Predicting Human Judgement in Online Debates with Argumentation

Beauty Oluokun, Guilherme Paulino-Passos*, Antonio Rago* and Francesca Toni

Department of Computing, Imperial College London, UK

Abstract

Debates and disagreements are fundamental to human progression as they challenge assumptions and foster personal growth. We have witnessed public discourse migrating online rapidly, with platforms like Reddit's `r/AmITheAsshole` subreddit offering a space for users to seek judgements on their actions and engage in moral debates. However, automated methods for modelling such debates and making predictions regarding human judgement therein are lacking in the literature. In this paper, we investigate how to model and predict within these online debates using computational argumentation, a set of formalisms known to excel in representing and reasoning with knowledge in a human-like manner. Concretely, we introduce a pipeline for modelling and predicting human judgement within Reddit threads using argument mining and quantitative bipolar argumentation frameworks under gradual semantics to imitate an outside observer or arbitrator. We demonstrate that our approach achieves a reasonable degree of accuracy in this domain and, interestingly, that our model's behaviour when different gradual semantics are applied correlates fairly well with their theoretical properties.

Keywords

bipolar argumentation, gradual semantics, online debate, human judgement

1. Introduction

Debates and disagreements are common in everyday life, from the seemingly insignificant, e.g. arguments between young siblings over toys, to those which concern billions of dollars, e.g. corporate legal battles. Though they can often be unpleasant at that moment, disputes, i.e. processes of argumentation, are pivotal to human progression as they underpin all human reasoning [1], enabling us to present new ideas to each other and foster personal growth. When people with differing perspectives converse, they can challenge each other's assumptions which can lead to better informed decisions, especially in the political and legal realms but also on a smaller-scale, e.g. domestic issues such as disputes between family and friends.

Nowadays, public discourse is commonly held online, particularly on social media, where participants are almost invisible to one another. This means that, when taking part in such discourse, one cannot always be sure how their contribution to a debate affected other participants or the outcome of the dispute. Thus, in order to predict such effects, it would be beneficial to have models for human debate that easily assist in investigating the effect of the debate

The 24th International Workshop on Computational Models of Natural Argument (CMNA'24), September 17th, 2024, Hagen, Germany

*Corresponding authors.

beauty.oluokun20@imperial.ac.uk (B. Oluokun); gppassos@imperial.ac.uk (G. Paulino-Passos); a.rago@imperial.ac.uk (A. Rago); ft@imperial.ac.uk (F. Toni)

structures and individual arguments in the conclusion of disputes. One such set of formalisms for modelling debates is computational argumentation (see [2, 3] for overviews). For example, argumentative techniques have previously been used to gather insights on online debates in [4], focusing on what views were being expressed and why.

To understand why argumentation has been so effective in this task, we need to consider its core components, as well as those of the application domain. Online debates normally consist of chains of comments and replies which naturally create a tree-like structure. Argumentation Frameworks (AFs) enable us to model a set of arguments that attack or support one another, most commonly as graphs and often as trees, and have been beneficial in areas of explainable AI, e.g. in modelling disagreements between an AI model's output and human observer by evaluating the dialectical acceptability or strength of the arguments [5]. Baroni et. al. in [6] outline the various extensions of a basic AF and unify them into a more generalised framework: Quantitative Bipolar Argumentation Frameworks (QBAFs). Following this, Cocarascu et al. built on Baroni's work to deploy QBAFs in an Argumentative Dialogical Agent (ADA) model to improve review aggregation explanations in sites such as Metacritic and Rotten Tomatoes [7]. ADA utilised sentiment analysis, argument mining, and gradual evaluation for QBAFs, showing how QBAFs can be successful in modelling a network of reviews that interact with one another.

In this work, we propose an argumentative approach to modelling an outside observer (an arbitrator, a judge, or simply the audience) of an online debate in order to create an automatic model of human judgement. We leverage QBAFs to model human debates and provide tools to further investigate the effect of the debate structure and individual arguments in the final decision or conclusion of the dispute. We deploy our approach in the particularly intriguing corner of the digital space that is the Reddit `r/AmITheAsshole` (AITA) forum. This online community provides a space for users to explain a situation in their lives where they may have acted like an "asshole" and receive a judgement from other users on their morality. These online debates involve thousands of participants, from across the globe, offering their opinions in response to the original poster (OP) or other comments. We model AITA debates as QBAFs, showing the framework closely mimics the flow of a thread of comments under a Reddit post that would be read by a user of the app. As our argumentative approach is able to predict the verdict of an AITA thread with a high success rate, we tentatively posit that it could be a plausible model of human reasoning during the debate, which is worthy of extensive future investigations.

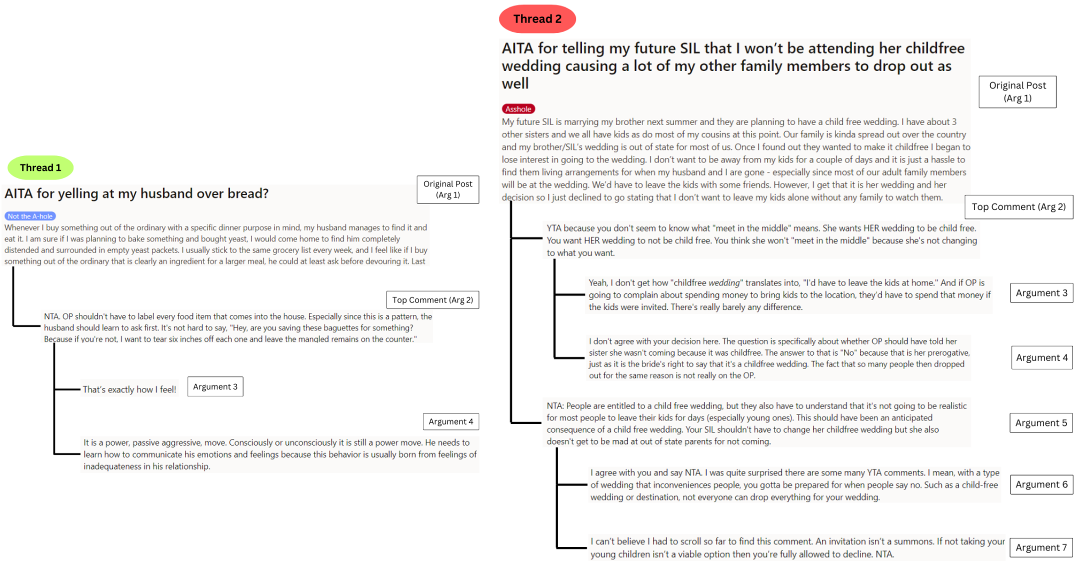
Our contributions are as follows:

- We present a comprehensive pipeline for modelling Reddit's AITA threads using QBAFs to predict their verdicts. This pipeline facilitates the mining of arguments and their relations and the analysis of argumentative structures from online discussions.
- We evaluate our pipeline's prediction performance wrt accuracy and discuss the advantages and limitations of employing QBAF and different gradual evaluation methods in existing literature for modelling online AITA debates.
- We provide a public dataset of 823 example threads from the AITA subreddit that have the verdicts labelled[1].

[1] Dataset and code available at github.com/BOluokun/reddit-argumentation.

2. Background and Related Work

AITA In this online forum, the OP's situation can be judged as one of four tags: You are The Asshole (YTA), indicating that the OP is the only one in the wrong; Not The Asshole (NTA), meaning the OP has done no wrong and the other party in the conflict is an asshole; No Assholes Here (NAH), stating that no one in the situation can be rightly labelled an asshole; and Everyone Sucks Here (ESH), expressing that both (or all) sides in the situation have acted as assholes. They capture the four possible combinations between the OP being in the wrong or not, and the other party (or parties) of the conflict being in the wrong or not. In order to simplify the problem, we focus only on whether the OP is in the wrong or not. That is, we map both YTA and ESH into YTA, and both NTA and NAH into NTA. The overall verdict comes from the judgement indicated in the top-level comment with the highest upvote score. Nonetheless, there may be disagreements between comments supporting the NTA and NAH verdicts, likewise between YTA and ESH verdicts, which we leave to future work. This is similar to the approach in [8], where Efstathiadis et. al. utilised a BERT model [9] that achieved an accuracy of 62% when classifying posts and an accuracy of 86% when classifying comments. In that work, issues may have been caused by isolating the comments from the original posts. Thus, our approach of modelling the interactions between comments and the post, through AFs, could yield better results.



AITA for yelling at my husband over bread?

Whenever I buy something out of the ordinary with a specific dinner purpose in mind, my husband manages to find it and eat it. I am sure if I was planning to bake something and bought yeast, I would come home to find him completely distended and surrounded in empty yeast packets. I usually stick to the same grocery list every week, and I feel like if I buy something out of the ordinary that is clearly an ingredient for a larger meal, he could at least ask before devouring it. Last

NTA. OP shouldn't have to label every food item that comes into the house. Especially since this is a pattern, the husband should learn to ask first. It's not hard to say, "Hey, are you saving these baguettes for something? Because if you're not, I want to tear six inches off each one and leave the mangled remains on the counter."

That's exactly how I feel!

It is a power, passive aggressive, move. Consciously or unconsciously it is still a power move. He needs to learn how to communicate his emotions and feelings because this behavior is usually born from feelings of inadequateness in his relationship.

(a) AITA thread 1 with the NTA verdict.

AITA for telling my future SIL that I won't be attending her childfree wedding causing a lot of my other family members to drop out as well

My future SIL is marrying my brother next summer and they are planning to have a child free wedding. I have about 3 other sisters and we all have kids as do most of my cousins at this point. Our family is kinda spread out over the country and my brother/SIL's wedding is out of state for most of us. Once I found out they wanted to make it childfree I began to lose interest in going to the wedding. I don't want to be away from my kids for a couple of days and it is just a hassle to find them living arrangements for when my husband and I are gone - especially since most of our adult family members will be at the wedding. We'd have to leave the kids with some friends. However, I get that it is her wedding and her decision so I just declined to go stating that I don't want to leave my kids alone without any family to watch them.

YTA because you don't seem to know what "meet in the middle" means. She wants HER wedding to be child free. You want HER wedding to not be child free. You think she won't "meet in the middle" because she's not changing to what you want.

Yeah, I don't get how "childfree *wedding*" translates into, "I'd have to leave the kids at home." And if OP is going to complain about spending money to bring kids to the location, they'd have to spend that money if the kids were invited. There's really barely any difference.

I don't agree with your decision here. The question is specifically about whether OP should have told her sister she wasn't coming because it was childfree. The answer to that is "No" because that is her prerogative, just as it is the bride's right to say that it's a childfree wedding. The fact that so many people then dropped out for the same reason is not really on the OP.

NTA: People are entitled to a child free wedding, but they also have to understand that it's not going to be realistic for most people to leave their kids for days (especially young ones). This should have been an anticipated consequence of a child free wedding. Your SIL shouldn't have to change her childfree wedding but she also doesn't get to be mad at out of state parents for not coming.

I agree with you and say NTA. I was quite surprised there are some many YTA comments. I mean, with a type of wedding that inconveniences people, you gotta be prepared for when people say no. Such as a child-free wedding or destination, not everyone can drop everything for your wedding.

I can't believe I had to scroll so far to find this comment. An invitation isn't a summons. If not taking your young children isn't a viable option then you're fully allowed to decline. NTA.

(b) AITA Reddit thread 2 with the YTA verdict.

Figure 1: Examples of online debate in the AITA threads in the dataset.

Quantitative Bipolar Argumentation Frameworks QBAFs [10] are quadruples $\langle \mathcal{X}, \mathcal{A}, \mathcal{S}, \tau \rangle$ where $\mathcal{X}$ is a finite set (of *arguments*), $\mathcal{A} \subseteq \mathcal{X} \times \mathcal{X}$ and $\mathcal{S} \subseteq \mathcal{X} \times \mathcal{X}$ are the *attack* and *support* relations, respectively, and $\tau : \mathcal{X} \rightarrow \mathbb{I}$ assigns *base scores* to arguments, representing the arguments' intrinsic strengths from within a given evaluation range $\mathbb{I}$ (we use $\mathbb{I} = [0,1]$ throughout). In our application, the arguments represent the comments in the AITA Reddit thread. For any $a \in \mathcal{X}$, the *attackers* of a are $\mathcal{A}(a) = \{b \in \mathcal{X} \mid (b,a) \in \mathcal{A}\}$ and the *supporters* of a are $\mathcal{S}(a) = \{b \in \mathcal{X} \mid (b,a) \in \mathcal{S}\}$. QBAFs can be visualised as graphs, with arguments as nodes labelled by their base scores and relations as edges labelled by + (for support) or - (for attack).

Then, we say that $\mathcal{F} = \langle \mathcal{X}, \mathcal{A}, \mathcal{S}, \tau \rangle$ is a *QBAF for* $e \in \mathcal{X}$ iff $\nexists(e,a) \in \mathcal{A} \cup \mathcal{S}$ for any $a \in \mathcal{X}$, for all $a \in \mathcal{X} \setminus \{e\}$ there is a path in $\mathcal{F}$ from a to e, and $\nexists a \in \mathcal{X}$ with a path from a to a. Argument e is called the *explanandum*. In our application, e is the specific statement in which the initial stances of the participants are polar, and always refer to the statement "OP is NTA". In a QBAF for e, all other arguments are 'related to' e and there are no circular paths between arguments. We can visualise an agent's QBAF as a tree rooted at e.

Gradual semantics QBAFs can be paired with a *gradual semantics* σ [6] which credits arguments with a dialectical strength within $\mathbb{I}$. Dialectical strengths encapsulate a participant's views on the quality of or belief in arguments within a QBAF based on a combination of the base scores and perceived strengths of the arguments' attackers and supporters. We focus on four gradual semantics: the *Quantitative Argumentation Debate (QuAD)* [11] and *Discontinuity-Free Quantitative Argumentation Debate (DF-QuAD)* [12] algorithms, *Quadratic Energy Model (QEM)* [13] and *Exponent-based restricted semantics (Ebs)* [14]. We omit most of the formal definitions for lack of space, but, for illustration, we give the definition of Ebs. This formula shows how this semantics calculates the dialectical strength of an argument, a, using its base score, $\tau(a)$, and an aggregation of the dialectical strengths of its attackers and supporters - in this case given by $E(a)$.

Definition 1 (Ebs Gradual Semantics). *For $\mathcal{F} = \langle \mathcal{X}, \mathcal{A}, \mathcal{S}, \tau \rangle$ and any $a \in \mathcal{X}$,*

$$\sigma(\mathcal{F}, a) = 1 - \frac{1 - \tau(a)^2}{1 + \tau(a) \cdot 2^{E(a)}} \quad \textit{where} \quad E(a) = \sum_{x \in \mathcal{S}(a)} \sigma(\mathcal{F}, x) - \sum_{y \in \mathcal{A}(a)} \sigma(\mathcal{F}, y).$$

Different gradual semantics may satisfy different properties, characterising their functionality, including *attainability (A)* [7], informally enforcing that all strength scores in $\mathbb{I}$ are possible for an argument, given any base score and a suitable choice of attackers and supporters; *(strict) bivariate monotony ((S)M)* [14], informally describing that attacks cannot benefit their targets and equally, supports cannot harm their targets; and *(strict) Franklin ((S)F)* [14], informally that a supporter is never more important than an attacker of equal strength. We decided to focus on the four chosen semantics, as they provided a good variation in our selected properties, all used the same evaluation range, and they allowed for variable base score functions. More gradual semantics could be tested in future work, e.g. the *Restricted Euler-based semantics* [15].

Relation-based Argument Mining To obtain QBAFs from `r/AmITheAsshole` threads, we use relation-based argument mining, amounting to classifying pairs of texts (a child and a parent) as 'Attack', 'Support' or 'No' (depending on whether the child disagrees, agrees, or is unrelated to the parent, respectively). Specifically, in our experiments, we use Large Language Models (LLMs) [16] to perform relation-based argument mining using the few-shot prompting method outlined by Gorur et al. [17].

3. Online Debates as Argumentation Frameworks

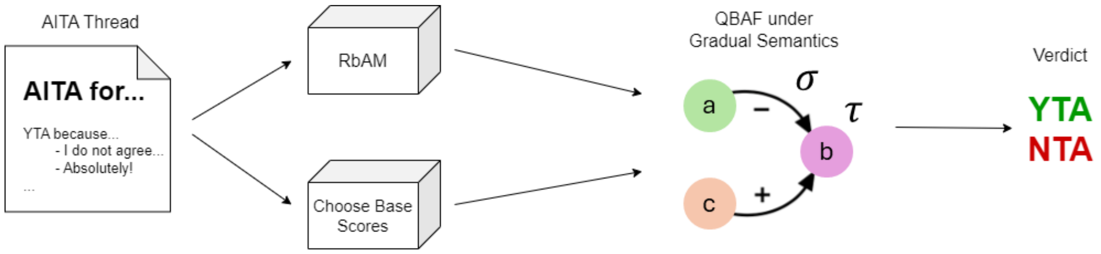


Figure 2: The proposed pipeline. The initial thread, with the original post and comments are passed through the LLM for relation-based argument mining (RbAM), which returns the graph structure of the QBAF, that is, the arguments, and the attack and support relations. The thread is also sent through a base scorer, which defines the base score of each argument. The graph structure with the base scores together define the QBAF. Then the evaluation of the QBAF with some gradual semantic defines a score for the explanandum, which is used to give the final classification.

In this section we detail our method for modelling an "arbitrator" for the AITA debates using QBAFs under gradual semantics, as outlined in Figure 2.

First, we obtain a QBAF for "OP is NTA" (the explanandum) as follows. The explanadum has a single supporter representing the original post. The nested comments under the post are then added to the QBAF recursively as attackers or supporters of their parent (post or comment) depending on the classification by the relation-based argument mining component. We discard a comment and its replies when it has a 'No' relation to its parent. This algorithm for converting AITA threads into QBAFs is guaranteed to produce well-formed QBAFs for the explanandum "OP is NTA", as each edge between arguments is exclusively either an attack or a support, all comments are connected to the explanandum through a reply chain and no argument attacks or supports itself. These QBAFs are also guaranteed to be acyclic (in that there is no path from any argument to itself).

In the QBAF we need to assign appropriate base scores (τ) for all arguments. We propose two elementary methods for this. A basic method is to use a 'fixed' τ which assigns each argument (including the explanandum) the same value:

$$\tau_{\text{fixed}}(a) = \alpha \tag{1}$$

for a chosen value $\alpha \in \mathbb{I}$. This can be interpreted as every comment in the thread being equally influential and trustworthy. This is quite a naive approach but a useful baseline to evaluate the importance of the structure of the debate - chains of supports and attacks - in determining the verdict rather than features of the comments themselves. Figure 3a shows an example when using τ_{fixed} for an AITA thread with DF-QuAD gradual semantics (σ_1) and different values of α: $\alpha_1 = 0.1$, $\alpha_2 = 0.25$ and $\alpha_3 = 0.4$.

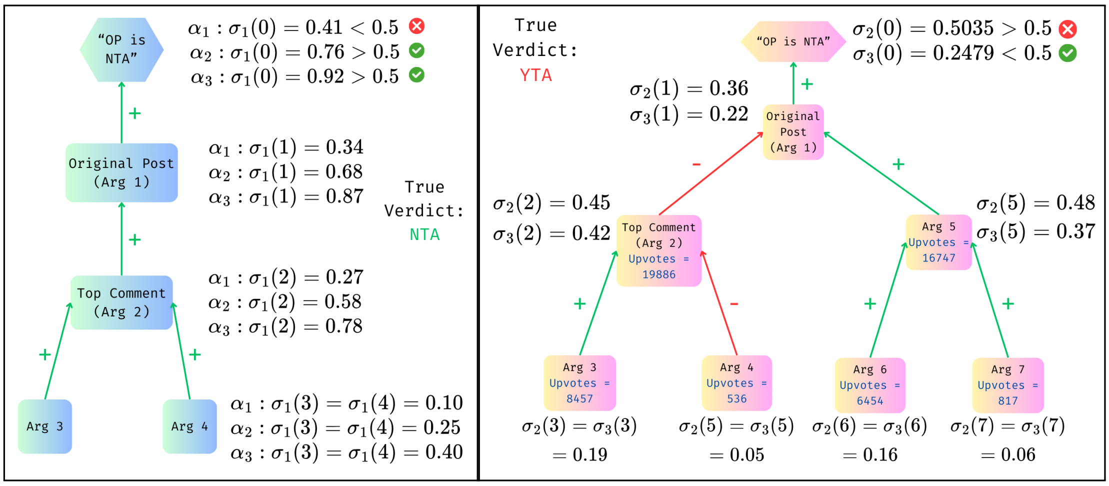
The second method is setting τ to vary with the number of upvotes a comment has. A comment having a greater upvote score indicates more participants who are reading the thread agree with the content of that comment. This suggests that a comment with a high upvote score should have a high base score, and a comment with a low upvote score should have a low base score. The range of upvotes varies between AITA threads and is unbounded so it is not appropriate to select an absolute scale with specific 'high' and 'low' values. Instead, for a thread t, We find the maximum and minimum upvotes (v_{max} and v_{min} respectively) and define a base score function such that for an arbitrary comment a with $v_{\text{max}} \neq v_{\text{min}}$:

$$\tau_{\text{upvote}}(a) = \alpha + \beta \times \frac{v_a - v_{\text{min}}}{v_{\text{max}} - v_{\text{min}}} \tag{2}$$

where α and β are parameters to be chosen and v_a is the upvote score of comment a. Using eq. (2) to determine base scores gives the comment with the lowest number of upvotes a base score of α and the comment with the highest number of upvotes a base score of $\alpha + \beta$. For any two comments a, b, if $\text{score}(a) < \text{score}(b)$, then $\tau_{\text{upvote}}(a) < \tau_{\text{upvote}}(b)$ and if $\text{score}(a) = \text{score}(b)$, then $\tau_{\text{upvote}}(a) = \tau_{\text{upvote}}(b)$. The explanandum and the original post do not have upvote scores so we set, for a representing one of these arguments:

$$\tau_{\text{upvote}}(a) = \alpha + \frac{\beta}{2} \tag{3}$$

Equation (3) is also used if $v_{\text{max}} = v_{\text{min}}$. Figure 3b shows an example of evaluating the stance of the explanandum in a QBAF with QuAD (σ_2) and Ebs (σ_3) with τ_{upvote} where $\alpha = 0.05$ and $\beta = 0.35$. For all four gradual semantics, if an argument has no supporters or attackers its strength is equal to its base score (τ), we see argument 7 has fewer upvotes than argument 6 and thus is assigned a lower base score with τ_{upvote}. Note that, if $\alpha, \beta \in \mathbb{I}$ and $\alpha + \beta \in \mathbb{I}$ then τ_{fixed} (defined in eq. (1)) and τ_{upvote} (defined in eq. (2) and *eq.* (3)) are well-defined.


"OP is NTA"
$\alpha_1 : \sigma_1(0) = 0.41 < 0.5$ ☒
$\alpha_2 : \sigma_1(0) = 0.76 > 0.5$ ✓
$\alpha_3 : \sigma_1(0) = 0.92 > 0.5$ ✓
+
Original Post (Arg 1)
$\alpha_1 : \sigma_1(1) = 0.34$
$\alpha_2 : \sigma_1(1) = 0.68$
$\alpha_3 : \sigma_1(1) = 0.87$
True Verdict: NTA
+
Top Comment (Arg 2)
$\alpha_1 : \sigma_1(2) = 0.27$
$\alpha_2 : \sigma_1(2) = 0.58$
$\alpha_3 : \sigma_1(2) = 0.78$
+
+
Arg 3
Arg 4
$\alpha_1 : \sigma_1(3) = \sigma_1(4) = 0.10$
$\alpha_2 : \sigma_1(3) = \sigma_1(4) = 0.25$
$\alpha_3 : \sigma_1(3) = \sigma_1(4) = 0.40$
True Verdict: YTA
"OP is NTA"
$\sigma_2(0) = 0.5035 > 0.5$ ☒
$\sigma_3(0) = 0.2479 < 0.5$ ✓
+
$\sigma_2(1) = 0.36$
$\sigma_3(1) = 0.22$
Original Post (Arg 1)
-
+
$\sigma_2(2) = 0.45$
$\sigma_3(2) = 0.42$
Top Comment (Arg 2) Upvotes = 19886
Arg 5 Upvotes = 16747
$\sigma_2(5) = 0.48$
$\sigma_3(5) = 0.37$
+
-
+
+
Arg 3 Upvotes = 8457
Arg 4 Upvotes = 536
Arg 6 Upvotes = 6454
Arg 7 Upvotes = 817
$\sigma_2(3) = \sigma_3(3) = 0.19$
$\sigma_2(5) = \sigma_3(5) = 0.05$
$\sigma_2(6) = \sigma_3(6) = 0.16$
$\sigma_2(7) = \sigma_3(7) = 0.06$

(a) Thread 1 arguments' strengths with τ_{fixed}

(b) Thread 2 arguments' strengths with τ_{upvote}

Figure 3: Example QBAFs, obtained from the AITA threads in Fig. 1, and argument evaluation with different base scores and gradual semantics.

Once the relations between comments and replies in a thread are assigned and the base scores decided, the *QBAF classifier*, i.e. the AITA "arbitrator", or audience, can be built specific to that thread. The QBAF classifier predicts the verdict of the AITA thread using a chosen gradual semantics to evaluate the explanandum, giving a score in the evaluation range. If the score is strictly above the neutral value in $\mathbb{I}$ (0.5), we predict the stance NTA (the positive class) as the QBAF has calculated a strong belief in the argument "OP is NTA". Contrastingly, if the score is below or equal to the neutral value we predict the stance YTA (negative class) because the QBAF has calculated a lack of belief in the explanandum, suggesting the OP has likely been an asshole. Normally, an argument can have a neutral stance (strength) but we have restricted our use of strength scores so the QBAFs act as binary classifiers.

Implementation details We build QBAFs by applying a depth-first algorithm on a JSON representation of AITA threads which produces QBAF objects encapsulating NetworkX [18] directed graphs (representing the arguments in QBAFs as nodes, with the edges having either an 'Attack' or 'Support' label). Each QBAF object has `tau`, `semantics` and `eval_range` attributes which are functions corresponding to the base score function (τ), gradual semantics (σ) and evaluation range ($\mathbb{I}$), respectively.

For the relation-based argument mining, we used the method of [17] (off-the-shelf) with a 4-bit quantisation of the Mistral-7B-Instruct-v0.2, due to hardware limitations. However, memory and storage limitations did not reduce the effectiveness of the LLM as [17] reported similar accuracy and macro F_1 as Llama70B-4bit. Moreover, the message sent to the LLM consists of a 7-shot prompt primer followed by the parent comment (`Arg1`), the child comment (`Arg2`) and the line "Relation:" in an identical form to the examples.

4. Evaluation

Experiment Set-Up Numerous databanks of posts from `r/AmITheAsshole` exist publicly[2]. However, many of these separate the original post content from the comments underneath and thus lose the structure of the discussion about the post that is integral to our research. Therefore, we decided to use Reddit's API, through PRAW [3], to scrape snippets of AITA threads. Our dataset is stored in a TSV file and has a total of 823 entries. Reddit's public API was employed to scrape threads from the AITA subreddit and from an initial sample of 1000 threads; we only retained those which were correctly labelled with a verdict leaving an arbitrary number of 823 entries. For each, we recorded the title, the verdict, the filename of the JSON containing the thread contents, and the total number of comments (including the original post). Overall, the dataset has 390 NTA entries and 433 YTA entries, which is not a significant class imbalance.

Of the 823 example threads in the data set, 165 examples were set aside for testing. 5-fold cross-validation, performed on the remaining 658 examples, was used to determine the optimal values of α and β for eq. (1) and eq. (2) for each gradual semantics. All eight combinations of gradual semantics (σ) and base score functions (τ) for QBAF classifiers were tuned and tested.

To tune α for the QBAF classifiers using τ_{fixed}, we tested each classifier with 100 values for α ranging from 0.01 to 0.5. This range was chosen because preliminary testing suggested that the performance of $\alpha > 0.5$ would be too low. The optimal α chosen for a classifier was the one which produced the highest mean F_1 score across the 5 folds. Likewise, to tune α and β

[2] E.g. dvc.ai/blog/a-public-reddit-dataset.

[3] praw.readthedocs.io/en/latest/

for the τ_{upvote} classifiers, we tested 1600 combinations of (α, β) with α ranging from 0.01 to 0.2 and β ranging from 0.01 to 0.8. These ranges were chosen to explore performance over the full evaluation range $\mathbb{I} = [0, 1]$ as $0.2 + 0.8 = 1$ is the maximum possible base score and $0.01 + 0.01 = 0.02$ is very close to the minimum.

We did not utilise an LLM as a baseline in our experiments as we were aiming for explainability: an LLM could be tuned to perform well on our dataset, however its black-box nature would likely make it difficult to interpret how its prediction was formulated. Exploring the performance of a purely LLM-based system is left for future work.

Experiment Results The optimal parameters for α and β are as in Table 1. The performance of the eight QBAF classifiers with respect to their optimal values for α and β are shown in Table 2.

Note that DF-QuAD, QuAD and Ebs have an optimal α below 0.25 whereas QEM prefers an α above 0.4. With τ_{upvote}, all four gradual semantics prefer a low value for α (below 0.05) paired with a higher value for β. However, like with τ_{fixed}, QEM finds higher base scores more optimal with $\beta > 0.7$ while the other three semantics prefer smaller ranges. Figure 3a highlights how the choice of α can greatly influence the prediction of the QBAF arbitrator. All arguments support their parent, which should result in a prediction that agrees with the true verdict of NTA, however, DF-QuAD with $\alpha_1 = 0.1$ is not capable of building strength in the explanandum which is greater than 0.5. Furthermore, $\alpha_3 = 0.4$ provides the correct prediction but with an extremely high strength, placing too much trust in individual arguments. Thus the aggregation of attackers and supporters may cause too extreme changes from the base score of an argument.

	Df-QuAD	QEM	QuAD	Ebs
τ_{fixed}	$\alpha = 0.2377$	$\alpha = 0.4159$	$\alpha = 0.1634$	$\alpha = 0.1882$
τ_{upvote}	$\alpha = 0.01,\ \beta = 0.4962$	$\alpha = 0.0490,\ \beta = 0.7190$	$\alpha = 0.01,\ \beta = 0.3746$	$\alpha = 0.01,\ \beta = 0.3544$

Table 1
Optimal Parameters for each QBAF classifier (gradual semantics and base score functions).

	DF-QuAD, τ_{fixed}	QEM, τ_{fixed}	QuAD, τ_{fixed}	Ebs, τ_{fixed}	DF-QuAD, τ_{upvote}	QEM, τ_{upvote}	QuAD, τ_{upvote}	Ebs, τ_{upvote}
F_1	0.8187	0.8065	0.7487	0.7374	0.8249	0.8156	0.8047	**0.8261**
Precision	**0.7609**	0.7009	0.6481	0.6134	0.7449	0.7300	0.7556	0.7238
Recall	0.8861	0.9494	0.8861	0.9241	0.9241	0.9241	0.8608	**0.9620**
ROC-AUC	0.8151	0.7886	0.7221	0.6946	**0.8167**	0.8050	0.8025	0.8134

Table 2
Performances of the QBAF classifiers (wrt their optimal values of α and β) on the held-out test set.

Overall, Table 2 highlights that QBAF classifiers can be successful at determining the verdicts of AITA threads, with the lowest F_1 score being 0.7374 for Ebs with τ_{fixed} and the best performing QBAF classifier using Ebs and τ_{upvote} with an F_1 score of 0.8261, closely followed by DF-QuAD with the highest ROC-AUC score. The variation in performance across the gradual semantics seems to be correlated with the properties each satisfy (see Table 3)[4], as discussed below.

[4] The proofs for the semantics' satisfaction of these properties are shown in [6, 13, 14] except for Attainability (A) for QEM, which is satisfied but we omit the proof for lack of space.

	A	M	SM	F	SF
DF-QuAD	✓	✓	✗	✓	✓
QEM	✓	✓	✓	✓	✓
QuAD	✓	✓	✗	✗	✗
Ebs	✗	✓	✓	✓	✓

Table 3
Properties satisfied by gradual semantics for acyclic QBAFs (and, for Ebs, with arguments all having non-maximal base scores), namely: Attainability, Monotonicity, Strict Monotonicity, Franklin and Strict Franklin.

Discussion Predictably, all τ_{fixed} versions of the QBAF classifiers performed worse than their τ_{upvote} counterparts. However, the differences for DF-QuAD and QEM were less significant, suggesting that with the appropriate gradual semantics, the structure of the debate can be reasonably sufficient in determining a arbitrator's verdict. When using τ_{fixed}, each argument is equally weighted and all four gradual semantics satisfy monotony, though not necessarily the strict version, therefore an argument's strength is affected solely by the number of attackers and supporters it has, not the quality of its attacking and supporting arguments.

The performance of the Ebs classifiers were the most interesting since, depending on the choice of base score, results were disparate. That is, using τ_{upvote} produced the best results among all semantics for F_1 score, while using τ_{fixed} gave the worst. As shown in Table 2, recall was high using both τ_{fixed} and τ_{upvote}, but precision increased with τ_{upvote}. QuAD's performance was also significantly improved by changing from τ_{fixed} to τ_{upvote}. Like Ebs, τ_{upvote} did not really affect its recall but improved its precision; however, QuAD's average recall is quite low so it still performs the worst overall out of the four gradual semantics. DF-QuAD improves on the QuAD gradual semantics by removing the discontinuities and introducing the strict Franklin property while maintaining attainability and monotony. For example, in Figure 3b, Ebs is able to correctly predict the YTA verdict as the attack from argument 2 (the top comment) overcomes the support from argument 5 and decreases the strength of the original post - $\tau_{\text{upvote}}(1) = 0.2250$ by eq. (3). Contrastingly, QuAD narrowly fails and predicts NTA as the strength of supporting argument 5 is higher than the strength of the attacking argument 2, despite argument 2 having a greater base score due to a higher upvote score. This imbalance in the affects of supports and attacks is likely due to QuAD not satisfying the Franklin property.

Across all QBAF classifiers tested, recall is higher than precision, suggesting that argumentation frameworks are particularly useful for identifying the positive NTA verdict but it struggles to align with the negative YTA verdict. DF-QuAD and QEM have the most stable performance with the two versions of τ presented, exhibiting less significant changes in recall and precision. This is probably due to their satisfaction of attainability, bivariate montony and strict Franklin.

Note that the lack of an off-the-shelf method (based on LLMs) for the relation-based argument mining, not requiring training, could lead to some unpredictable results and errors in the prediction or relations between comments. For example, if in our first example (Figure 3a), the LLM determined that the top comment (argument 2) attacked the original post instead, all the calculation in the QBAF would have the wrong effect. Even if all other relations are correctly determined, all the gradual semantics we have explored would decrease the strength of the original post (argument 1) based on the strength of the top comment, causing the QBAF to calculate $\sigma(0) < 0.5$ and to incorrectly identify the OP as an asshole.

5. Conclusions and Future Work

In this paper, we have presented a novel approach to using symbolic techniques to model and explore human behaviour during debate and moral judgement. We outline a method for modelling an online debate, such as a thread on Reddit's `r/AmITheAsshole` subreddit, and explore how various choices of base scores (τ) and gradual semantics (σ) affect the performance of the QBAF as a classifier due to argumentation properties. Our work can be seen as steps towards building explainable and accurate models of online debate and human judgement which are lacking in the literature. Our QBAF classifier for Reddit threads was fairly successful and appears to rival other NLP-focused methods such as those in [8]. Moreover, it maintains a beneficial level of human-interpretability, as a user may select any argument and calculate its strength to determine how important it may have been to the verdict, in addition to efficiently examining the structure of the debate.

We suggest future work should first focus on refining the relation-based argument mining specifically for the purpose of mining social media threads. A fine-tuned transformer model could be developed to decrease the error in QBAF models of the Reddit threads. The experiments could then be redone to determine if an improvement in performance is possible. Moreover, the methodology we have outlined in this project is transferable and can be easily tweaked to be applied to other written debates. Firstly, it would be simple to apply the pipeline to other subreddits or other social media sites such as X (Twitter) and Facebook. Other suggested areas, away from online settings, include legal debates and cases, in which one could use argumentation frameworks to predict a judge's verdict on a case after the prosecution and defence have laid out their arguments. For example, a model could be developed to assign base scores to pieces of evidence and legal arguments in a case based on how trustworthy they are or how influential they could be. Then, a QBAF (with an appropriate gradual semantics) could be applied to predict if a jury or judge would likely pass a guilty verdict or not. This may be useful in the legal field in aiding decisions on which cases to take to trial or determining which need more evidence to be gathered. When applying QBAFs and gradual semantics in other contexts, it would thus be ideal to explore more informative algorithms, e.g. based on NLP, to determine the base score function τ in specific situations. Additionally, more attention needs to be placed on improving the relation-based argument mining for the chosen debate context.

Acknowledgments

Paulino-Passos, Rago and Toni were partially funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101020934). Rago and Toni were partially funded by J.P. Morgan and by the Royal Academy of Engineering under the Research Chairs and Senior Research Fellowships scheme. Any views or opinions expressed herein are solely those of the authors.

References

[1] H. Mercier, D. Sperber, Why do humans reason? arguments for an argumentative theory, Behavioral and brain sciences 34 (2011) 57–74.

[2] P. Baroni, D. Gabbay, M. Giacomin, L. van der Torre (Eds.), Handbook of Formal Argumentation, College Publications, 2018.

[3] K. Atkinson, P. Baroni, M. Giacomin, A. Hunter, H. Prakken, C. Reed, G. R. Simari, M. Thimm, S. Villata, Towards artificial argumentation, AI Magazine 38 (2017) 25–36.

[4] J. Lawrence, C. Reed, Using complex argumentative interactions to reconstruct the argumentative structure of large-scale debates, in: ArgMining@EMNLP, 2017, pp. 108–117.

[5] K. Cyras, A. Rago, E. Albini, P. Baroni, F. Toni, Argumentative XAI: A survey, in: IJCAI, 2021, pp. 4392–4399.

[6] P. Baroni, A. Rago, F. Toni, From fine-grained properties to broad principles for gradual argumentation: A principled spectrum, Int. J. Approx. Reason. 105 (2019) 252–286.

[7] O. Cocarascu, A. Rago, F. Toni, Extracting Dialogical Explanations for Review Aggregations with Argumentative Dialogical Agents, in: AAMAS, 2019.

[8] I. S. Efstathiadis, G. Paulino-Passos, F. Toni, Explainable patterns for distinction and prediction of moral judgement on reddit, CoRR abs/2201.11155 (2022). arXiv:2201.11155.

[9] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: NAACL-HLT, 2019, pp. 4171–4186.

[10] P. Baroni, A. Rago, F. Toni, How many properties do we need for gradual argumentation?, in: AAAI, 2018, pp. 1736–1743.

[11] P. Baroni, M. Romano, F. Toni, M. Aurisicchio, G. Bertanza, Automatic evaluation of design alternatives with quantitative argumentation, Argument Comput. 6 (2015) 24–49.

[12] A. Rago, F. Toni, M. Aurisicchio, P. Baroni, Discontinuity-free decision support with quantitative argumentation debates, in: KR, 2016, pp. 63–73.

[13] N. Potyka, Continuous dynamical systems for weighted bipolar argumentation, in: KR, 2018, pp. 148–157.

[14] L. Amgoud, J. Ben-Naim, Evaluation of arguments in weighted bipolar graphs, Int. J. Approx. Reason. 99 (2018) 39–55.

[15] L. Amgoud, J. Ben-Naim, Evaluation of arguments in weighted bipolar graphs, in: ECSQARU, 2017, pp. 25–35.

[16] B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heintz, D. Roth, Recent advances in natural language processing via large pre-trained language models: A survey, ACM Comput. Surv. 56 (2024) 30:1–30:40.

[17] D. Gorur, A. Rago, F. Toni, Can large language models perform relation-based argument mining?, CoRR abs/2402.11243 (2024). arXiv:2402.11243.

[18] A. Hagberg, D. Schult, P. Swart, J. M. Hagberg, Exploring Network Structure, Dynamics, and Function using NetworkX, in: SciPy, 2008.