted to merge Score 25 with Score 20 and Score 15 with Score 10, resulting in three labels (with Score 0 remaining unchanged).

After recalculating Fleiss' kappa score, we observed an increase only for Cogency (from 0.68 to 0.86). Therefore, we decided to use a three-label score (i.e., 0, 15, 25) for Cogency prediction, while keeping the more fine-grained score (i.e., 0, 10, 15, 20, 25) for Reasonableness. Annotators then engaged in a reconciliation phase, where they resolved disagreements through discussion. One of the expert annotators

performed the remaining annotation. Table IV and Table V report on the statistics of the final dataset.

Table IV. Statistics of the dataset, reporting on the percentage and type of Rhetorical arguments.

No Rhetoric	***Ethos***	***Logos***	***Pathos***
76.04%	11.51%	6.79%	5.66%

Table V. Statistics of the dataset, reporting on the percentage of Cogency and Reasonableness for each score.

Dimension	**Score 0**	**Score 10**	**Score 15**	**Score 20**	**Score 25**
Cogency	19.70%	9.38%	19.14%	31.71%	20.08%
Reasonableness Counterargument	27.27%	25.45%	26.36%	16.64%	7.27%
Reas. Rebuttal	79.82%	9.65%	4.39%	3.51%	2.63%

4 Automatic assessment of argumentation

An overview of the automatic argument quality assessment framework we propose is visualized in Fig. 2. Starting from the persuasive essays where argument components and their relations are identified, the goal is to assess the quality of each argument (i.e., the quality of each claim). Three scores are computed: a *cogency* score in the range {0, 15, 25}, an *argumentation rhetoric* label among *ethos*, *logos*, and *pathos*, and a *reasonableness* score in the range {0, 10, 15, 20, 25}. Two different methods are combined to assess the quality dimensions of the arguments: *(i)* the cogency score and the argumentation rhetoric labels are predicted using an attention-based neural architecture which employs the argumentation graphs through graph embeddings, and *(ii)* the reasonableness score is computed by means

of an algorithm, combining the cogency score predicted at step *(i)* and the graph structure of each persuasive essay.

To feed the argumentative texts into the computational models we need to, first, convert them into vectorial representations called *embeddings*. In the following, we present the textual and graph embeddings we extracted from the persuasive essays to predict the cogency score and argumentation rhetoric labels, the architecture we define to predict these two quality dimensions, and conclude with the reasonableness algorithm used to assess this score.

Table VI summarizes our findings on the technical experiments for the automatic assessment of all quality dimensions, which are further discussed in (Marro et al., 2022)

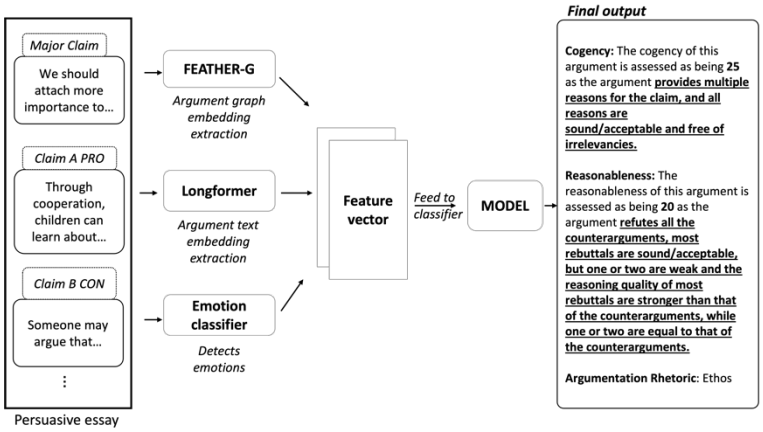

Figure 2. Overview of our natural language argumentation quality prediction model.

4.1 Embeddings

We generate features using *textual and graph embeddings*. Embeddings are low-dimensional, continuous vector representations of real-world data, such as text or graphs, which capture semantics and similarity. For text, the meaning of the words is encoded in such a way that words that are close in the vector space are expected to be similar in meaning. Graph embeddings similarly transform graph properties

into vectors to capture topology, vertex-to-vertex relationships, and any other relevant information.

To generate textual features, we employ various embedding approaches, such as GloVe (Pennington et al., 2014) for static methods and Longformer (Beltagy et al., 2020) for contextualized embeddings, among others. To create a textual representation of an argument, we considered not only the sentences in the claim but also those from related components that are linked to the claim by a support or attack relations (e.g., premises, counterarguments, and rebuttals) to reflect human evaluation of argument quality. For graph embeddings, we used the state-of-the-art FEATHER-G (Rozemberczki and Sarkar, 2020) as our primary model.

To enrich our features for the Rhetoric dimension, we explored a way to obtain representations for the emotions present in the arguments. We utilized a state-of-the-art system trained for the Emotion Recognition downstream task. This approach enabled us to obtain an emotion label from a set of six basic emotions (sadness, joy, love, anger, fear, or surprise), which was subsequently used to generate a word embedding via various techniques discussed earlier in this section.

4.2 Cogency and rhetoric scoring assessment

Following feature generation, we proceed to perform an automated assessment of each quality attribute. With regards to Cogency and Reasonableness, we present a range of models in our experimentation, including both standard baselines and advanced methods. Specifically, we evaluate our models using textual embeddings alone, as well as a combination of textual and graph embeddings.

Our findings, presented in Table VI, demonstrate a significant improvement in the performance of our system upon the inclusion of argument graph features. However, in our assessment of rhetorical strategies, we did not observe any such improvement with the incorporation of graph features, whereas the inclusion of emotion embeddings resulted in a positive impact.

4.3 Reasonableness scoring assessment

As counterarguments and rebuttals are scarce in our dataset, our models struggle to properly classify reasonableness. To address this, we propose a new approach that takes into account the structure of the argumentation graph, which plays a significant role in assessing reasonableness.

The reasonableness dimension (Stapleton and Wu, 2015) considers *(i)* the cogency of the counterarguments attacking the argument we want to assess the reasonableness of, *(ii)* the cogency of the rebuttals to these counterarguments, and *(iii)* the relative number of rebuttals and counterarguments. This means that to effectively compute the reasonableness dimension, we need to combine the cogency-based quality of the argument components and the structure of the argumentation graph.

In (Marro et al. 2022) we propose an algorithm to compute the reasonableness score of the arguments in the persuasive essays. In this Rebuttal Reasonableness Score algorithm, we define how each score is evaluated by combining the cogency values of the pertinent arguments and the relevant properties of the argument graph.

Table VI. Results for automatic assessment of Cogency, Reasonableness and Rhetoric given in macro F1 (Marro et al., 2022).

Cogency Assessment	
Model	***F1 Score***
textual features	0.72
textual & graph features	**0.77**

Reasonableness assessment	
Model	***F1 Score***
majority baseline	0.18
Reasonableness algorithm	**0.54**

Rhetorical assessment	
Model	***F1 Score***
textual features	0.57
textual & emotion features	**0.63**

4.4 Final outcome

After automatically assessing the Cogency, Rhetoric, and Reasonableness dimensions, our system leverages the obtained scores to assist students in improving their essays. The pipeline concludes by automatically generating scores based on the following template:

The [QUALITY DIMENSION] of this argument is assessed as being [*PREDICTED SCORE*] as the argument [DEFINITION] (see Figure 2).

5. Concluding remarks

We presented a novel approach to the task of automatic quality assessment of natural language argumentation. We built a new resource of 402 students' persuasive essays annotated with 3 different quality dimensions. We show that our neural architecture relying on a transformer-based model and graph embeddings can successfully classify arguments along with these quality dimensions. Our quality assessment method conjugates the empirical evaluation of the cogency dimension with the graph-based computation of the reasonableness one, which encompasses the quality (expressed in terms of cogency) of the counterarguments and the argumentation structure.

In the context of AI in education, we aim to include our automatic argument quality assessment pipeline into a larger framework where the system engages the student into an explanatory rule-based dialogue to assess her essays, explain why they obtained a certain quality score and how to improve them along with the considered quality dimensions.

Acknowledgements

This work has been supported by the French government, through the 3IA Côte ’Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR- 19-P3IA-0002

References

Amgoud, L., Doder, D., & Vesic, S. (2022). Evaluation of argument strength in attack graphs: Foundations and semantics. Artificial Intelligence, 302, 103607.

Aristotle (2004). Rhetoric. Translated by Roberts. Mineola, NY: Dover Publications

Baroni, P., Caminada, M., & Giacomin, M. (2011). An introduction to argumentation semantics. The knowledge engineering review, 26(4), 365-410.

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150.

Bench-Capon, T. J. (2003). Persuasion in practical argument using value-based argumentation frameworks. Journal of Logic and Computation, 13(3), 429-448.

Cabrio, E., & Villata, S. (2018, July). Five years of argument mining: A data-driven analysis. In IJCAI (Vol. 18, pp. 5427-5433).

Coertjens, L., Lesterhuis, M., Verhavert, S., Van Gasse, R., & De Maeyer, S. (2017). Teksten beoordelen met criterialijsten of via paarsgewijze vergelijking: een afweging van betrouwbaarheid en tijdsinvestering= Judging texts with rubrics and comparative judgement: Taking into account reliability and time investment. Pedagogische Studien, 94(4), 283-303.

da Costa Pereira, C., Tettamanzi, A. G., & Villata, S. (2011, June). Changing one's mind: Erase or rewind? possibilistic belief revision with fuzzy argumentation based on trust. In Twenty-Second International Joint Conference on Artificial Intelligence.

Habernal, I., & Gurevych, I. (2016, August). Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1589-1599).

Lauscher, A., Wachsmuth, H., Gurevych, I., & Glavaš, G. (2022). Scientia Potentia Est—On the Role of Knowledge in Computational Argumentation. Transactions of the Association for Computational Linguistics, 10, 1392-1422.

Lawrence, J., & Reed, C. (2020). Argument mining: A survey. Computational Linguistics, 45(4), 765-818.

Marro, S., Cabrio, E., & Villata, S. (2022). Graph Embeddings for argumentation Quality Assessment. In Findings of the Association for Computational Linguistics: EMNLP 2022, pages 4154–4164, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

Rozemberczki, B., & Sarkar, R. (2020, October). Characteristic functions on graphs: Birds of a feather, from statistical descriptors to parametric models. In Proceedings of the 29th ACM international conference on information & knowledge management (pp. 1325-1334).

Stab, C., & Gurevych, I. (2017). Parsing argumentation structures in persuasive essays. Computational Linguistics, 43(3), 619-659.

Stapleton, P., & Wu, Y. A. (2015). Assessing the quality of arguments in students' persuasive writing: A case study analyzing the relationship between surface structure and substance. Journal of English for Academic Purposes, 17, 12-23.Tindale, C. W. (2007). Fallacies and argument appraisal. Cambridge: Cambridge University Press.

Toledo, A., Gretz, S., Cohen-Karlik, E., Friedman, R., Venezian, E., Lahav, D., ... & Slonim, N. (2019, November). Automatic Argument Quality Assessment-New Datasets and Methods. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 5625-5635).

Wachsmuth, H., Naderi, N., Habernal, I., Hou, Y., Hirst, G., Gurevych, I., & Stein, B. (2017, July). Argumentation quality assessment: Theory vs. practice. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 250-255).