Graph Embeddings for Argumentation Quality Assessment

Santiago Marro and **Elena Cabrio** and **Serena Villata**
Université Côte d'Azur, CNRS, Inria, I3S, France
{firstname.lastname}@univ-cotedazur.fr

Abstract

Argumentation is used by people both internally, by evaluating arguments and counterarguments to make sense of a situation and take a decision, and externally, e.g., in a debate, by exchanging arguments to reach an agreement or to promote an individual position. In this context, the assessment of the quality of the arguments is of extreme importance, as it strongly influences the evaluation of the overall argumentation, impacting on the decision making process. The automatic assessment of the quality of natural language arguments is recently attracting interest in the Argument Mining field. However, the issue of automatically assessing the quality of an argumentation largely remains a challenging unsolved task. Our contribution is twofold: first, we present a novel resource of 402 student persuasive essays, where three main quality dimensions (i.e., *cogency*, *rhetoric*, and *reasonableness*) have been annotated, leading to 1908 arguments tagged with quality facets; second, we address this novel task of argumentation quality assessment proposing a novel neural architecture based on graph embeddings, that combines both the textual features of the natural language arguments and the overall argument graph, i.e., considering also the support and attack relations holding among the arguments. Results on the persuasive essays dataset outperform state-of-the-art and standard baselines' performance.

1 Introduction

Argumentation is the process by which arguments are constructed, compared, evaluated in several respects and judged in order to establish whether any of them is warranted. Argumentation is an effective approach for solving various theoretical and practical problems (Simari and Rahwan, 2009; Atkinson et al., 2017), like explaining and justifying the decision making outcomes and reasoning under inconsistent and incomplete information. Roughly, each argument is a set of premises or assumptions that, together with a claim, is obtained by a reasoning process. The overall goal of argumentation is to increase or decrease the acceptability of claims by supporting or attacking them with new arguments.

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 (Hunter, 2021; 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). In particular, despite some approaches addressing the automatic assessment of natural language arguments (Wachsmuth et al., 2017a; Wachsmuth and Werner, 2020; Saveleva et al., 2021), 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 characterise 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, 2019; Lauscher et al., 2021) approach to iden-

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 4183 - 4193

tify 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). More specifically, *cogency* estimates the acceptability of the premises that are relevant to the argument's conclusion and their sufficiency to draw the conclusion, *rhetoric* determines the rhetorical strategy employed in the argument's conclusion (if any), and *reasonableness* rates if the argument adequately rebuts its counterarguments. We then train a transfomer-based neural classifier with an attention mechanism called Longformer (Beltagy et al., 2020) empowered with graph embeddings 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 transformer-based model architecture, exploiting the structure of the argument graph through graph embeddings, and we address an extensive evaluation obtaining good results. 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.

The work we present in this paper is motivated by the lack of existing resources of natural language argumentation annotated with quality dimensions, and the need for effective methods to address this task. Our contribution advances the state of the art with a novel resource and an effective method.

2 Related Work

Recent approaches in Argument(ation) Mining (AM) (Cabrio and Villata, 2018; Lawrence and Reed, 2019; Lauscher et al., 2021) tackle specific argument qualities features, such as argument relevancy (Wachsmuth et al., 2017b), convincing arguments (Habernal and Gurevych, 2016) and overall argument quality (Toledo et al., 2019). Previous work on student essays aimed to assess clarity (Persing and Ng, 2013), organization (Persing et al., 2010) and argument strength (Persing and Ng, 2015). (Luo and Litman, 2016) target the automatic prediction of the quality of student reflective responses, showing how expert-coded quality ratings and quality predictions based on their features positively correlate with student learning gain.

Defining the characteristics of a good and successful argument is a hard task. Different approaches have been proposed to assess logical, rhetorical, and dialectical quality dimensions of natural language arguments. (Wachsmuth et al., 2017a) 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 and Werner, 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. In another text-only approach, (Lauscher et al., 2020) describe a large argument quality corpus with data extracted from forums. They propose the first computational model to automatically evaluate Cogency, Reasonableness, Effectiveness and overall quality.

(Saveleva et al., 2021) present an argument quality assessment method defined as a graph classification task. The authors reconstruct the graph structure of the arguments within the argument quality dataset of (Wachsmuth et al., 2017a), showing that this is feasible only in some cases. The reconstructed structures are composed of claims and evidence connected by a support relation, disregarding important elements like counterarguments and rebuttals. Results indicate that discourse-based argument structures reflect qualitative properties of the arguments. For rhetorical aspects, (Duthie et al., 2016) show the impact of the different rhetorical strategies used in political discourse. For Automatic Essay Scoring, (Zhang and Litman, 2021) show how human-labelled evidence scores can be replaced with other automated essay quality signals, such as word count and topic distribution similarity.

In this paper, we advance the state of the art of natural language argument quality assessment by investigating three main quality properties of *persuasive essays* grounding on social science argument quality assessment scores (Stapleton and Wu, 2015). Moreover, we propose a novel method to evaluate the reasonableness of an argument by combining 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). This results in 402 argument graphs where the argument components are the nodes of the graph, and the argumentative relations are the edges of the graph. 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*, following the taxonomy proposed by (Wachsmuth et al., 2017a). Taking advantage of the relation annotations, we use the *argument graph* (i.e., argument components and their relations) to assist annotators in their annotation process.

Annotation guidelines. Given that our goal is to assess persuasive essays written by students, we rely on the quality evaluation process proposed in social sciences showing how these essays are assessed by professors. (Stapleton and Wu, 2015) propose a scoring rubric for persuasive writing that integrates the assessment of both argumentative structural elements and reasoning quality by manually analyzing argumentative essays made by 125 students in Hong Kong. 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 1, 2 and 3 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.

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., 2017a). Annotators were provided with Table 1 to assess the cogency dimension. Following this definition, 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 conclusionThese criteria are considered in point (b) (Table 1) whilst the structural information about the argument graph is addressed in point (a). Example 1 shows the cogency annotation on a persuasive essay from (Stab and Gurevych, 2017). The first sentence is 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.

Score: 25	Score: 20	Score: 15
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

Score: 10	Score: 0
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)

Table 1: Analytic Scoring Rubric to assess Cogency (Stapleton and Wu, 2015).

> **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*] $_3$.

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., 2017a). The Analytic Scoring Rubric for Reasonableness (Tables 2 and 3 (Stapleton and Wu, 2015)) integrates these concepts and follows the idea of evaluating the argumentation graph with a focus on the counterarguments and their respective rebuttals. Annotators were asked to annotate both the Reasonableness Counterargument and the Reasonableness Rebuttal for each claim in the essays which is attacked by a counterargument. Whilst the definitions of Reasonableness and Cogency are similar, the key difference is that with Cogency we evaluate the premises of the argument and with Reasonableness the whole argumentation graph involving the argument to be assessed (including its counterarguments and their rebuttals).

Assessing the Reasonableness of an argument implicates analysing its counterarguments and the related rebuttals. In the example of Figure 1, we

assess the reasonableness quality dimension following Tables 2 and 3: for counterargument *Claim E*, we can see that no reasons, or premises, provided to support it. This falls under the criteria for Score 0 for Reasonableness Counterargument. For the rebuttal, we can see that *Claim F* correctly points out the weakness of the counterargument, providing an acceptable and sound premise and with a reasoning quality stronger than of the counterargument, therefore falling under the criteria for Score 25.

Score: 25	Score: 20	Score: 15
a. Provides multiple reasons for the counterargument claim(s) /alternative view(s), and b. All counterarguments/reasons for the alternative view(s) are sound/acceptable and free of irrelevancies	a. Provides multiple reasons for the counterargument claim(s)/alternative view(s), and b. Most counterarguments/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) /alternative view(s), and b. Some counterarguments/reasons for the alternative view(s) are sound/acceptable, but some are weak or irrelevant

Score: 10	Score: 0
a. Provides only one reason for the counterargument claim(s)/alternative view(s), or b. The counterargument/reason for the alternative view is weak or irrelevant	a. No reasons are provided for the counterargument claim(s)/alternative view(s); or b. None of the reasons are relevant to/support the counterargument claim(s)/alternative view(s)

Table 2: Analytic Scoring Rubric for assessing Reasonableness Counterargument (Stapleton and Wu, 2015).

Score: 25	Score: 20	Score: 15
a. Refutes/points out the weaknesses of all the counterarguments, and b. All rebuttals are sound/acceptable c. The reasoning quality of all the rebuttals are stronger than that of the counterarguments	a. Refutes/points out the weaknesses 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 weaknesses 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 than that of the counterarguments

Score: 10	Score: 0
a. Refutes/points out the weaknesses 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 counterarguments

Table 3: Analytic Scoring Rubric for assessing Reasonableness Rebuttal (Stapleton and Wu, 2015).

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). *Logos* is the act of appealing to the audience through reasoning or logic, by citing facts and statistics, historical and literal analogies. *Ethos* is the act of appealing to the audience through the credibility of the author's beliefs or authority. *Pathos* means to persuade an audience by appealing to their emotions. Some examples of persuasive essays annotated with argumentation rhetoric are available in the Appendix.

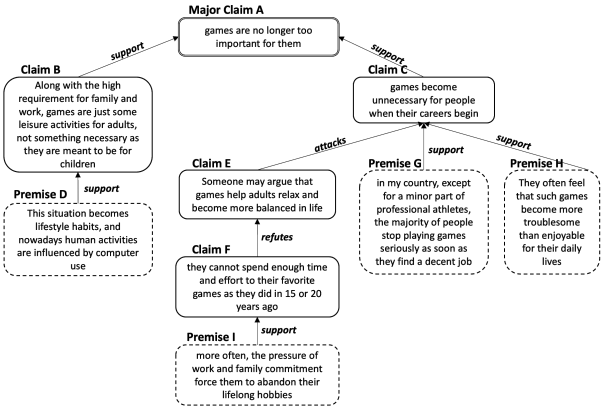

Figure 1: Example of an argument graph of a persuasive essay (Stab and Gurevych, 2017).

In Example 2 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 3 the authors employ *Ethos*, we can notice that the author refers to personal experiences in premises 1 and 2. Example 4 employs *Logos*, the author refers to a formal study, in premise 2, in order to support its claim.

Example 2 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 3 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 4 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*]$_3$.

Before starting the annotation process, three annotators (English speakers and experts in Argumentation Mining) 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 (a Claim or Major Claim component) and its full argument graph, and they had to annotate the argument quality following the rubric scores. To prove the reliability of the annotation task, the inter-annotator agreement (IAA) has been calculated on an unseen set of 33 essays, obtaining 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, an issue for the annotators was the difficulty to opt for a precise score, like 25 or 20. To minimize subjectivity issues in the manual annotation and the consequent noise in the training and testing phases for the automatic assessment of these scores, we decided to merge Score 25 with Score 20, and Score 15 with Score 10, reducing the number of labels to 3 (Score 0 is kept as is). We then proceeded to recompute Fleiss' kappa score, obtaining an increment for Cogency (from 0.68 to 0.86) only. For this reason, we decided to rely on a three-label score for Cogency prediction (i.e., 0, 15, 25), and to keep the more fine-grained score for Reasonableness (i.e., 0, 10, 15, 20, 25). The annotators performed then a reconciliation phase, during which they discussed to reach an agreement on the cases of disagreements. The rest of the annotation was carried out by one of the expert annotators. Tables 4 and 5 report on the statistics of the final dataset.[1]

Score	Cogency	Reas. Counterargument	Reas. Rebuttal
0	19.70%	27.27%	79.82%
10	9.38%	25.45%	9.65%
15	19.14%	26.36%	4.39%
20	31.71%	13.64%	3.51%
25	20.08%	7.27%	2.63%

Table 4: Statistics of the dataset, reporting on the percentage of Cogency and Reasonableness for each score.

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

Table 5: Statistics of the dataset, reporting on the percentage and type of Rhetorical arguments.

[1]The annotated dataset will be released upon paper acceptance. The guidelines are available as Supplementary Material.

4 Automatic assessment of argumentation

An overview of the automatic argument quality assessment framework we propose is visualized in Figure 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 effectively 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. In the following, we present the features we extracted from the persuasive essays to predict the cogency score and argumentation rhetoric labels, the neural architecture we define to predict these two quality dimensions, and we conclude with the reasonableness algorithm used to assess this score.

4.1 Cogency and rhetoric scoring assessment

For feature generation, we employ different embeddings methods ranging from static methods like GloVe (Pennington et al., 2014) to contextualized embeddings, such as BERT (Devlin et al., 2019) and Longformer (Beltagy et al., 2020). To obtain the textual representation of an argument, we take all the sentences in the claim, but also those present in the related argument components. This mirrors the way humans evaluate the quality of an argument, meaning that a claim is assessed not only relying on the sentence(s) composing it but also on the sentences from the related components (i.e., those components linked to this claim by a support or an attack relation). Given that joining all these sentences for every argument component results in a document no longer than 2000 tokens per argument, we use the pre-trained model Longformer (Beltagy et al., 2020) which allows us to process documents up to 4096 tokens with state-of-the-art results. For the graph embeddings, we utilize FEATHER-G (Rozemberczki and Sarkar, 2020) as our main model. To describe node neighbourhoods, this approach combines the characteris-

tic functions of node attributes with random walk weights. These node-level features are pooled by mean pooling to create graph level statistics.

Transformers can be also be used to fine-tune the pre-trained model on a target dataset. To enrich our features for the Rhetoric dimension, we explored a way to obtain representations for the emotions present in the arguments. We use the model T5 (Raffel et al., 2019) fine-tuned on the emotion recognition dataset by (Saravia et al., 2018) for the Emotion Recognition downstream task. This approach allows us to obtain an emotion label amongst sadness, joy, love, anger, fear, or surprise. We then obtain a word embedding as a feature vector by either directly extracting the label representation from the fine-tuned model or employing the label to obtain a word embedding using GloVe.

After feature generation, we automatically assess each quality attribute. For Cogency and Reasonableness, Support Vector Machines (SVM) (Chang and Lin, 2011), Random Forests (Cutler et al., 2012), Bidirectional LSTM-CRF (Huang et al., 2015) and fine-tuned Longformer (Beltagy et al., 2020) with an added dense layer for classification models were investigated. In our experiments, we evaluate different combinations of these methods with different combinations of the previously mentioned embeddings as an input vector.

As the majority of the arguments in our dataset have a non-rhetorical structure (Table 5), the automatic Argumentation Rhetoric assessment task was divided into two different steps. First, a binary classification task to distinguish between a *rhetorical* and a *non-rhetorical* argument, and then a multi-label classification task to classify a rhetorical argument into *ethos*, *logos* or *pathos*. For both tasks, the implemented architectures are the same.

4.2 Reasonableness scoring assessment

Given the fact that in our dataset the majority of the essays did not present any counterarguments or, for a given counterargument, there was no rebuttal, our models did not have enough data to learn how to classify the reasonableness quality dimension. Motivated by this and by the consideration that the structure of the argumentation graph plays a main role in assessing reasonableness, we propose a novel approach to address this task. The reasonableness dimension (Stapleton and Wu, 2015) takes into account *(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 (i.e., the arguments attacking the counterargument), 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. We define the cogency function CV which assigns to each argument component A a cogency value in $\{0, 10, 15, 20, 25\}$, using the SVM plus graph embeddings approach we proposed.

Based on (Stapleton and Wu, 2015), we propose an algorithm (Algorithm 1 in the Appendix) to compute the reasonableness score of the arguments in our argumentation graphs. In this Rebuttal Reasonableness Score algorithm, the reasonableness score of the argument component A is 0 if *(i)* no attack to the counterarguments in CA of A holds (line 19), or *(ii)* the cogency value of the argument components defending A, i.e., attacking the counterarguments of A, is 0 (line 16).

For the remaining reasonableness scores, the reasonableness score of A is 10 if *(i)* at least one and less than half of its counterarguments are attacked (line 25), or *(ii)* the cogency score of more than half of the argument components defending it is 10 (line 22), or *(iii)* the cogency score of more than a half of the argument components defending it is lower than the cogency score of the counterarguments of A (line 29). The reasonableness of A is 15 if *(i)* all the counterarguments of A are attacked (line 32), and *(ii)* the cogency score of at least one of the argument components defending A is equal to or higher than 15 and at least one of the argument components defending A has a cogency score lower than 15 (line 33), and *(iii)* the cogency score of at least one of the argument components defending A is higher than the cogency score of the counterarguments of A and at least one of the argument components defending A has a cogency score lower than the cogency score of the counterarguments of A (lines 34 and 35, respectively). The reasonableness score of A is 20 if *(i)* all the counterarguments of A are attacked (line 32), and *(ii)* the cogency score of more than half of the argument components defending A is equal to or higher than 20, and at least one of the argument components defending A has cogency score equal to or lower than 10 (line 40), and *(iii)* the cogency score of more than half of the argument components de-

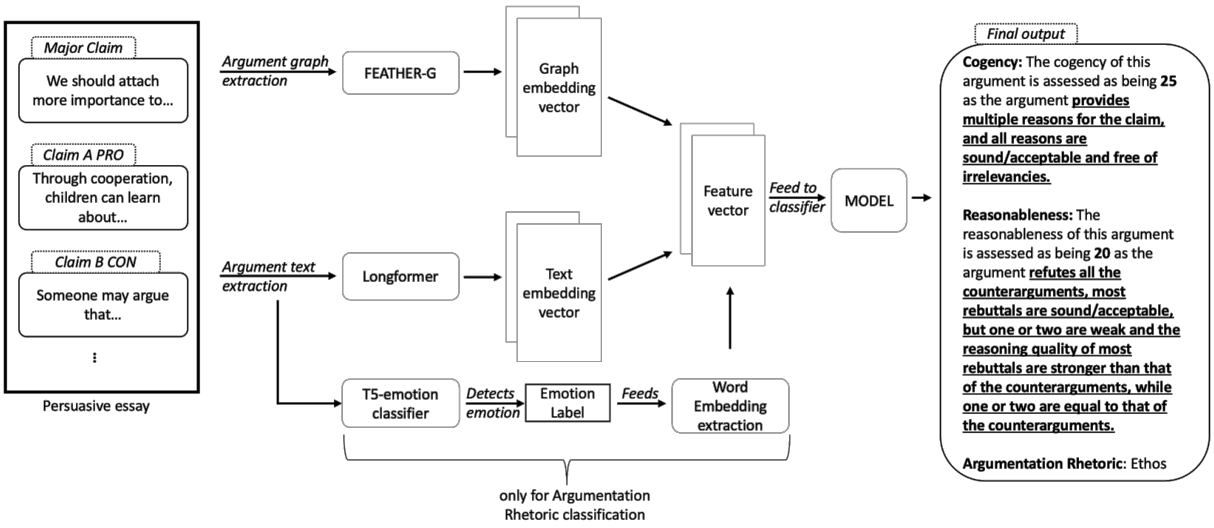

Figure 2: Overview of our natural language argumentation quality prediction model.

fending A is higher than the cogency score of the counterarguments of A while one or two of the argument components defending A has cogency score equal to that of the counterarguments of A (line 41). Finally, the reasonableness score of A is 25 if *(i)* all the counterarguments of A are attacked (line 32), and *(ii)* the cogency score of all of the argument components defending A is 25 (line 45), and *(iii)* the cogency score of all of the argument components defending A is higher than the one of all of the counterarguments of A (line 46).

Let us consider the example in Figure 1. We aim to assess the reasonableness score of claim C. It holds that $CV(E) = 0$ (E is the counterargument of C) and $CV(F) = 10$ (F is the rebuttal of E). Starting from the cogency scores of all the counterarguments and rebuttals of our target argument component C, we can see that if the cogency value of every rebuttal is 10 (the cogency score of claim F), then the reasonableness of claim C is 10.

After the automatic assessment of the Cogency, Rhetoric, and Reasonableness dimensions, the obtained scores are used to help the student to improve the essay. Our pipeline ends with the automatic generation of the scores using this template: The [QUALITY DIMENSION] of this argument is assessed as being **[PREDICTED SCORE]** as the argument **[DEFINITION]** (see Figure 2).

5 Evaluation

In the following, we report on the experimental setup, the obtained results and the error analysis.

Experimental Setup. For argument quality prediction, the embeddings (see Section 4) were combined with either *(i)* a Random Forest, *(ii)* a LSTM, *(iii)* a dense layer, or *(iv)* a SVM. Additionally, the best performing static and dynamic embeddings were concatenated and evaluated as if they were one single embedding. The PyTorch framework (Paszke et al., 2019) version 1.10 was used for implementing the LSTM model with a learning rate selected from 0.05, 0.1, RNN layers 1, 2, dropout 0.1, 0.3, 0.5, and batch size from 8, 16, 32 and a hidden size of 128. For Longformer, BERT and T5 pre-trained models, we use the PyTorch implementation of huggingface (Wolf et al., 2019) version 4.16.2. For the graph FEATHER-G embeddings, the Karate Club framework (Rozemberczki et al., 2020) was used with the standard hyperparameters. The Scikit-learn (Pedregosa et al., 2011) framework was employed for the implementation of the Random Forest and SVM models. We trained the SVM models for each quality attribute, optimizing the Gamma and C hyperparameter (tested C range: 10^{-4} to 10^{3}, Gamma range: 10^{-4} to 10^{0}) on the training data set given by the original split (Stab and Gurevych, 2017) in the dataset. For the rhetoric attribute, we trained the SVM models concatenating the Longformer and FEATHER-G embeddings with the emotion word embedding. The latest was obtained by *(i)* running the fine-tuned T5 model to detect emotions, and *(ii)* either using that label as an input on Glove, or extracting directly from the model the representation of the labels by summarizing the hidden states of the last four layers in the model. To train the binary classification, we con-

Algorithm 1 Rebuttal Reasonableness score

Require: Argument Component A, CA a set with all the argument components that directly attack A.
Ensure: Returns 0, 10, 15, 20 or 25 as a prediction for the Reasonableness Score. ReasonablenessScore$A$, CA

$Y \leftarrow 0$
$cogScores \leftarrow []$
$CACogScores \leftarrow$ get the cogency values for each arg. in CA
for each argument C in CA **do**
  $DA \leftarrow$ get all the arg. components that attack C
  $cogScores \leftarrow$ append the cogency values for each arg. in DA
  if $len(DA) > 0$ **then**
    $Y \mathrel{+}= 1$
  end if
end for
$CV10 \leftarrow$ Count how many of the rebuttal scores in $cogScores$ are 10
$CV20 \leftarrow$ Count how many of the rebuttal scores in $cogScores$ are 20
$Q \leftarrow$ Count how many rebuttals have a cogency score lower to the counterarguments.
$X \leftarrow$ Count how many rebuttals have a cogency score higher than all of the counterarguments.
$Z \leftarrow$ Count how many rebuttals have a cogency score equal to the counterarguments.
if max($cogScores$) $= 0$ **then**
  return Score 0
end if
if $Y = 0$ **then**
  return Score 0
end if
if $CV10 > \frac{len(cogScores)}{2}$ **then**
  return Score 10
end if
if $1 \leq Y \leq \frac{len(CA)}{2}$ **then**
  return Score 10
end if
if $Q > \frac{len(cogScores)}{2}$ **then**
  return Score 10
end if
if $Y = len(CA)$ **then**
  if max($cogScores$) ≥ 15 and min($cogScores$) < 15 **then**
    if max($cogScores$) $>$ max($CACogScores$) **then**
      if min($cogScores$) $<$ min($CACogScores$) **then**
        return Score 15
      end if
    end if
  end if
  if $len(CV20) > \frac{len(cogScores)}{2}$ and min($cogScores$) ≤ 10 **then**
    if $X > \frac{len(cogScores)}{2}$ and $1 < Z \leq 2$ **then**
      return Score 20
    end if
  end if
  if min($cogScores$)=25 and max($cogScores$)=25 **then**
    if min($cogScores$) $>$ max($CACogScores$) **then**
      return Score 25
    end if
  end if
end if

verted all of the *ethos*, *pathos* and *logos* labels to *rhetorical*, while for the multi-label classification all the non-rhetorical arguments were discarded.

Embedding	Model	f1	F1
Longformer	RandomForest	0.72	0.74
Longformer	LSTM	0.55	0.51
finetunning Longformer	dense layer	0.43	0.33
Longformer	SVM	0.74	0.72
Long. + FEATHER-G	RandomForest	0.73	0.75
Long. + FEATHER-G	SVM	**0.78**	**0.77**

Table 6: Results for the Cogency score of the 3-class sequence tagging task are given in weighted F1 (f1) and macro F1 (F1).

Results. Table 6 and 7 report on the results for the best performing models and embedding combi-

	Binary Clf.		Multi-label Clf		Full Pipeline	
Embedding	f1	F1	f1	F1	f1	F1
Longformer	0.78	0.69	0.70	0.62	0.91	0.57
Long.+ FEATHER-G	0.78	0.69	0.66	0.58	0.91	0.57
Long.+ FEATHER-G+ T5	0.80	0.73	0.70	0.62	0.89	0.62
Long.+ T5	0.80	0.73	0.80	0.72	0.89	0.62
Long.+ T5w/GloVe	**0.80**	**0.73**	**0.80**	**0.77**	**0.89**	**0.63**

Table 7: Results of the Argumentation Rhetoric sequence tagging task training a SVM model (weighted F1 (f1) and macro F1 (F1)).

nations. Performances are given on the test set in weighted average and macro multi-class F1-score. Each run was repeated five times with different random seeds to assess the stability of the results and the average score is reported. For Cogency classification, a significant improvement (from .72 to .77 macro F1-score) can be seen when the FEATHER-G graph embeddings are combined with the Longformer embeddings. The best performing model (in bold) is composed of these embeddings along with a SVM model for the quality prediction scores.

For Reasonableness, due to the scarceness of counterarguments and rebuttals, no deep learning model showed a significant learning success. Following Algorithm 1, we obtain the Rebuttal Reasonableness score for each argument (computed starting from the cogency values obtained by our model, not the golden labels) yielding an accuracy of .80 and a macro F1 of .54 while a majority baseline obtains an accuracy of .78 and a macro F1 score of .18.

Table 7 shows the results for the two steps and the full pipeline of the argumentation Rhetoric classification task. The T5 fine-tuned model with Glove embeddings shows the best performance with a .73 macro F1-score for the first step of the pipeline (i.e., the binary classification *rhetoric/non-rhetoric*), a .77 macro F1-score for the multi-label classification (i.e., the multi-class classification *ethos/logos/pathos*), and a .63 macro F1-score for the full pipeline. We observe that the performance improves for every model when we add the emotion embeddings to the input feature vector, supporting our choice of integrating a general emotion dimension into the rhetorical classification for a better embedding representation. We can also notice that the graph embeddings are not really contributing to this task, leading to a detriment of macro F1-score. This result can be explained as the persuasive rhetorical strategies relies mainly on the textual formulation of the argument component itself, without being impacted by the support and

attack relations involving this component.

We addressed a comparison with the state-of-the-art approaches for the Cogency assessment, despite the fact that we focus on a different dataset and divergent features (e.g., graph embeddings in our case). We retrained our model with the dataset of (Wachsmuth and Werner, 2020) following their same configuration. In this dataset of forum data, each argument instance is associated to 3 different gold labels for Cogency, one for each annotator. They also separate them into 16 different topics and train each model with 15 of them, testing on the excluded one. We followed the same process for each annotator with our baseline model (Longformer embedding + SVM). Given that they do not provide any graph structure we cannot test our best model on their data to compare. However, the results obtained are a Mean Absolute Error of .64, .38 and .52 for Expert #1, #2 and #3, respectively. Comparing with (Wachsmuth and Werner, 2020), we can see that for Expert #2 our baseline model outperforms their best model (.38 vs .57). In the case of Expert #1, we obtain the same result as their baseline, and for Expert #3 we perform similarly (.52 vs .50). The results we obtained on Cogency (.78 f1) are, to the best of our knowledge, the best result obtained so far in the literature (Wachsmuth and Werner, 2020; Saveleva et al., 2021).

Error Analysis. A common mistake for Cogency is that the scores 0 and 25 are more often correctly classified than score 15. This is due to the imbalance of score 15 given by the nature of the essays, and the complexity of the task for human annotators, as it is easier to distinguish bad from good cogency quality, but more difficult to assess a more subtle distinction. For the argumentation Rhetoric binary classification task, the model tends to misclassify the arguments as *non rhetorical*. This results from the imbalanced dataset, where 76% of the arguments are non-rhetorical. For the multi-label classification task, the model tends to confuse pathos arguments with ethos. This can be explained by the fact that *ethos* and *pathos* are the majority and minority classes, respectively. A further extension of the dataset with the spans of text in the argument that justify the annotated rhetorical structure could yield an improvement in the performance of sequence tagging. For Reasonableness, disagreements between the results given by the algorithm and the gold labels mostly lie in a wrong classification of the cogency score for the counterarguments and rebuttals, leading to a propagation of the error to the reasonableness score.

6 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, i.e., cogency, rhetoric and reasonableness. Through our extensive evaluation, we show that our neural architecture relying on a transformer with an attention mechanism and graph embeddings is able to successfully classify arguments along with these quality dimensions, outperforming standard baselines and similar approaches in the literature. 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.

Limitations

In this section, we discuss the main limitations of the proposed approach to automatically assess argumentation quality along with the three quality dimensions of Cogency, Reasonableness and Rhetoric.

First, we cannot directly compare our results with the few existing approaches in the literature for this task (Wachsmuth and Werner, 2020; Saveleva et al., 2021) by empirically testing our approach on the existing datasets for argumentation quality. This is due to the fact that existing datasets annotated with argument quality dimensions do not have the graph structure available (i.e., argument components and their relations) but only the text of each argument component separately. A further annotation effort to include argument relations in these datasets would allow us to evaluate our approach on these datasets too.

Secondly, the three argument quality aspects we assessed are general to argumentation and fit well for characterizing persuasive essays but may lack

depth when evaluating arguments in other domains. For instance, a good argument in evidence-based medicine may not be characterised only in terms of Cogency, Reasonableness and Rhetoric. Further dimensions need to be defined according to each different domain and precise use case scenarios.

Finally, the lack of counterarguments and rebuttals in the persuasive essays dataset makes it difficult for any automatic method to learn the reasonableness assessment task. More data on these classes would improve not only the training of such models, but would also enable more evaluation cases of the proposed algorithm.

Acknowledgements

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

References

Leila Amgoud, Dragan Doder, and Srdjan Vesic. 2022. Evaluation of argument strength in attack graphs: Foundations and semantics. *Artif. Intell.*, 302:103607.

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

Katie Atkinson, Pietro Baroni, Massimiliano Giacomin, Anthony Hunter, Henry Prakken, Chris Reed, Guillermo Ricardo Simari, Matthias Thimm, and Serena Villata. 2017. Towards artificial argumentation. *AI Mag.*, 38(3):25–36.

Pietro Baroni, Martin Caminada, and Massimiliano Giacomin. 2011. An introduction to argumentation semantics. *Knowl. Eng. Rev.*, 26(4):365–410.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Trevor J. M. Bench-Capon. 2003. Persuasion in practical argument using value-based argumentation frameworks. *J. Log. Comput.*, 13(3):429–448.

Elena Cabrio and Serena Villata. 2018. Five years of argument mining: a data-driven analysis. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 5427–5433. ijcai.org.

Chih-Chung Chang and Chih-Jen Lin. 2011. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27.

Adele Cutler, D Richard Cutler, and John R Stevens. 2012. Random forests. In *Ensemble machine learning*, pages 157–175. Springer.

Célia da Costa Pereira, Andrea Tettamanzi, and Serena Villata. 2011. Changing one's mind: Erase or rewind? In *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*, pages 164–171. IJCAI/AAAI.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186.

Rory Duthie, Katarzyna Budzynska, and Chris Reed. 2016. Mining ethos in political debate. In *Computational Models of Argument: Proceedings from the Sixth International Conference on Computational Models of Argument (COMMA)*, pages 299–310. IOS Press.

Ivan Habernal and Iryna Gurevych. 2016. 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)*, pages 1589–1599.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Anthony Hunter. 2021. Argument strength in probabilistic argumentation using confirmation theory. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty - 16th European Conference, ECSQARU 2021, Prague, Czech Republic, September 21-24, 2021, Proceedings*, volume 12897 of *Lecture Notes in Computer Science*, pages 74–88. Springer.

Anne Lauscher, Lily Ng, Courtney Napoles, and Joel Tetreault. 2020. Rhetoric, logic, and dialectic: Advancing theory-based argument quality assessment in natural language processing. *arXiv preprint arXiv:2006.00843*.

Anne Lauscher, Henning Wachsmuth, Iryna Gurevych, and Goran Glavas. 2021. Scientia potentia est - on the role of knowledge in computational argumentation. *CoRR*, abs/2107.00281.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Comput. Linguistics*, 45(4):765–818.

Wencan Luo and Diane Litman. 2016. Determining the quality of a student reflective response. In *The twenty-ninth international FLAIRS Conference*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward

Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP 2014*, pages 1532–1543.

Isaac Persing, Alan Davis, and Vincent Ng. 2010. Modeling organization in student essays. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 229–239.

Isaac Persing and Vincent Ng. 2013. Modeling thesis clarity in student essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 260–269.

Isaac Persing and Vincent Ng. 2015. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 543–552.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Benedek Rozemberczki, Oliver Kiss, and Rik Sarkar. 2020. Karate Club: An API Oriented Open-source Python Framework for Unsupervised Learning on Graphs. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20)*, page 3125–3132. ACM.

Benedek Rozemberczki and Rik Sarkar. 2020. Characteristic functions on graphs: Birds of a feather, from statistical descriptors to parametric models. *CoRR*, abs/2005.07959.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

Ekaterina Saveleva, Volha Petukhova, Marius Mosbach, and Dietrich Klakow. 2021. Graph-based argument quality assessment. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1268–1280.

Guillermo Ricardo Simari and Iyad Rahwan, editors. 2009. *Argumentation in Artificial Intelligence*. Springer.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Paul Stapleton and Yanming (Amy) Wu. 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.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Automatic argument quality assessment–new datasets and methods. *arXiv preprint arXiv:1909.01007*.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017a. 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*, pages 176–187, Valencia, Spain. Association for Computational Linguistics.

Henning Wachsmuth, Benno Stein, and Yamen Ajjour. 2017b. "pagerank" for argument relevance. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1117–1127.

Henning Wachsmuth and Till Werner. 2020. Intrinsic quality assessment of arguments. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6739–6745.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Haoran Zhang and Diane Litman. 2021. Essay quality signals as weak supervision for source-based essay scoring. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 85–96, Online. Association for Computational Linguistics.