

Argumentation ranking semantics as a feature for classification – On automatic evaluation of argumentative essays

Roberto Barile¹, Claudia d'Amato¹, Nicola Di Mauro¹, Stefano Ferilli¹ and Nunzia Lomonte¹

¹University of Bari, Via E. Orabona 4, Bari, 70125, Italy

Abstract

In this paper we focus on the automatic evaluation of argumentative essays, for scaffolding improvements in writing skills. Our goal is providing an automated approach to classify argumentative elements as "effective", "adequate", or "ineffective". We propose the usage of an additional feature, called ranking score, in the training process of a text-based classifier. The ranking score is obtained by performing argumentative reasoning on the different argumentative elements of an essay. We experimentally show that the introduction of this feature leads to improved performance of both Ada boost classifier and biLSTM neural network.

Keywords

digital libraries, automatic evaluation of argumentative essay argumentative reasoning

1. Introduction

The attention to suitable educational support tools is increasing, particularly in the last few years, when facing the COVID-19 pandemic emergency. Among the others, automatic feedback writing solutions [1] have been considered, since it has been observed that, with automated guidance, students resulted to be able to complete increased assignments and ultimately become more confident and proficient writers.

Despite several automated writing feedback tools are currently available, most of them have limitations with argumentative writing, as they often fail to evaluate the quality of argumentative elements, such as organization, evidence, and idea development ¹.

In this paper we focus on the automatic evaluation of argumentative essays², for

¹AI³ 2022: 6th Workshop on Advances in Argumentation in Artificial Intelligence, Nov 28 - Dec 2, Udine, Italy

✉ r.barile17@studenti.uniba.it (R. Barile); claudia.damato@uniba.it (C. d'Amato); nicola.dimauro@uniba.it (N. Di Mauro); stefano.ferilli@uniba.it (S. Ferilli); n.lomonte1@studenti.uniba.it (N. Lomonte)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

¹<https://www.kaggle.com/competitions/feedback-prize-effectiveness>

²This task was a competition on Kaggle, details on the evaluation metric and the data used in this work can be found at <https://www.kaggle.com/competitions/feedback-prize-effectiveness>

scaffolding improvements in writing skills. Particularly, each argumentative essay can be split into several components, called *discourse elements*. Each discourse element play a specific role in the argumentation. The definitions of the roles are summarized in the following [2]³:

Lead: an introduction that begins with a statistic, a quotation, a description, or some other device to grab the reader’s attention and point toward the thesis

Position or Primary claim: an opinion or conclusion on the main question

Claim: a claim that supports the position

Counterclaim: a claim that refutes another claim or gives an opposing reason to the position

Rebuttal: a claim that refutes a counterclaim

Evidence or Data: ideas or examples that support claims, counterclaims, or rebuttals.

Concluding statement: a concluding statement that restates the claims.

The evaluation of an argumentative essay is meant at predicting the quality of each *discourse element*, which can be judged as *Ineffective*, *Adequate*, or *Effective*.

While most of the existing solutions are basically grounded on the exploitation of text classification techniques [1], we aim at improving the evaluation of argumentative essay by integrating argumentative reasoning solutions within standard multi-class classification methods. Specifically, we aim at improving the performance of text classification models by adopting additional features: the discourse element type and its preliminary assessment (represented as a numeric feature) obtained by performing argumentative reasoning, that computes the strength propagation ranking semantics (sp-ranking) of a Bipolar Weighted Argumentation Framework (BWAFF) [3].

This research direction is motivated by the fact that the quality of a *discourse element* does not depend solely on its textual features, grammatical or syntactical quality, rather, it also depends on how the *the discourse element* "attacks" or "supports" other discourse elements. We show how the performance of state of the art models can be actually improved by the adopting argumentative reasoning solutions for the purpose.

The rest of the paper is organized as follows. The next section synthesizes the state of the art on automated evaluation of argumentative writing. Basics on the adopted argumentation framework are provided in Sect. 3. The modelled solution is illustrated in Sect. 4 whilst experiments are presented in Sect. 5. Conclusions are drawn in Sect. 6 along with some proposals for further research directions.

³We extend the different roles, as reported in [2], with the the definition of *Lead* that stands for an introductory element not providing argumentative impact.

2. Related works

In automated evaluation of (student) argumentative writing, text-based classification solutions falling in one of the three following categories are generally employed [4]:

- **feature-based** where off-the-shelf algorithms are used with additional hand-crafted features, such as:
 - **lexical** features which aim is to capture information at the level of words (e.g. n-grams and words frequency).
 - **syntactic** features which commonly rely on parse trees (e.g. number of sub-clauses found in a tree or part-of-speech tags) [5, 6].
 - **structural** features, which generally describe the position and frequency of a piece of text (e.g. the position of a token, a punctuation character or an argumentative component) [5].
 - **embedding** which rely on the representation of words as vectors in a continuous space [5].
 - **discourse** which captures how sentences or clauses are connected together. Discourse features can be obtained analyzing discourse markers or discourse parser [7]. For instance the marker "therefore", suggests the relationship between a current text span and its adjacent text span. As for discourse parser, the sentences are parsed into discourse roles that are then used as input features.
- **neural-based** where neural architectures such as long short-term memory (LSTM) networks and convolutional neural networks (CNN) are adopted. More recently also transformer based models, such as BERT, were explored [8, 9, 10]. In particular, this architectures are adopted in a transfer leaning fashion, starting from pretrained language models, obtained from general domain corpora containing large amount of texts and adapting such models to specific downstream tasks.
- **unsupervised** methods which use heuristics for bootstrapping a small set of labels and then training the text-based classification model in a self-training fashion [11].

Differently from the state of the art (independently on the specific category), we consider split essays where each discourse element has an evaluation label allowing us to build argumentation frameworks to be used for obtaining an additional feature to be exploited for coming up to a possibly improved automatic evaluation of argumentative essays. Hence, we concentrate our attention on assessing if the additional feature can bring, in the evaluation of a specific discourse element, a value added, independently on the specific classifier that is adopted.

Very few approaches related to our solution can be found. In [12], where annotations involve both argumentative role and evaluation label, relationships between elements are only of support type. While in [13] annotations also includes directed labels (support, attack or detailing), but, differently from our solution, the evaluation labels for the text spans are not provided.

3. Basics

Argumentation is the inferential strategy for practical and uncertain reasoning aimed at coping with partial and inconsistent knowledge, in order to justify one of several contrasting positions in a discussion [14].

Abstract argumentation, in particular, focuses on the resolution of the dispute based only on ‘external’ information about the arguments (notably, the inter-relationships among them), neglecting their internal structure or interpretation.

An argumentation semantics is the formal definition of a method ruling the argument evaluation process. The standard acceptability semantics, introduced by Dung [5], characterizes admissible sets of arguments. Traditional Abstract Argumentation Frameworks (AFs for short) can express only attacks among arguments [15]:

Definition 1. An argumentation framework (or AF) is a pair $\mathcal{F} = \langle \mathcal{A}, \mathcal{R} \rangle$, where \mathcal{A} is a finite set of arguments and $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$ is an attack relationship (meaning that, given $\alpha, \beta \in \mathcal{A}$, if $\alpha \mathcal{R} \beta$ then α attacks β).

While sufficient to tackle many cases (because the attack relationship is indeed the very core and driving feature in a debate), this is a limitation in expressiveness. So, more recent studies tried to introduce additional features to be considered in the argumentation frameworks, as summarized below.

Bipolar AF (BAF) [16]: allows two kinds of interactions between arguments, that are *attack* and *support* relationship;

Weighted AF (WAF) [17]: allows specifying a weight for each attack between arguments, indicating its relative strength;

Bipolar Weighted Argumentation Framework (BWAFF) [3]: embed the notions of *attack* and *support* into the weights. In [18], weights are associated to arguments and the evaluation method transforms them into an overall strength (to be interpreted as an acceptability degree) based on the attacks and supports received. It deals only with acyclic graphs.

General Argumentation Framework (GAF) [19]: extends traditional AFs with bipolarity, weights on both attacks and supports, and weights on the arguments.

GAF provides a general and powerful setting, allowing to express all the other frameworks. It is formally defined as follows:

Definition 2. A General Argumentation Framework (GAF) is defined as a tuple $F = \langle \mathcal{A}, \mathcal{S}(\mathcal{A}), w_{\mathcal{A}}, w_{\mathcal{R}} \rangle$ where:

- \mathcal{A} is a finite set of arguments,
- $\mathcal{S}(\mathcal{A})$ is a system providing external information on the arguments in \mathcal{A} ,
- $w_{\mathcal{A}}: \mathcal{A} \times \mathcal{S}(\mathcal{A}) \mapsto [0, 1]$ assigns a weight to each argument, to be considered as its strength, also based on $\mathcal{S}(\mathcal{A})$,
- $w_{\mathcal{R}}: \mathcal{A} \times \mathcal{A} \mapsto [-1, 1]$ assigns a weight to each pair of arguments.

4. Methodology

In this section we illustrate our proposed approach for performing the automatic evaluation of argumentative essays. It is grounded on the integration of argumentative reasoning solutions and standard multi-class classification methods. Specifically, we employ argumentative reasoning for computing the strength propagation ranking semantics (sp-ranking) of a BWAF, that is then used as the additional numeric feature (besides the textual data) to be exploited for the classification process.

In the following we first illustrate the choices that have done for actually using the adopted AF, hence we describe how it is applied for computing the additional numeric feature.

4.1. The Argumentative Framework

In order to perform argumentative reasoning given the essay discourse elements, the BWAF has been adopted. It is formally defined as follows:

Definition 3. A BWAF is a triplet $\mathcal{G} = \langle \mathcal{A}, \mathcal{R}, w_R \rangle$, where \mathcal{A} is a finite set of arguments, $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$ is either an attack or support relationship between arguments and $w_R: \mathcal{R} \mapsto [-1, 0[\cup]0, 1]$ is a function assigning a weight to each relation.

Attack relations are defined as

$$\mathcal{R}_{att} = \{ \langle a, b \rangle \in \mathcal{R} \mid w_R(\langle a, b \rangle) \in [-1, 0[\}$$

and support relations as

$$\mathcal{R}_{sup} = \{ \langle a, b \rangle \in \mathcal{R} \mid w_R(\langle a, b \rangle) \in]0, 1] \}$$

Initially, only two levels of evaluations for arguments have been considered (arguments are either accepted or rejected). Nevertheless, for several real world applications, this may represent an actual limitation. Hence, the notion of ranking-based semantics [20] has been proposed. It allows to use semantics for capturing arguments with larger levels of acceptability, thus making it possible to also rank them.

For our purpose, the strength propagation semantics [3] has been adopted. It is formally defined as follows:

Definition 4. Let $\mathcal{G} = \langle \mathcal{A}, \mathcal{R}, w_R \rangle$ be a BWAF and let $a, b \in \mathcal{A}$ be two arguments such that there exists a simple path $\langle a \dots b \rangle$. The strength propagation (sp) from a towards b is defined as:

$$sp(a, b) = \sum_{\langle a \dots b \rangle} (pw(\langle a \dots b \rangle)) \times \prod_{c \in \langle a \dots b \rangle} infl(c)$$

where $pw(\cdot)$ (path weight) computes the strength of a simple path by multiplying every weight relation in it, while the function $infl(\cdot)$ (influence) computes the influence of an element within the simple path, on the basis of cycles to which it belongs.

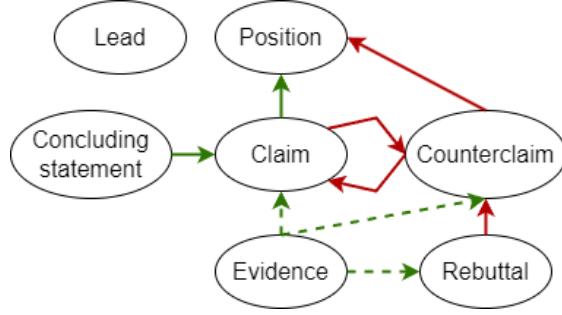


Figure 1: General structure of a BWAf built on the discourse elements of an argumentative essay. Green arrows represent supports, red arrows represent attacks, the green arrows starting from evidence are dashed because an evidence can express support only to one discourse element (of type claim, counterclaim or rebuttal), in particular, to the one with the highest value of the chosen similarity function between discourse elements.

Relying on the previous definitions, we define the *spr* function whose aim is computing the final ranking score of an argument.

Definition 5. Let $\mathcal{G} = \langle \mathcal{A}, \mathcal{R}, w_R \rangle$ be a BWAf, $a \in \mathcal{A}$ an argument, $sp(\cdot, a)$ the strength propagation of a path ending to argument a , $SP = \{sp(x_1, a), \dots, sp(x_m, a)\}$ the set of all the strength propagations on the different path ending to a and $\mathcal{P} = \{p_1, \dots, p_n\}$ the set of all directed paths towards a in \mathcal{G} , with $p_i = \langle x, \dots, a \rangle \in \mathcal{P}, \forall i \leq n$. The *spr* function $spr : \mathcal{A} \mapsto [0, 2]$ is defined as:

$$spr(a) = \begin{cases} 1 & \text{if } \forall x \in \mathcal{A} : \langle x, a \rangle \notin \mathcal{R} \\ \frac{1}{n} \sum_{sp(x_i, a) \in SP} 1 + sp(x_i, a) & \text{otherwise} \end{cases}$$

In the next session, we present how, given the formalized framework, a BWAf can be built from an argumentative essay.

4.2. Building a Bipolar Weighted Argumentation Framework from an Argumentative Essay

As illustrated in Sect. 1, argumentative essay can be split into several *discourse elements*, each one having a specific role within the essay. These roles may be employed for building a BWAf from an argumentative essay.

The general approach is modeled in Fig. 1, where green arrows represent *supports*, red arrows represent *attacks*, whilst the green arrows starting from evidence are dashed as an evidence can express support only to one discourse element (of type claim, counterclaim or rebuttal) that is generally the most similar one.

In order to obtain the weights of *attacks* and *supports*, a similarity function between discourse element needs to be computed. Specifically, the weights of *attacks* and *supports* are obtained computing a similarity function between document embeddings [21] which represents the semantics of the text in a continuous vector space. In

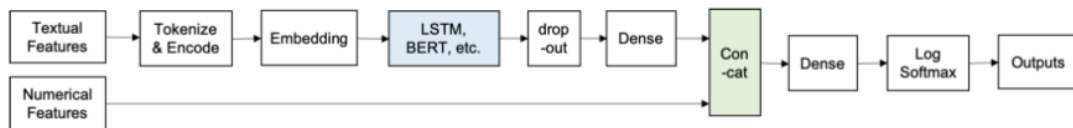


Figure 2: biLSTM additional layer for integration of numeric features

the specific setting, documents are actually the discourse elements. For the purpose, any pre-trained model (e.g. *word2vec* [22]) which maps each discourse element to an embedding vector can be used. The general procedure that is implemented is summarized as follows:

1. extract an embedding vector for each word in the discourse element by means of *word2vec* model for each word in the discourse element
2. compute the weighted average of the embedding vectors using, as weighting factors, the *tf-idf* score of each word.

The sp-ranking semantics can now be computed on each BWAf, obtaining the ranking of each discourse element with respect to other discourse elements in an argumentative essay.

4.3. Classifying Discourse Elements

As illustrated in Sect. 1, the final evaluation task is performed by using text classification. For the purpose we considered Ada Boost [23] and Long Short-Term Memory (LSTM) neural networks [24]. They are briefly summarized in the the following.

- **Ada Boost algorithm** [23] is the first practical contribution in the context of boosting algorithms. Boosting is an ensemble learning approach based on the idea of creating a highly accurate model through the combination of many, relatively weak, models.

For this task Ada Boost is trained on a bag of words representation of discourse elements, in which each word is associated to its *tf-idf* score.

- **biLSTM neural networks** [24] is a recurrent neural network (RNN) architecture that has been designed to address the vanishing and exploding gradient problems of conventional RNNs. Unlike feedforward neural networks, RNNs have cyclic connections making them powerful for modeling sequences. Indeed, they have been successfully used for sequence labeling and sequence prediction tasks, such as handwriting recognition, language modeling, phonetic labeling of acoustic frames. A further extension of this architecture is its bidirectional variant [25] which can be trained using all available input information, that, in the context of textual input, means using words before and after a specific one.

For this task the initial embedding layer of the biLSTM is initialized from a pretrained model ⁴. The loss function used in the training process is the negative log likelihood loss. This choice is justified by the similarity in the behaviour between the loss function and the evaluation metric adopted in the testing phase (see Sect. 5 for details).

We slightly modified both models in order to include the additional numeric features (sp-ranking). In the standard classification the integration is straightforward, the sp-ranking and the discourse type are just other numerical values, analogously to the tf - idf scores. For the biLSTM network, instead, we needed to add a concatenation step as represented in Fig. 2.

5. Evaluation

In this section we illustrate the experiments carried out for assessing the validity of our proposed solution. We first illustrate the dataset that has been adopted for the automatic evaluation of argumentative essays. Hence we specify the evaluation metric and the experimental setting for the considered classifier. We conclude the section with discussing the obtained results.

5.1. Dataset

The dataset that has been used for experiments is publicly available ⁵ and contains argumentative essays written by U.S students in grades 6-12. The essays were annotated by expert for elements commonly found in argumentative writing. The dataset overall counts 15594 texts divided into sub-sections, for a total of 144293 speeches. Each speech is composed of the following characteristics:

- `discourse_id` - ID code for discourse element
- `essay_id` - ID code for essay response. This ID code corresponds to the name of the full-text file in the train/ folder.
- `discourse_text` - Text of discourse element.
- `discourse_type` - Class label of discourse element.
- `discourse_type_num` - Enumerated class label of discourse element.
- `discourse_effectiveness` - Quality rating of discourse element, the target.

In Fig. 3, a pie chart depicting the percentages for each type of argumentative text is provided, while Fig. 4 shows the pie chart of the percentages for the three possible target feature values.

Given the initial dataset, we filter out some essays. In particular we kept only essays composed of less than 15 discourse elements. This choice was required

⁴In our experiments we downloaded the word2vec model trained on Google news and with embedding size equal to 300

⁵<https://www.kaggle.com/competitions/feedback-prize-effectiveness>



Figure 3: Distribution of data with respect to different types of speeches.

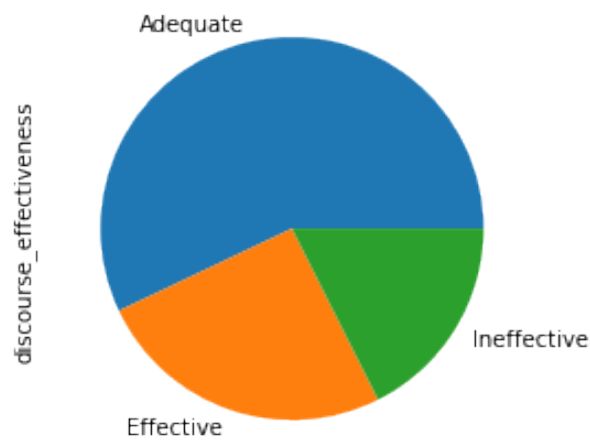


Figure 4: Distribution of data with respect to target feature.

to keep the experiments computationally feasible because the complexity of the argumentation semantics increases with the number of arguments in the graph (see Sect. 4 and Fig. 1 for details).

5.2. Evaluation Metric

The metric used for evaluation is the multi-class logarithmic loss, defined as:

Table 1

Ada Boost log losses in the two settings

iteration	without ranking	with ranking
1	1.0855	1.0844
2	1.0855	1.0844
3	1.0856	1.0844
4	1.0857	1.0845
5	1.0854	1.0844
6	1.0854	1.0844
7	1.0854	1.0844
8	1.0856	1.0844
9	1.0854	1.0844
10	1.0856	1.0844

$$\log_loss = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^M y_{ij} \log(p_{ij})$$

where N is the number of rows in the test set, M is the number of class labels, \log is the natural logarithm, y_{ij} is 1 if observation i is in class j and 0 otherwise, and p_{ij} is the predicted probability that observation i belongs to class j .

In the following we separately discuss the evaluation of the two approaches.

5.3. Ada Boost classifier

The setup adopted for this approach consists in:

1. Perform $n(= 10)$ iterations of cross-validation in which the model is trained **without** taking into account the sp-ranking
2. Perform $n(= 10)$ iterations of cross-validation in which the model is trained taking into account the sp-ranking
3. Perform a corrected resampled t-test to check if there is a significant difference between the two approaches

The performance, in terms of log loss, for the ten iterations are reported in table 1. We notice in all iterations a decrease in the log loss value, on average this difference is equal to 0.0011.

To perform the statistical test we compute the value of t as follows:

$$t = \frac{m_d}{\sqrt{(\frac{1}{k} + \frac{n_2}{n_1})\sigma^2}}$$

where:

- m_d is the mean of the differences between the log losses of the two setups

Table 2

biLSTM 5-fold cross-validation with hp-tuning

batch size	learning rate	epochs number	use sp-ranking	log loss
64	0,005	9	No	0,859
128	0,0005	10	Yes	0,853
64	0,005	8	No	0,840
32	0,0005	9	Yes	0,848
32	0,005	7	No	0,861

- $k = 100$ since we perform a 10-fold cross-validation for 10 times
- $n_2 = 0, 1$ and $n_1 = 0, 9$ because the set is split in 10 folds
- σ^2 is the variance of the differences between the log losses of the two setups

The value of t is -6.466, we fix as confidence level for the statistical test $c = 5\%$ so we look out for the value z corresponding to $\frac{c}{2}$ on the Student's distribution with $k - 1$ degrees of freedom, so we have $z = 1.984$; since t is less than $-z$ we can reject the null hypothesis of the test.

Although the performance difference is not really high, it shows that the integration has an impact on the task; so this addition may be further explored using different argumentation semantics or different approaches to build the argumentation frameworks.

5.4. biLSTM

During the training process of this model we adopted a train set - validation set split in order to perform the tuning of the hyper-parameters, in particular to choose an appropriate value of batch size, learning rate and number of epochs. For the batch size the possible values are $\{32, 64, 128\}$ and for learning rate the possible values are $\{0, 0005, 0, 005\}$; while for the number of epochs, instead of using a set of candidate values, we fix the values to 10 and we adopt an early stopping if the log loss on the validation set increases due to overfitting. In addition, to decide whether to introduce the sp-ranking or not we consider this choice as a boolean hyper-parameter to tune contextually to the other parameters. So we report in table 2, for each test fold of a 5-fold cross validation the parameters selected after the tuning and the performance obtained after training the model with such parameters.

We show that in 2 out of 5 folds the sp-ranking is taken into account.

6. Conclusions

We explored the task of automatic evaluation of argumentative essays, proposing a text classification approach enriched with the usage of new feature computed by exploiting argumentative reasoning on a BWAf built on each essay following a general structure. We evaluated the usefulness of this feature (the sp-ranking

feature) when performing standard classification with Ada Boost and a biLSTM neural network and comparing the performance with and without the sp-ranking feature, showing improved results when considering the additional feature.

A drawback of this approach is represented by the complexity of applying argumentation, which may limit the computation of the ranking to small essays (ultimately structured as graphs).

To further investigate this topic, a more complex reasoning strategy could be considered, as for the case of a general argumentation framework [19]. Additionally, different measures of similarity could be taken into account for assessing the weights of the relations in the AF.

References

- [1] J. Wilson, G. N. Andrada, Using automated feedback to improve writing quality: Opportunities and challenges., in: *Handbook of Research on Technology Tools for Real-World Skill Development*, IGI Global, 2016, pp. 679–704.
- [2] S. Crossley, Y. Tian, Q. Wan, Argumentation features and essay quality: Exploring relationships and incidence counts, *Journal of Writing Research* 14 (2022) 1–34. URL: <https://www.jowr.org/index.php/jowr/article/view/831>. doi:10.17239/jowr-2022.14.01.01.
- [3] A. Pazienza, S. Ferilli, F. Esposito, Constructing and evaluating bipolar weighted argumentation frameworks for online debating systems, in: *Proceedings of the 1st Workshop on Advances In Argumentation In Artificial Intelligence (AI³ 2017)*, co-located with the XVI International Conference of the Italian Association for Artificial Intelligence (AI*IA 2017), volume 2012 of *Central Europe (CEUR) Workshop Proceedings*, 2017, pp. 111–125.
- [4] X. Wang, Y. Lee, J. Park, Automated evaluation for student argumentative writing: A survey, 2022.
- [5] C. Stab, I. Gurevych, Parsing Argumentation Structures in Persuasive Essays, *Computational Linguistics* 43 (2017) 619–659. URL: https://doi.org/10.1162/COLI_a_00295. doi:10.1162/COLI_a_00295.
- [6] C. Stab, I. Gurevych, Recognizing insufficiently supported arguments in argumentative essays, in: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 980–990. URL: <https://aclanthology.org/E17-1092>.
- [7] B. Beigman Klebanov, B. Gyawali, Y. Song, Detecting good arguments in a non-topic-specific way: An oxymoron?, in: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 244–249. URL: <https://aclanthology.org/P17-2038>. doi:10.18653/v1/P17-2038.
- [8] J. W. G. Putra, S. Teufel, T. Tokunaga, Parsing argumentative structure in English-as-foreign-language essays, in: *Proceedings of the 16th Workshop on*

- Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics, Online, 2021, pp. 97–109. URL: <https://aclanthology.org/2021.bea-1.10>.
- [9] Y. Ye, S. Teufel, End-to-end argument mining as biaffine dependency parsing, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 669–678. URL: <https://aclanthology.org/2021.eacl-main.55>. doi:10.18653/v1/2021.eacl-main.55.
- [10] H. Wang, Z. Huang, Y. Dou, Y. Hong, Argumentation mining on essays at multi scales, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 5480–5493. URL: <https://aclanthology.org/2020.coling-main.478>. doi:10.18653/v1/2020.coling-main.478.
- [11] I. Persing, V. Ng, Unsupervised argumentation mining in student essays, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 6795–6803. URL: <https://aclanthology.org/2020.lrec-1.839>.
- [12] W. Carlile, N. Gurrupadi, Z. Ke, V. Ng, Give me more feedback: Annotating argument persuasiveness and related attributes in student essays, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 621–631. URL: <https://aclanthology.org/P18-1058>. doi:10.18653/v1/P18-1058.
- [13] J. W. G. Putra, S. Teufel, T. Tokunaga, Annotating argumentative structure in english-as-a-foreign-language learner essays, Natural Language Engineering (2021) 1–27. doi:10.1017/S1351324921000218.
- [14] S. E. Toulmin, The Uses of Argument, University Press, 1958. URL: <https://books.google.it/books?id=WffWAAAAMAAJ>.
- [15] P. M. Dung, On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games, Artificial intelligence 77 (1995) 321–357.
- [16] C. Cayrol, M.-C. Lagasquie-Schiex, On the acceptability of arguments in bipolar argumentation frameworks, in: European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty, Springer, 2005, pp. 378–389.
- [17] P. E. Dunne, A. Hunter, P. McBurney, S. Parsons, M. Wooldridge, Weighted argument systems: Basic definitions, algorithms, and complexity results, Artificial Intelligence 175 (2011) 457–486.
- [18] L. Amgoud, J. Ben-Naim, Evaluation of arguments in weighted bipolar graphs, in: A. Antonucci, L. Cholvy, O. Papini (Eds.), Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU 2017), Springer, Cham, 2017, pp. 25–35.
- [19] S. Ferilli, Introducing general argumentation frameworks and their use, in: AIXIA 2020 (reboot) - The 19th International Conference of the Italian Asso-

- ciation for Artificial Intelligence, volume 12414 of *Lecture Notes in Artificial Intelligence*, Springer, 2021, pp. 136–153.
- [20] L. Amgoud, J. Ben-Naim, Ranking-based semantics for argumentation frameworks, in: *Scalable Uncertainty Management*, volume 8078 of *Lecture Notes in Computer Science (LNAI)*, Springer, 2013, pp. 134–147.
- [21] Q. V. Le, T. Mikolov, Distributed representations of sentences and documents, 2014. URL: <https://arxiv.org/abs/1405.4053>. doi:10.48550/ARXIV.1405.4053.
- [22] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, 2013. URL: <https://arxiv.org/abs/1301.3781>. doi:10.48550/ARXIV.1301.3781.
- [23] R. E. Schapire, Explaining adaboost, in: *Empirical inference*, Springer, 2013, pp. 37–52.
- [24] H. Sak, A. Senior, F. Beaufays, Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition, 2014. URL: <https://arxiv.org/abs/1402.1128>. doi:10.48550/ARXIV.1402.1128.
- [25] M. Schuster, K. K. Paliwal, Bidirectional recurrent neural networks, *IEEE transactions on Signal Processing* 45 (1997) 2673–2681.