Explaining Online Debate Evolution under Bipolar Gradual Argumentation Semantics

Caren Al Anaissy[a] and Nicolas Maudet[b]
LIP6, Sorbonne Université - CNRS, Paris, France
{*caren.al-anaissy, nicolas.maudet*}*@lip6.fr*

Keywords: Abductive Explanations, Online Debates, Argumentation Semantics.

Abstract: Online debates allow for online collective discussions essential for forming opinions, decisions, and policies within society. Computational argumentation plays an important role in structuring online debates and inferring conclusions in these debates. It allows to represent the arguments exchanged by the participants, the interactions between the arguments and the participants' votes on each argument. It proposes argumentation semantics that can be used to infer a debate's outcome by evaluating the exchanged arguments' strengths. However, in massive online debates, where a large number of arguments is exchanged, it becomes difficult for a participant to navigate through the debate and to understand the reasoning and computation behind the semantics. In this paper, we address this issue by generating abductive explanations (i.e. sufficient reasons) for the debates' evolution of outcomes computed by bipolar gradual semantics. We define intuitive strategies and heuristics to produce explanations addressing the questions: "why is a debate issue's final weight higher/lower than its initial weight under a specific semantics?". We illustrate our methodology, compare and analyze the different heuristics and strategies with respect to the size of their corresponding generated explanations, by conducting several experiments on a real-world dataset.

1 INTRODUCTION

Online collective discussions play an important role in enhancing the participants critical thinking and justification standards, understanding the public opinion and reaching collective decisions in online deliberation for e-democracy. Online debate platforms allow for collective discussions where web users can create debates, argue for or against a specific issue by posting "pro" and "con" arguments, and vote on other users' arguments. There are numerous online debate platforms, like Idebate's debatabase[1], Kialo[2], ProCon[3] or Make.org[4]. During the discussions on online debate platforms, many arguments are raised. Those arguments are the key part of the discussion and contain valuable information. However, due to the large number of arguments, it is time consuming to read them all. Discussions can sometimes shift away from the topic or be manipulated to spread misinformation. The more participants we have, the less structured, meaningful and rational the debates can become (Shortall et al., 2022), hence the need to provide sense-making and decision-aiding functionalities on these platforms. Abstract argumentation (Dung, 1995; Cayrol and Lagasquie-Schiex, 2005) is a popular approach to represent online debates in a structured manner by modeling a debate in a weighted bipolar argumentation framework where the exchanged "pro" and "con" arguments are represented as nodes and the attack and support relations between the arguments are represented as edges in this argumentation graph. When votes are allowed on arguments, they can be aggregated into a weight denoting the popularity or plausibility of the argument. This argument's initial weight can be considered as an initial bias for its "acceptability", independently of its interactions with the other arguments. Computational argumentation has developed several techniques that evaluate the final acceptability of arguments after consideration of the impact of the other arguments. Specifically, bipolar gradual semantics (Baroni et al., 2015; Rago et al., 2016; Amgoud and Ben-Naim, 2018; Potyka, 2018) compute for each argument its *final weight* based on its initial weight and on

[a] https://orcid.org/0000-0002-8750-1849
[b] https://orcid.org/0000-0002-4232-069X
[1] https://idebate.net/resources/debatabase
[2] https://www.kialo.com/
[3] https://www.britannica.com/procon
[4] https://make.org/NL

the strengths of its attackers and supporters.

It is crucial to ensure transparent and explainable decision-aiding processes for the participants in online debates (Grimmelikhuijsen, 2023). Although computational argumentation is considered interpretable—since it is possible in principle to trace the reasoning process through an argumentation graph—when these graphs become very large, it becomes difficult to present the entire reasoning path and computation to the user. In this paper, we started with the following research question: can we explain to an online debate's participant why a given bipolar gradual semantics produces a particular evaluation on a specific debate? However, instead of explaining the exact value computation, which is unlikely to make sense for the user anyway, we focus on the more qualitative question of the evolution of the weight of the initial claim after considering arguments put forward during the debate. In other words, did the debate overall contribute to strengthen the initial claim, or to weaken it?

The type of explanations we are after are *abductive explanations* (i.e. sufficient reasons) allowing to justify why the final weight of a debate main claim is higher or lower than its initial weight under a given computation technique (i.e. argumentation semantics). Technically, the abductive explanations we propose in this paper are *sub-tree-based*. A sub-tree of arguments connected to the debate issue, affects the debate issue's acceptability by strengthening or weakening it. In an online debate, participants can extend the debate's existing sub-trees at any time, by presenting new arguments. Most of the online debate platforms are thus designed in a way that constrains graphs to be tree-structured. Users can also add new votes, modifying the initial weights of arguments within the sub-trees. Hence, these evolving sub-trees affect the final weight of the debate issue. We believe that argument-sub-trees constitute natural elementary blocks of explanation in this context. However, it may not be necessary to explicitly show all the sub-trees to justify the evolution of a debate, and there may be several ways to put forward those sub-trees. This work explores these questions.

The contributions of this paper are the following:

- We define two heuristics and two strategies that generate abductive explanations under a well-known bipolar gradual argumentation semantics.
- We illustrate our methodology by providing experiments on Kialo debates.
- We analyze and compare the different heuristics and strategies with respect to the size of the produced explanations.
- We analyze the size of the generated explanations with respect to the size of the debate graphs following the different heuristics and strategies.
- We study the correlation between the debate issue's weight change induced by the semantics and the size of the returned explanation.

Our methodology is a contribution allowing participants to understand how online debates are assessed, hopefully increasing users' trust and engagement. The full code and experimental results are available at this GitHub repository.

2 PRELIMINARIES

We first give some basic notions.

Definition 1 (Weighted Bipolar Argumentation Framework)**.** *A weighted bipolar argumentation framework WBAF is a quadruple $\langle Ar, att, sup, w\rangle$ where Ar is a finite set of arguments, $att, sup \subseteq Ar \times Ar$ are binary relations over Ar called attack and support respectively, and w is a weighting function $w : Ar \to [0,1]$, that associates for each argument $t \in Ar$ its initial weight $w(t) \in [0,1]$. We denote by $\mathtt{Sup}(t)$ (resp. $\mathtt{Att}(t)$) the set of t's supporters (resp. attackers).*

Definition 2 (Restriction)**.** *The restriction of a WBAF $\mathcal{F} = \langle Ar, att, sup, w\rangle$ to a subset of its arguments $Args \subseteq Ar$, denoted as $\mathcal{F}\downarrow_{Args}$ is $\langle Args, att \cap (Args \times Args), sup \cap (Args \times Args), w \cap (Args \times [0,1])\rangle$.*

Definition 3 (Path)**.** *Let $\mathcal{F} = \langle Ar, att, sup, w\rangle$ be a WBAF, $t, u \in Ar$. A path from t to u, denoted by $P(t,u)$ is a sequence of arguments $t_0, \ldots, t_n \in Ar$ such that $t = t_0$ and $u = t_n$, and for each i, $0 \leq i \leq n-1$, t_i attacks or supports t_{i+1}.*

Bipolar gradual semantics evaluate the strength of each argument in a *WBAF* by aggregating its initial weight and the strengths of its attackers and supporters, using an iterative procedure that updates all arguments' strengths until they converge to their final weights. Most of the state-of-the-art bipolar gradual semantics are defined for acyclic graphs, where the arguments are hence processed in a topological order.

Definition 4 (Bipolar Gradual Semantics)**.** *A bipolar gradual semantics σ for WBAFs is a function that maps a WBAF $\mathcal{F} = \langle Ar, att, sup, w\rangle$ into a weighting on Ar, $\sigma_{\mathcal{F}} : Ar \to [0,1]$. $\sigma_{\mathcal{F}}(t)$ stands for the weight assigned by $\sigma_{\mathcal{F}}$, called final weight of t.*

Definition 5 (Wothless/Significant Argument)**.** *Let $\mathcal{F} = \langle Ar, att, sup, w\rangle$ be a WBAF. Let $\sigma_{\mathcal{F}}$ be a bipolar gradual semantics. An argument $t \in Ar$ is called worthless (resp. significant) if $\sigma_{\mathcal{F}}(t) = 0$ (resp. $\neq 0$).*

We denote by $\texttt{sAtt}(t)$ *(resp.* $\texttt{sSup}(t)$*) the set of all t's significant attackers (resp. significant supporters).*

Several bipolar gradual semantics have been defined in the literature namely, Quantitative Argumentation Debate (QuAD) (Baroni et al., 2015), Discontinuity-Free Quantitative Argumentation Debate (DF-QuAD) (Rago et al., 2016), Exponent-based (Exb) (Amgoud and Ben-Naim, 2018), and Quadratic Energy Model (QEM) (Potyka, 2018). In this paper, we illustrate our work using the QEM semantics since it does not present the undesired behaviors of the QuAD, DF-QuAD and Exb semantics (Amgoud and Ben-Naim, 2018; Potyka, 2018).

Definition 6 (QEM, Quadratic Energy Model (Potyka, 2018))**.** *Let* $\mathcal{F} = \langle Ar, att, sup, w\rangle$ *be a WBAF,* $t \in Ar$*,* $e \in \mathbb{R}$*, the impact of e is given by h such that:*

$$h\colon \mathbb{R} \to [0,1]$$

$$h(e) = \frac{max(e,0)^2}{1+max(e,0)^2}$$

The final weight of t computed by QEM is defined as:

$$QEM_{\mathcal{F}}(t) = \begin{cases} w(t) + (1-w(t)) \cdot h(E(t)) & \text{if } E(t) > 0 \\ w(t) - w(t) \cdot h(-E(t)) & \text{otherwise} \end{cases}$$

where E is the energy at argument t computed as $E(t) = \sum_{u \in \texttt{Sup}(t)} QEM_{\mathcal{F}}(u) - \sum_{u \in \texttt{Att}(t)} QEM_{\mathcal{F}}(u)$

Amgoud et al. define several principles (Amgoud and Ben-Naim, 2018) to evaluate bipolar gradual semantics. In this paper, we present only those necessary for our work. Due to space restrictions, we explain them briefly and we invite the reader to check their formal definitions. *Bi-variate Independence* states that an argument's final weight is independent of the arguments with which it has no path (ignoring the direction of the edges). *Bi-variate Directionality* states that an argument's final weight depends only on the arguments connected by paths directed to it. *Weakening* (resp. *Strengthening*) states that an argument's final weight becomes lower (resp. higher) than its initial weight if its attackers are stronger (resp. weaker) than its supporters. *Reinforcement* states that an argument's final weight does not decrease if the attackers become weaker and the supporters become stronger. *Neutrality* states that an argument's final weight is unaffected by worthless attackers or worthless supporters. Note that QEM satisfies all these principles for any *WBAF* (Potyka, 2018).

3 DEBATES AS WBAFs

Let $\mathcal{F} = \langle Ar, att, sup, w\rangle$ be a **tree** *WBAF*, $\sigma_{\mathcal{F}}$ be a bipolar gradual semantics that satisfies bi-variate directionality, bi-variate independence, weakening, strengthening, reinforcement and neutrality. Let $t \in Ar$ be the debate issue. Figure 1 shows the graphical representation of a Kialo debate as a *WBAF*.

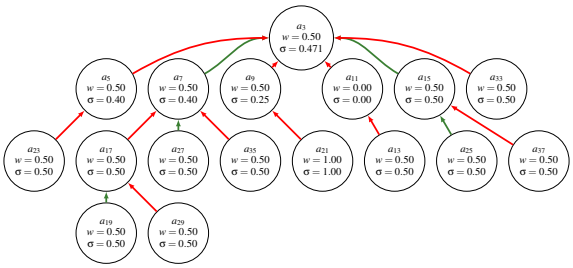

Figure 1: A debate from Kialo represented as a *WBAF*. The arguments are represented by nodes and the green (resp. red) arrows represent the supports (resp. attacks) between the arguments. For each argument, the first, second and third row of labels represent its id, its initial weight computed by aggregating the votes on it and its final weight computed by QEM. The debate issue a_3: It takes a nation to defend a nation.

A sub-tree is called a top-level sub-tree iff its root attacks or supports the debate issue. A sub-tree is called full iff it contains all the arguments connected by a path to its root otherwise, partial.

Example 1. *Consider the tree WBAF* $\mathcal{F} = \langle Ar, att, sup, w\rangle$ *depicted in Figure 1.* $\{a_7, a_{17}, a_{27}, a_{35}, a_{19}, a_{29}\}$ *is a full top-level sub-tree of* $\mathcal{F}$*,* $\{a_7, a_{17}, a_{27}, a_{19}\}$ *is a partial top-level sub-tree of* $\mathcal{F}$*,* $\{a_{17}, a_{19}, a_{29}\}$ *is a full sub-tree of* $\mathcal{F}$ *and* $\{a_{17}, a_{19}\}$ *is a partial sub-tree of* $\mathcal{F}$*.*

Definition 7 (Neutral Argument)**.** *An argument* $x \in Ar$ *is called a neutral argument iff it is not connected by any path to t, or is a worthless argument, or is connected to a worthless argument by a path.*

Note that since $\sigma_{\mathcal{F}}$ satisfies bi-variate independence, bi-variate directionality and neutrality, the removal of a neutral argument from $\mathcal{F}$ does not affect t's final weight.

Definition 8 (Pro/Con-argument)**.** *An argument* $x \in Ar$ *is called a pro-argument (resp. con-argument) iff it is not a neutral argument and its removal from* $\mathcal{F}$ *does not increase (resp. does not decrease) t's final weight, i.e.* $\sigma_{\mathcal{F}\downarrow_{Ar\setminus\{x\}}}(t) \leq \sigma_{\mathcal{F}}(t)$ *(resp.* $\sigma_{\mathcal{F}\downarrow_{Ar\setminus\{x\}}}(t) \geq \sigma_{\mathcal{F}}(t)$*).*

Example 2. *Consider* $\mathcal{F}$ *of Figure 1, let* $X, Y \in Ar$ *such that* $X = \{a_{23}, a_7, a_{27}, a_{29}, a_{21}, a_{15}, a_{25}\}$ *and* $Y = \{a_5, a_{17}, a_{19}, a_{35}, a_9, a_{37}, a_{33}\}$*. X contains all the pro-arguments of* $\mathcal{F}$ *and Y contains all the con-arguments of* $\mathcal{F}$*.* a_{11} *and* a_{13} *are neutral arguments since* a_{11} *is a worthless attacker of* a_3 *and* a_{13} *is connected to* a_{11}*.*

Note that top-level sub-trees can be classified as supporting or attacking depending on whether their root supports or attacks the main claim of the debate.

Definition 9 (Pro/Con-top-level Sub-trees)**.** *A pro-top-level sub-tree Pst_i (resp. con-top-level sub-tree Cst_i) of $\mathcal{F}$ is a full top-level sub-tree whose root is a significant supporter (resp. significant attacker) of t.*

Definition 10 (Unweakened Pro/Con-top-level Sub-trees)**.** *An unweakened pro-top-level sub-tree $UnwPst_i$ (resp. unweakened con-top-level sub-tree $UnwCst_i$) of $\mathcal{F}$ is a maximal top-level sub-tree of $\mathcal{F}$ (w.r.t. set inclusion), whose root is a significant supporter (resp. significant attacker) of t and whose all arguments are pro-arguments (resp. con-arguments).*

Definition 11 (Pro/Con-weakening Sub-trees)**.** *A pro-weakening (resp. con-weakening) sub-tree $PWst_i$ (resp. $CWst_i$) of $\mathcal{F}$ is a full sub-tree of $\mathcal{F}$ whose root is a significant attacker of one of the pro-arguments (resp. con-arguments) of an unweakened pro-top-level (resp. unweakened con–top-level) sub-tree of $\mathcal{F}$.*

Definition 12 (Weakened Pro/Con-top-level Sub--trees)**.** *A weakened pro-top-level sub-tree (resp. con-top-level sub-tree) of $\mathcal{F}$ is the union of an unweakened pro-top-level sub-tree (resp. unweakened con-top-level sub-tree) with at least one of its corresponding pro-weakening (resp. con-weakening) sub-trees.*

Example 3. *Consider the WBAF $\mathcal{F}$ of Figure 1, $\mathcal{F}$'s pro/con-top-level sub-trees, unweakened pro/con-top-level sub-trees, and pro/con-weakening sub-trees are:*

Pst_i	$\{a_7, a_{17}, a_{19}, a_{29}, a_{27}, a_{35}\}$	$\{a_{15}, a_{25}, a_{37}\}$	
Cst_i	$\{a_5, a_{23}\}$	$\{a_9, a_{21}\}$	$\{a_{33}\}$
$UnwPst_i$	$\{a_7, a_{27}\}$	$\{a_{15}, a_{25}\}$	
$UnwCst_i$	$\{a_5\}$	$\{a_9\}$	$\{a_{33}\}$
$PWst_i$	$\{a_{35}\}$	$\{a_{17}, a_{19}, a_{29}\}$	$\{a_{37}\}$
$CWst_i$	$\{a_{23}\}$	$\{a_{21}\}$	

Definition 13 (Sub-tree Strength)**.** *Let B be a sub-tree of $\mathcal{F}$, $x \in Ar$ be B's root, $a \in Ar$ be the argument to which x is connected. The strength of B is the effect of its removal from $\mathcal{F}$ on a's final weight. The stronger (resp. weaker) B is, the more (resp. less) a's final weight decreases/increases when removing B from $\mathcal{F}$.*

Observation 1. *Since we are working on tree graphs, the strength of B is determined by the effect of removing x from $\mathcal{F}$ on a's final weight. This is because when x is removed from $\mathcal{F}$, all sub-trees rooted at x are also removed. As a result, every argument connected to a through x will be eliminated. Since the effect of removing x from $\mathcal{F}$ on a's final weight is determined by x' s final weight, and since $\sigma_{\mathcal{F}}$ satisfies reinforcement, when x's final weight increases (resp. decreases), B's strength cannot decrease (resp. increase). More specifically, if $0 < \sigma_{\mathcal{F}}(a) < 1$, the higher (resp. lower) $\sigma_{\mathcal{F}}(x)$ is, the stronger (resp. weaker) B is.*

Definition 14 (Proponent/Opponent Sub-tree)**.** *A sub-tree is called proponent (resp. opponent) iff its removal from $\mathcal{F}$ does not increase (resp. does not decrease) t's final weight, i.e., iff its root is a pro-argument (resp. con-argument).*

Observation 2. *Since $\sigma_{\mathcal{F}}$ satisfies weakening and strengthening, the pro-top-level, unweakened and weakened pro-top-level, and con-weakening sub-trees are proponent sub-trees while the con-top-level, unweakened and weakened con-top-level, and pro-weakening sub-trees are opponent sub-trees.*

4 EXPLANATION GENERATION

We now address the question of generating explanations regarding the evolution of such debates. We distinguish between the *strengthening case* where t's final weight has increased and the *weakening case* where t's final weight has decreased (with respect to its initial weight) after applying σ to the graph. In a nutshell, a sufficient explanation for an increase would consist of exhibiting opponent sub-trees but showing sufficiently strong proponent sub-trees to justify an increase (and dually for a decrease). The exact choice of which sub-trees to present is the matter of heuristics and strategies. The main idea behind our approach is to build explanations on the basis of top-level sub-trees.

4.1 Heuristics

We define two heuristics, $H_{s\to w}$ and $H_{s\to l}$, that allow to determine the minimality condition for generating the explanation. Each heuristic classifies a category's sub-trees under a specified ranking. $H_{s\to w}$ classifies a category's sub-trees from the strongest to the weakest sub-tree. It allows to return a minimal explanation w.r.t. the number of sub-trees, i.e. to return the minimal number of sub-trees that let t's final weight surpass (resp. fall short of) its initial weight for the strengthening case (resp. weakening case). $H_{s\to l}$ classifies each category's sub-trees from the smallest sub-tree, i.e. having the smallest number of arguments, to the largest sub-tree, i.e. having the largest number of arguments. It allows to minimize the explanation based on the number of returned arguments.

4.2 Strategies

We define below, for each case (strengthening/weakening), the constructive strategy and the destructive strategy. Each strategy first selects the categories of sub-trees to consider, ranks the sub-trees

within each category according to one of the defined heuristics, then selects which sub-trees to return as an explanation following the chosen heuristic. This results in four explanation-heuristic combinations.

At a high level, we can see that there are three types of sub-trees that can be distinguished for explanation generation. Suppose without loss of generality that we want to explain why the weight of the debate issue has increased: we either start by considering the *worst-case situation* of the debate consisting of all the unweakened con-top-level sub-trees, then we present pro-top-level sub-trees, (and then weakening counter-arguments to the unweakened con-top-level sub-trees if necessary), or we start by considering the *maximal weakening scenario*, where we present first all the unweakened con-top-level sub-trees with all the counter-arguments to those sub-trees to weaken them, (and then pro-top-level sub-trees). The first approach corresponds to the *constructive* strategy, while the second corresponds to the *destructive* strategy. We denote by $E = Cons\_Exp$ (resp. $E = Des\_Exp$) the generated constructive (resp. destructive) explanation following a heuristic $H \in [H_{s\to w}, H_{s\to l}]$. Let Pst (resp. Cst, $UnwPst$, $UnwCst$, $PWst$, $CWst$) be the set of arguments of all the pro-top-level (resp. con-top-level, unweakened pro-top-level, unweakened con-top-level, pro-weakening, con-weakening) sub-trees of $\mathcal{F}$, $UnwCst \cup CWst = Cst$, $UnwPst \cup PWst = Pst$.

4.2.1 Constructive Strategy, $\sigma_{\mathcal{F}}(t) < w(t)$

The constructive strategy selects the following three categories (sets) of sub-trees: $UnwPst$ as the set of proponent sub-trees, then Cst and $PWst$ as the sets of opponent sub-trees. It uses H to rank each category of the opponent sub-trees.

Con-Top-Level Sub-Trees Addition Phase: The constructive strategy considers first all the unweakened pro-top-level sub-trees and the ranked con-top-level sub-trees. A constructive explanation for why $\sigma_{\mathcal{F}}(t) < w(t)$ consists of finding, following H, the minimal number of con-top-level sub-trees whose effect on t surpasses that of the unweakened pro-top-level sub-trees. This is equivalent to finding, following H, the minimal number of t's significant attackers whose effect on t is strictly larger than that of all of t's significant supporters when they are unattacked, hence unweakened. If we successfully find this minimal number of significant attackers (con-top-level sub-trees) following H, since $\sigma_{\mathcal{F}}$ satisfies strengthening, we have a guarantee that the presence of only these con-top-level sub-trees in the graph, alongside all the unweakened pro-top-level sub-trees, is sufficient to let t's final weight be smaller than its initial weight. We return as a constructive explanation for why $\sigma_{\mathcal{F}}(t) < w(t)$, the set of all the unweakened pro-top-level sub-trees and the set of the minimal number of con-top-level sub-trees selected following H.

Example 4. *We apply the con-top-level sub-trees addition phase of the constructive strategy to the WBAF of Figure 1. Since* $\sigma_{\mathcal{F}}(a_3) = 0.471 < w(a_3) = 0.5$, *we take the unweakened pro-top-level sub-trees of* $\mathcal{F}$*:* $UnwPst_1 = \{a_7, a_{27}\}$, $UnwPst_2 = \{a_{15}, a_{25}\}$, *the con-top-level sub-trees of* $\mathcal{F}$*:* $Cst_1 = \{a_5, a_{23}\}$, $Cst_2 = \{a_9, a_{21}\}$, $Cst_3 = \{a_{33}\}$ *and the pro-weakening sub-trees of* $\mathcal{F}$*:* $PWst_1 = \{a_{35}\}$, $PWst_2 = \{a_{17}, a_{19}, a_{29}\}$, $PWst_3 = \{a_{37}\}$. *We choose for an illustration* $H_{s\to w}$. *Following* $H_{s\to w}$, *the ranking of the con-top-level sub-trees is* $[Cst_3, Cst_1, Cst_2]$ *and that of the pro-weakening sub-trees is* $[PWst_2, PWst_1, PWst_3]$. *Note that we rank sub-trees with similar strength in a random manner. Figure 2 shows the resulting WBAF.*

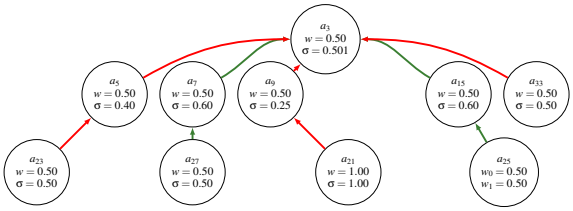

Figure 2: *WBAF* corresponding to the con-top-level sub-trees addition phase of the constructive strategy.

1. $cum\_Cst_0 = \emptyset$, $\mathcal{F}' = \mathcal{F} \downarrow_{UnwPst \cup \{a_3\}}$, $\sigma_{\mathcal{F}'}(a_3) \approx 0.795 > w(a_3)$
2. $cum\_Cst_1 = Cst_3$, $\mathcal{F}' = \mathcal{F} \downarrow_{UnwPst \cup cum\_Cst_1 \cup \{a_3\}}$ $\sigma_{\mathcal{F}'}(a_3) \approx 0.664 > w(a_3)$
3. $cum\_Cst_2 = cum\_Cst_1 \cup Cst_1$, $\mathcal{F}' = \mathcal{F} \downarrow_{UnwPst \cup cum\_Cst_2 \cup \{a_3\}}$, $\sigma_{\mathcal{F}'}(t) \approx 0.541 > w(a_3)$
4. $cum\_Cst_3 = cum\_Cst_2 \cup Cst_2, \mathcal{F}' = \mathcal{F} \downarrow_{UnwPst \cup cum\_Cst_3 \cup \{a_3\}}, \sigma_{\mathcal{F}'}(a_3) \approx 0.501 > w(a_3)$

Note that even after adding all the con-top-level sub-trees to the unweakened pro-top-level ones, a_3*'s final weight is still larger than its initial weight. To let* a_3*'s final weight fall below* $w(a_3)$, *we move to the pro-weakening sub-trees addition phase explained below.*

Pro-Weakening Sub-Trees Addition Phase: If the constructive explanation is not found during the con-top-level sub-trees addition phase, i.e. the effect of **all** the con-top-level sub-trees on t does not surpass that of all the unweakened pro-top-level sub-trees, the strategy considers all the unweakened pro-top-level and all the con-top-level sub-trees in the graph, then adds progressively the pro-weakening sub-trees to the unweakened pro-top-level sub-trees, following H. We consider hence weaker versions of the unweakened pro-top-level sub-trees. We stop adding pro-weakening sub-trees when the effect of all the con-top-level sub-trees on t is strictly greater than that

of the cumulative set of weakened pro-top-level sub-trees. Once we achieve this, the constructive explanation for why $\sigma_{\mathcal{F}}(t) < w(t)$ is then the union of the cumulative set of the weakened pro-top-level sub-trees, i.e. the unweakened pro-top-level sub-trees and some pro-weakening sub-trees, with the set of all the con-top-level sub-trees.

Example 5. *In Example 4, we applied the con-top-level sub-trees addition phase to the WBAF of Figure 1. Since the constructive explanation for why $\sigma_{\mathcal{F}}(a_3) < w(a_3)$ was not found during this phase, we apply below the pro-weakening sub-trees addition phase. Figure 3 shows the resulting WBAF.*

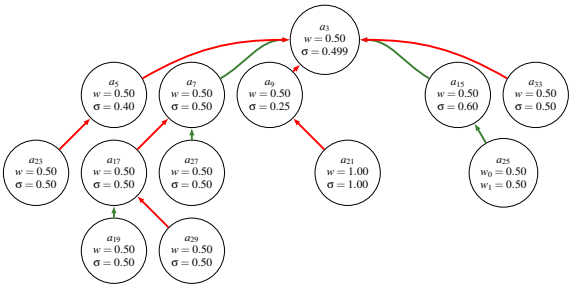

Figure 3: *WBAF* corresponding to the Pro-weakening sub-trees addition phase of the constructive strategy.

1. $cum\_PWst_0 = \emptyset$, $cum\_PWst_1 = PWst_2$, $\mathcal{F}' = \mathcal{F}\downarrow_{UnwPst \cup cum\_PWst_1 \cup Cst \cup \{a_3\}}$, $\sigma_{\mathcal{F}'}(a_3) \approx 0.499 < w(a_3) = 0.5$
2. $Cons\_Exp = UnwPst \cup cum\_PWst_1 \cup Cst = UnwPst \cup PWst_2 \cup Cst$.

4.2.2 Constructive Strategy, $\sigma_{\mathcal{F}}(t) > w(t)$

The constructive strategy selects $UnwCst$ as the set of opponent sub-trees, then Pst and $CWst$ as the sets of proponent sub-trees. It uses H to rank each category of the proponent sub-trees. It uses the same logic for when $\sigma_{\mathcal{F}}(t) < w(t)$. The constructive explanation consists of searching for the minimal number, following H, of pro-top-level sub-trees whose effect surpasses that of the unweakened con-top-level sub-trees (pro-top-level sub-trees addition phase). If not found, it searches for the minimal number, following H, of con-weakening sub-trees that weakens the unweakened con-top-level sub-trees enough to allow for all the pro-top-level sub-trees to surpass this effect (con-weakening sub-trees addition phase).

Algorithm 1 outlines the constructive strategy for generating $Cons\_Exp$ for $\sigma_{\mathcal{F}}(t) > w(t)$. The algorithm for $\sigma_{\mathcal{F}}(t) < w(t)$ follows the same algorithmic pattern and is therefore omitted. Let $[Pst_1, \cdots, Pst_n]$ and $[CWst_1, \cdots, CWst_m]$ be the rankings of all the pro-top-level sub-trees and all the con-weakening sub-trees of $\mathcal{F}$ following H. Let cum_Pst_i be the cumulative pro-top-level sub-trees set at an iteration $i = 0, \ldots, n$ and cum_CWst_j be the cumulative con-weakening sub-trees set at iteration $j = 0, \ldots, m$.

4.2.3 Destructive Strategy, $\sigma_{\mathcal{F}}(t) < w(t)$

The destructive strategy selects Pst (resp. Cst) as the set of proponent (resp. opponent) top-level sub-trees. It classifies the con-top-level sub-trees following H. The idea behind finding a destructive explanation for why $\sigma_{\mathcal{F}}(t) < w(t)$ is to search for the minimal number of con-top-level sub-trees, following H, whose effect surpasses that of all the pro-top-level sub-trees. Remember that the constructive strategy takes first t's significant supporters at their strongest state, i.e. the unweakened pro-top-level sub-trees, then if needed, it incrementally weakens the unweakened pro-top-level sub-trees with pro-weakening sub-trees. However, the destructive strategy considers the maximal weakening scenario by taking all of t's significant supporters at their weakest state, i.e. by taking all the pro-top-level sub-trees which consist of all the unweakened pro-top-level sub-trees and all their pro-weakening sub-trees $Pst = UnwPst \cup PWst$. Then its adds the con-top-level sub-trees incrementally following H, until t's final weight becomes strictly smaller than its initial weight. As a destructive explanation, it returns the set containing all the pro-top-level sub-trees, and the minimal number of con-top-level sub-trees needed following H.

Example 6. *We apply the destructive strategy to the WBAF of Figure 1. Since $\sigma_{\mathcal{F}}(a_3) < w(a_3)$, we take the pro-top-level sub-trees at their weakest strengths, $Pst_1 = \{a_7, a_{17}, a_{19}, a_{29}, a_{27}, a_{35}\}$ and $Pst_2 = \{a_{15}, a_{25}, a_{37}\}$. The con-top-level sub-trees of $\mathcal{F}$ are $Cst_1 = \{a_5, a_{23}\}$, $Cst_2 = \{a_9, a_{21}\}$ and $Cst_3 = \{a_{33}\}$. Following $H_{s\to l}$, we have the ranking: $[Cst_3, Cst_1, Cst_2]$. Note that we rank sub-trees that have similar number of arguments in a random manner. Figure 4 shows the resulting WBAF after the application of the destructive strategy.*

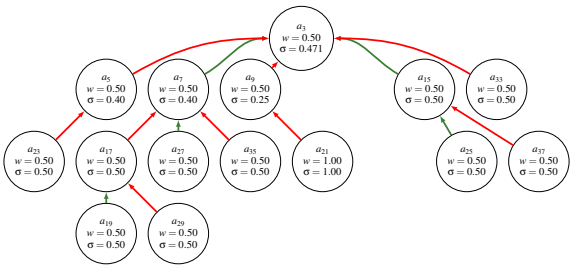

Figure 4: *WBAF* corresponding to the destructive strategy.

1. $cum\_Cst_0 = \emptyset$, $\mathcal{F}' = \mathcal{F}\downarrow_{Pst \cup \{a_3\}}$, $\sigma_{\mathcal{F}'}(a_3) \approx 0.724 > w(a_3) = 0.5$

Algorithm 1: Constructive strategy for generating a constructive explanation for the strengthening case.

Input: $\mathcal{F}, \sigma_{\mathcal{F}}, t, H$
Output: *Cons_Exp*
Create initial restriction: $\mathcal{F}' = \mathcal{F} \downarrow_{UnwCst \cup \{t\}}$;
Compute $\sigma_{\mathcal{F}'}(t)$;
$i \leftarrow 0$;
Initialize $cum\_Pst_i = cum\_Pst_0 = \emptyset$;
while $\sigma_{\mathcal{F}'}(t) \leq w(t)$ **do**
  // Pro-top-level sub-trees addition phase
  $i \leftarrow i+1$;
  Select next pro-top-level sub-tree Pst_i following H;
  Update $cum\_Pst_i = cum\_Pst_{i-1} \cup Pst_i$;
  Update restriction: $\mathcal{F}' = \mathcal{F} \downarrow_{UnwCst \cup cum\_Pst_i \cup \{t\}}$;
  Compute $\sigma_{\mathcal{F}'}(t)$;
  if $\sigma_{\mathcal{F}'}(t) > w(t)$ **then**
    return $Cons\_Exp = UnwCst \cup cum\_Pst_i$ // Success
  end
  if $\sigma_{\mathcal{F}'}(t) \leq w(t)$ *and* $cum\_Pst_i = Pst$ **then**
    // Con-weakening sub-trees addition phase
    $j \leftarrow 0$;
    Initialize $cum\_CWst_j = cum\_CWst_0 = \emptyset$;
    while $\sigma_{\mathcal{F}'}(t) \leq w(t)$ **do**
      $j \leftarrow j+1$;
      Select next con-weakening sub-tree $CWst_j$ following H;
      Update $cum\_CWst_j = cum\_CWst_{j-1} \cup CWst_j$;
      Update restriction: $\mathcal{F}' = \mathcal{F} \downarrow_{UnwCst \cup cum\_CWst_j \cup Pst \cup \{t\}}$;
      Compute $\sigma_{\mathcal{F}'}(t)$;
      if $\sigma_{\mathcal{F}'}(t) > w(t)$ **then**
        return $Cons\_Exp = UnwCst \cup cum\_CWst_j \cup Pst$ // success
      end
      if $\sigma_{\mathcal{F}'}(t) \leq w(t)$ *and* $cum\_CWst_j = CWst$ **then**
        return $Cons\_Exp = \emptyset$ // No constructive explanation found
      end
    end
  end
end

2. $cum\_Cst_1 = Cst_3$, $\mathcal{F}' = \mathcal{F} \downarrow_{Pst \cup cum\_Cst_1 \cup \{a_3\}}$ $\sigma_{\mathcal{F}'}(a_3) \approx 0.569 > w(a_3)$
3. $cum\_Cst_2 = Cst_3 \cup Cst_1$, $\mathcal{F}' = \mathcal{F} \downarrow_{Pst \cup cum\_Cst_2 \cup \{a_3\}}$ $\sigma_{\mathcal{F}'}(a_3) = w(a_3) = 0.5$
4. $cum\_Cst_3 = Cst_3 \cup Cst_1 \cup Cst_2$, $\mathcal{F}' = \mathcal{F} \downarrow_{Pst \cup cum\_Cst_3 \cup \{a_3\}}$, $\sigma_{\mathcal{F}'}(a_3) \approx 0.471 < w(a_3)$
5. $Des\_Exp = Pst \cup Cst$.

In this example, we could not minimize the returned explanation with $H_{s \to l}$, so we returned all the pro-top-level and con-top-level sub-trees as an explanation.

4.2.4 Destructive Strategy, $\sigma_{\mathcal{F}}(t) > w(t)$

The destructive strategy also selects *Pst* (resp. *Cst*) as the set of proponent (resp. opponent) top-level sub-trees. It searches however for the minimal number of pro-top-level sub-trees, following H, whose effect surpasses that of all the con-top-level sub-trees. Algorithm 2 outlines the destructive strategy for generating Des_Exp for $\sigma_{\mathcal{F}}(t) > w(t)$. The algorithm for $\sigma_{\mathcal{F}}(t) < w(t)$ follows the same algorithmic pattern and is therefore omitted.

Algorithm 2: Destructive strategy for generating a destructive explanation for the strengthening case.

Input: $\mathcal{F}, \sigma_{\mathcal{F}}, t, H$
Output: *Des_Exp*
Create initial restriction $\mathcal{F}' = \mathcal{F} \downarrow_{Cst \cup \{t\}}$;
Compute $\sigma_{\mathcal{F}'}(t)$;
$i \leftarrow 0$;
Initialize $cum\_Pst_i = cum\_Pst_0 = \emptyset$;
while $\sigma_{\mathcal{F}'}(t) \leq w(t)$ **do**
  $i \leftarrow i+1$;
  Select next pro-top-level sub-tree Pst_i following H;
  Update $cum\_Pst_i = cum\_Pst_{i-1} \cup Pst_i$;
  Update restriction: $\mathcal{F}' = \mathcal{F} \downarrow_{Cst \cup cum\_Pst_i \cup \{t\}}$;
  Compute $\sigma_{\mathcal{F}'}(t)$;
  if $\sigma_{\mathcal{F}'}(t) > w(t)$ **then**
    return $Des\_Exp = Cst \cup cum\_Pst_i$ // Success
  end
  if $\sigma_{\mathcal{F}'}(t) \leq w(t)$ *and* $cum\_Pst_i = Pst$ **then**
    return $Des\_Exp = \emptyset$ // No destructive explanation found
  end
end

Observation 3. *Since $\sigma_{\mathcal{F}}$ satisfies weakening and strengthening, when $\sigma_{\mathcal{F}}(t) > w(t)$ (resp. $\sigma_{\mathcal{F}}(t) < w(t)$), it means that t's significant supporters are*

stronger (resp. weaker) that t's significant attackers, i.e. the pro-top-level sub-trees are stronger (resp. weaker) than the con-top-level sub-trees. This guarantees that for both cases (strengthening/ weakening), the constructive and destructive strategies always return non-null explanations, since each strategy will return, in the worst case, the maximal explanation showing all the pro-top-level and all the con-top-level sub-trees of $\mathcal{F}$. When the constructive strategy returns a maximal explanation, this explanation coincides with the generated destructive explanation.

5 APPROACH

5.1 Dataset

We illustrate our work on debates from the Kialo online platform. Each debate starts with an issue or a question presented by a user as the initial claim. Users can then contribute by adding claims that either respond directly to the issue—if the debate is closed-ended—or address specific alternatives—if the debate is open-ended. Each claim supports or attacks the claim to which it responds. The underlying structure of any debate in Kialo is a tree. Users can vote on each claim using a 5-point scale to rate the argument's impact[5], so each claim is presented with its corresponding vote distribution. When the debate is open-ended, the alternatives are not connected to the debate issue by any relation. In this work, we consider each tree graph rooted at an alternative and its connected arguments as a sub-debate, where the alternative is the sub-debate issue. This results in 5273 (sub-)debates extracted from 2959 scraped debates. We aggregate the votes on each claim (argument) into it's initial weight according to Young et al.'s formula (Young et al., 2021), then we apply QEM to all the debates. Among these debates, we consider for our experimentation only the ones whose issue's initial weight has changed after applying *QEM*, i.e., 4326 debates.

5.2 Method

We applied both constructive and destructive algorithms to the 4326 debate graphs. Among these graphs, 2529 graphs fall into the strengthening case and the 1797 other graphs fall into the weakening case. For each of the 4326 graphs, we computed four explanations; for each of the two heuristics, we computed a constructive and a destructive explanation.

[5]https://support.kialo-edu.com/en/hc/about-voting/#understanding-impact

Table 1: Descriptive statistics of the argument coverage by the generated explanations on the Kialo dataset under the QEM semantics. The first and the second columns represent respectively the strategy and the heuristic used when generating the explanation. The third, fourth and fifth columns give respectively the mean, median and standard deviation of the argument coverage by the generated explanation across the 4326 graph debates of the Kialo dataset.

Strategy	Heuristic	Mean	Median	Std Dev
Constructive	$H_{s\to w}$	72.84	75	22.33
Constructive	$H_{s\to l}$	69.68	72.97	25.75
Destructive	$H_{s\to w}$	80.72	85.71	20.84
Destructive	$H_{s\to l}$	76.9	81.82	24.92

6 RESULTS AND DISCUSSION

6.1 Explanation Size

The explanation size is one of the key properties to assess the quality of an explanation in the explainable AI (XAI) literature (Nauta et al., 2023). The shorter an explanation is, the easier it is to understand. We define the size of E, $S(E)$, as the total number of returned arguments in the explanation: $S(E) = |E| + 1$ (we add the debate issue since we return it along with the explanation).

We define the argument coverage by E, $P(E)$, as the percentage of arguments covered by E, which is computed relative to the total number of arguments in the graph: $P(E) = \frac{S(E)}{|Ar|} \times 100$. For each of the four explanations generated for each debate graph, we computed $S(E)$ and $P(E)$. We computed the mean, the median and the standard deviation for the argument coverage $P(E)$ across all the 4326 debates' explanations that share the same type (constructive/destructive) and the same heuristic. We present these results in Table 1.

We draw from Table 1 the following observations: We can observe that for each explanation (constructive and destructive), $H_{s\to l}$ achieves a lower mean and a lower median than the mean and the median of $H_{s\to w}$. This is expected since the goal behind using $H_{s\to l}$ is to minimize the number of the returned arguments. However, note that when computing an explanation, if two sub-trees have the same number of arguments, $H_{s\to l}$ selects randomly one of the two sub-trees regardlessly of their strengths. Always selecting the strongest sub-tree among the two sub-trees might reduce the number of returned arguments, because the addition of the strongest sub-tree might be sufficient to let the debate issue's final weight surpass/fall short of its initial weight. We consider exploring this strategy in the future work.

For each heuristic, the argument coverage for con-

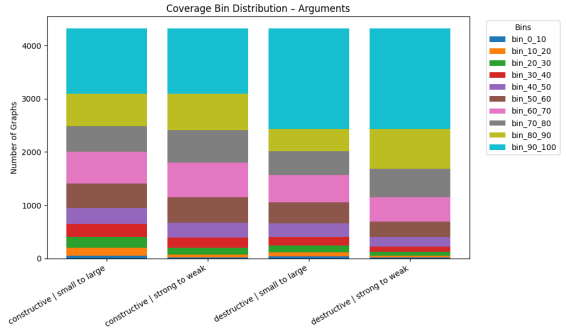

Figure 5: Distribution of argument coverage bins for explanation-heuristic combination. Each stacked bar shows, across the 4326 debate graphs, the number of graphs whose argument coverage falls into 10% -wide intervals $(0–10\%, 10–20\%, \cdots, 90–100\%)$ for the four explanation-heuristic combinations.

structive explanations, on average, is less than that for destructive explanations. This is expected since in the strengthening (resp. weakening) case, contrarily to the constructive strategy, the destructive strategy considers the *maximal weakening scenario* by considering *all* the con-top-level (resp. pro-top-level) sub-trees before cumulating the pro-top-level (resp. con-top-level) sub-trees.

For each explanation-heuristic combination, we can observe that the median is larger than the mean, which means that although more than half of the debates return a high argument coverage, we have some debates that return very small-sized explanations, which drags down the mean values. For every explanation-heuristic combination, we report a standard deviation around $20-26$ percentage points which shows that the argument coverage is not consistent from a debate to another, and that some debates return very small argument coverages while others return very high ones.

In order to visualize the overall distribution of the debates along the different ranges of argument coverage, we present in Figure 5 the stacked histogram of bin counts $(0–10\% \cdots 90–100\%)$ for each explanation-heuristic combination.

Although all the explanation-heuristic combinations are dominated by the $90-100\%$ coverage bin, we can observe more balanced distribution across bins for constructive explanations than for destructive explanations, which are heavily concentrated in the $90-100\%$ coverage bin. For each heuristic, constructive explanations give much taller low-coverage bins $(0-50\%)$ and much shorter high-coverage bins $(90-100\%)$ when compared to destructive explanations. Among all the four explanation-heuristic combinations, constructive explanation combined with $H_{s\to l}$ has the tallest early bars, about 21.8% of debates in 0–50% bins, and it has, along with constructive explanation combined with $H_{s\to w}$, the shortest $90-100\%$ bar, about 28.6% of debates. While constructive explanations combined with $H_{s\to l}$ shows the highest efficiency (in terms of argument coverage by the explanation), with a mean coverage of 69.68% compared to 80.72% for the least efficient explanation-heuristic combination, we still have around 78.2% of debates which need high argument coverage $(50-100\%)$. In order to check whether argument coverage decreases with debate size, we analyze in Section 6.2, how argument coverages vary across debates of different sizes.

Note that we also study the explanation size w.r.t the number of sub-trees. This study shows that $H_{s\to w}$ gives shorter explanations w.r.t. the number of sub-trees than $H_{s\to l}$. This is expected since the goal of $H_{s\to w}$ is to reduce the number of returned sub-trees. To check the details and results of this study, we invite the reader to check the GitHub repository.

6.2 Impact of Debate Size

In this section, we study how argument coverages vary across debates of different sizes. We first categorized the 4326 debates based on their total number of arguments. We conducted an empirical distribution analysis of the debates, revealing a median of 23 arguments and a mean of 82. Since about half of the debates have less than 23 arguments, we set the first threshold at 23 arguments. To create balanced statistical analysis and meaningful size categories for comparison, we divided the remaining debates into four equal-sized categories of approximately 12.5% each, using percentile-based thresholds computed via linear interpolation at 43, 80, and 204 arguments. Hence we determined the following five size categories: Very Small (<23 arguments, 49.3% of debates, $n=2133$), Small ($23-42$ arguments, 13.8% of debates, $n=598$), Medium ($43-79$ arguments, 11.7% of debates, $n=505$), Large ($80-203$ arguments, 13.4% of debates, $n=580$), and Very Large ($204+$ arguments, 11.8% of debates, $n=510$). Then we computed the mean argument coverages (±standard deviation) for all explanation-heuristic combinations across these five size categories. Figure 6 presents these results.

We can observe that on average, the very small debates (<23 arguments) consistently require the highest argument coverage ($\approx 77-82\%$) across all the explanation-heuristic combinations, while larger debates ($23+$ arguments) require significantly lower coverage ($\approx 62-80\%$). In Section 6.1, we reported that even the most efficient combination, which is constructive explanation combined with $H_{s\to l}$, reported a high proportion of debates ($\approx 78.2\%$) re-

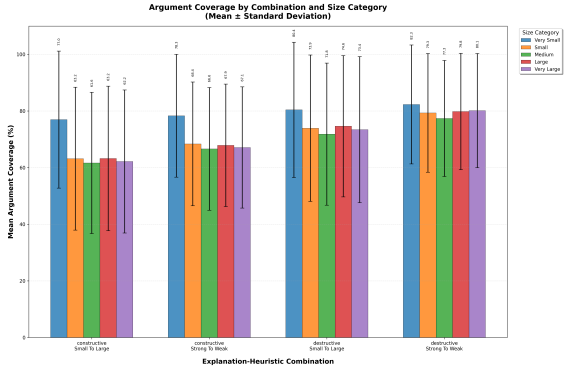

Figure 6: Mean argument coverage (±std dev) by explanation-heuristic combination and debate size category. Sample sizes: Very Small ($n = 2133$), Small ($n = 598$), Medium ($n = 505$), Large ($n = 580$), Very Large ($n = 510$).

quiring high argument coverage. This can be explained by the large proportion of very small debates in our dataset ($\approx 49.3\%$). Notably, this combination achieves argument coverage of approximately $62-63\%$ for small, medium, large, and very large categories, representing the minimum coverage threshold across all explanation-heuristic combinations and demonstrating the optimal performance in larger debates. These results show that the debate size affects the explanation size, and that the large proportion of very small debates (49.3%) contributes to the overall high argument coverage, though all debate sizes still need considerable argument coverage. In order to further reduce the argument coverages across all debate sizes, we will investigate in the future work a refined version of the constructive strategy. For the strengthening case, this version decomposes the con-top-level sub-trees in the same manner the current constructive strategy does. However, it also decomposes the pro-top-level sub-trees into two categories of sub-trees. While it considers the con-top-level sub-trees at their strongest state (unweakened), it takes the pro-top-level sub-trees at their weakest (unstrengthened) state. The same reasoning applies dually to the weakening case. We expect this refined strategy to reduce the number of returned arguments, as it returns for the strengthening (resp. weakening) case, minimal sub-trees from the pro-top-level (resp. con-top-level) sub-trees rather than the entire pro-top-level (resp. con-top-level) sub-trees, which are often larger.

6.3 Impact of Weight Change

In order to study the relationship between the debate issue's weight change and the argument coverage by an explanation $E \in [Cons\_Exp, Des\_Exp]$ following heuristic $H \in [H_{s\to w}, H_{s\to l}]$ across 4326 debates, we generated four scatter plots (one for each

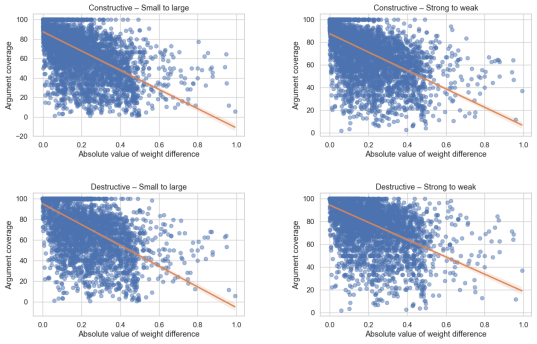

Figure 7: Scatter plots of $|\sigma_{\mathcal{F}}(t) - w(t)|$ vs. $P(E)$ for each explanation-heuristic combination across 4326 debates.

Table 2: Pearson's and Spearman's correlation coefficients by explanation-heuristic combination (4326 debates). All correlations are significant at p-value< 0.001.

Strategy	Heuristic	Pearson's r	Spearman's ρ
Constructive	$H_{s\to l}$	-0.59	-0.64
Constructive	$H_{s\to w}$	-0.56	-0.63
Destructive	$H_{s\to l}$	-0.62	-0.69
Destructive	$H_{s\to w}$	-0.56	-0.67

explanation-heuristic combination) in order to show the relationship between the absolute weight change of a debate issue t, $|\sigma_{\mathcal{F}}(t) - w(t)|$, and the argument coverage $P(E)$. Figure 7 shows the four scatter plots. We can observe in every scatterplot that the fitted regression line shows a negative slope, indicating that larger weight change consistently produces smaller explanations (in terms of argument coverage). Table 2 provides numerical confirmation of this visual pattern, where Pearson's r ranges from around –0.56 to around –0.62 and Spearman's ρ ranges from around –0.63 to around –0.69 across all four groups, with p-values effectively zero. These strong, highly significant negative correlations confirm an approximately linear decrease in explanation size as weight change increases, and this effect is a bit strongest for $H_{s\to l}$ yet remaining robust in all cases.

7 RELATED WORK

Several works in the literature propose explanations for argumentation semantics' outcomes. For extension-based semantics, Fan and Toni (Fan and Toni, 2015a; Fan and Toni, 2015b) provide explanations for the acceptance and the non-acceptance of arguments under admissible semantics in abstract and assumption-based argumentation frameworks. Borg and Bex (Borg and Bex, 2024) define basic explanations for the acceptance and non-acceptance of arguments under several extension-based semantics, in

both abstract and ASPIC$^+$ frameworks. They define three notions to select among these explanations, minimal, sufficient and necessary ones. Amgoud (Amgoud, 2024) defines factual, counterfactual, and contrastive explanations based on sufficient, necessary, and influential attacks that affect the acceptability status of an argument a, under a family of extension-based semantics. Ulbricht and Wallner (Ulbricht and Wallner, 2021) define a strong explanation for the acceptance of a set S of arguments under an extension-based semantics as a set of arguments such that S is acceptable in every subframework that contains the explaining set.

As for gradual semantics, several impact measures have been recently defined in the literature (Al Anaissy et al., 2025; Kampik et al., 2024b) to explain the outcomes of gradual semantics. Inspired by the feature attribution explainability notion in XAI, an impact measure computes the contribution of an individual argument or a set of arguments to an argument's final weight. Apart from impact measures, Kampik et al. (Kampik et al., 2024a) define sufficient, necessary and counterfactual explanations as sets of arguments that account for changes in the partial ordering of arguments' final weights in a $WBAF$. These changes result from adding or removing some arguments or edges, or from modifying the arguments' initial weights. Yin et al. (Yin et al., 2024; Yin et al., 2025) focus on counterfactual explanations for why an argument's final weight is different than a desired value. They study how the initial weights of arguments in a $WBAF$ can be changed in order to change a specific argument's final weight under a bipolar gradual semantics (Yin et al., 2024). They define and implement an algorithm that iteratively updates the initial weights of all the arguments in the $WBAF$ until the target argument reaches the desired final weight. In another work (Yin et al., 2025), they study how changes to edge weights in edge-weighted $WBAF$s can similarly achieve a desired final weight. Morveli-Espinoza et al. (Morveli-Espinoza and Nieves, 2024) define two contrastive explanations for the strength ranking over arguments produced by gradual semantics. These explanations address the questions: "why is an argument a ranked in position x rather than position y?" and "why is an argument a ranked in position x while argument b is ranked in position y?".

When it comes to explaining why an argument a's final weight is higher (resp. lower) than its initial weight, Čyras et al. (Čyras et al., 2022) define the notion of a sufficient Quantitative Dispute Tree (QDT) for a as a sub-graph of an acyclic $WBAF$, that is sufficient to ensure that a's final weight does not decrease (resp. does not increase). They define such trees in terms of pro/con-arguments, determined according to four contribution functions (impact measures) that they define. However, it has been shown in (Kampik et al., 2024b) that among these four impact measures, only the "removal-based" one satisfies the "counterfactuality" principle, which we consider an axiom for intuitive behavior of an impact measure used for explainability. In our work, we define pro/con-arguments w.r.t. the effect of their removal from the graph on the debate issue's final weight. This is equivalent to the removal-based impact measure defined in (Čyras et al., 2022). While Čyras et al. (Čyras et al., 2022) suggest that a sufficient QDT for a can be progressively constructed by inspecting how adding different arguments and relations, starting with a, affects a's final weight, our methodology introduces novel formal strategies and heuristics for constructing abductive (sufficient) sub-tree-based explanations, that address the same problem, explaining weight change, through a systematic approach.

8 CONCLUSION AND FUTURE WORK

In this paper, we addressed the explainability of online debates' evolution of outcomes computed by bipolar gradual semantics. We introduced two strategies and two heuristics to generate sub-tree-based abductive explanations. We applied our methodology on Kialo debate graphs. Although all the explanation-heuristic combinations produce on average high argument coverages, the constructive strategy produces smaller values than the destructive strategy, for each heuristic. Also, $H_{s\to l}$ produces the smallest argument coverages within each strategy. We studied how argument coverage varies across different debate size categories. Results show that the constructive strategy consistently outperforms the destructive strategy, and that $H_{s\to l}$ consistently outperforms $H_{s\to w}$, across all debate size categories. This study also shows that the high proportion of very small debates in the dataset contributes to the overall high argument coverage, yet considerable argument coverages are still required across all debate sizes. Finally, we studied the effect the debate issue's weight change on the argument coverage. Results show that arguments with large weight changes can be explained with fewer arguments.

For future work, we will investigate a refined version of the constructive strategy, introduced in Section 6.2, which we expect to reduce the argument coverage across all debate sizes and for all the heuristics. The next step would be to conduct user studies to evaluate users' satisfaction with our proposed ex-

planations. We will also explore a different type of abductive explanations for bipolar gradual semantics' outcomes that answers the question: "why is the final weight of an argument a below a specific threshold?". Finally, an interesting future work direction would be to generate personalized explanations by taking into account the user's preferences. The main idea would be to model the user's preferences over topics that we infer through topic modeling. Each argument being associated with a specific topic, the generated explanation will prioritize the arguments that are aligned with the user's most preferred topics.

ACKNOWLEDGMENTS

This work has been supported by the ANR-22-EXEN-0005 (PEPR eNSEMBLE / PC4 CONGRATS) and the ANR-23-IACL-0007 (AI Cluster PostGenAI) projects.

REFERENCES

Al Anaissy, C., Delobelle, J., Vesic, S., and Yun, B. (2025). Impact measures for gradual argumentation semantics. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, pages 69–77.

Amgoud, L. (2024). Post-hoc explanation of extension semantics. In *27TH EUROPEAN CONFERENCE ON ARTIFICIAL INTELLIGENCE*.

Amgoud, L. and Ben-Naim, J. (2018). Evaluation of arguments in weighted bipolar graphs. *International Journal of Approximate Reasoning*, 99:39–55.

Baroni, P., Romano, M., Toni, F., Aurisicchio, M., and Bertanza, G. (2015). Automatic evaluation of design alternatives with quantitative argumentation. *Argument & Computation*, 6(1):24–49.

Borg, A. and Bex, F. (2024). Minimality, necessity and sufficiency for argumentation and explanation. *International Journal of Approximate Reasoning*, 168:109143.

Cayrol, C. and Lagasquie-Schiex, M.-C. (2005). Gradual valuation for bipolar argumentation frameworks. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pages 366–377. Springer.

Čyras, K., Kampik, T., and Weng, Q. (2022). Dispute trees as explanations in quantitative (bipolar) argumentation. In *ArgXAI 2022, 1st International Workshop on Argumentation for eXplainable AI, Cardiff, Wales, September 12, 2022*, volume 3209. CEUR-WS.

Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77(2):321–357.

Fan, X. and Toni, F. (2015a). On computing explanations in argumentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Fan, X. and Toni, F. (2015b). On explanations for non-acceptable arguments. In *Theory and Applications of Formal Argumentation: Third International Workshop, TAFA 2015, Buenos Aires, Argentina, July 25-26, 2015, Revised Selected Papers 3*, pages 112–127. Springer.

Grimmelikhuijsen, S. (2023). Explaining why the computer says no: Algorithmic transparency affects the perceived trustworthiness of automated decision-making. *Public Administration Review*, 83(2):241–262.

Kampik, T., Čyras, K., and Alarcón, J. R. (2024a). Change in quantitative bipolar argumentation: sufficient, necessary, and counterfactual explanations. *International Journal of Approximate Reasoning*, 164:109066.

Kampik, T., Potyka, N., Yin, X., Cyras, K., and Toni, F. (2024b). Contribution functions for quantitative bipolar argumentation graphs: A principle-based analysis. *CoRR*, abs/2401.08879.

Morveli-Espinoza, M. and Nieves, J. C. (2024). Generating contrastive explanations from gradual semantics rankings. In *Ibero-American Conference on Artificial Intelligence*, pages 250–261. Springer.

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., and Seifert, C. (2023). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s):1–42.

Potyka, N. (2018). Continuous dynamical systems for weighted bipolar argumentation. In *KR*, pages 148–157.

Rago, A., Toni, F., Aurisicchio, M., Baroni, P., et al. (2016). Discontinuity-free decision support with quantitative argumentation debates. *KR*, 16:63–73.

Shortall, R., Itten, A., Meer, M. v. d., Murukannaiah, P., and Jonker, C. (2022). Reason against the machine? future directions for mass online deliberation. *Frontiers in Political Science*, 4:946589.

Ulbricht, M. and Wallner, J. P. (2021). Strong explanations in abstract argumentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6496–6504.

Yin, X., Potyka, N., Rago, A., Kampik, T., and Toni, F. (2025). Contestability in quantitative argumentation. *arXiv preprint arXiv:2507.11323*.

Yin, X., Potyka, N., and Toni, F. (2024). Ce-qarg: counterfactual explanations for quantitative bipolar argumentation frameworks. In *Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning*, pages 697–707.

Young, A. P., Joglekar, S., Boschi, G., and Sastry, N. (2021). Ranking comment sorting policies in online debates. *Argument & Computation*, 12(2):265–285.