Artificial Intelligence 340 (2025) 104291

Contents lists available at ScienceDirect

Artificial Intelligence

journal homepage: www.elsevier.com/locate/artint

Argumentative review aggregation and dialogical explanations

Antonio Rago [a,*], Oana Cocarascu [b], Joel Oksanen [a], Francesca Toni [a]

[a] Department of Computing, Imperial College London, UK
[b] Department of Informatics, King's College London, UK

ARTICLE INFO

MSC:
00-01
99-00

Keywords:
Argumentation
Argument mining
Review aggregation
Dialogical interaction
Conversational explanation

ABSTRACT

The aggregation of online reviews is one of the dominant methods of quality control for users in various domains, from retail to entertainment. Consequently, *explainable aggregation of reviews* is increasingly sought-after. We introduce quantitative argumentation technology to this setting, towards automatically generating reasoned review aggregations equipped with dialogical explanations. To this end, we define a novel form of *argumentative dialogical agent* (ADA), using ontologies to harbour information from reviews into argumentation frameworks. These agents may then be evaluated with a quantitative argumentation semantics and used to mediate the generation of dialogical explanations for item recommendations based on the reviews. We show how to deploy ADAs in three different contexts in which argumentation frameworks are mined from text, guided by ontologies. First, for hotel recommendations, we use a human-authored ontology and exemplify the potential range of dialogical explanations afforded by ADAs. Second, for movie recommendations, we empirically evaluate an ADA based on a bespoke ontology (extracted semi-automatically, by natural language processing), by demonstrating that its quantitative evaluations, which are shown to satisfy desirable theoretical properties, are comparable with those on a well-known movie review aggregation website. Finally, for product recommendation in e-commerce, we use another bespoke ontology (extracted fully automatically, by natural language processing, from a website's reviews) to construct an ADA which is then empirically evaluated favourably against review aggregations from the website.

1. Introduction

In an age in which e-commerce, social media and streaming platforms are dominant markets for consumers, items' online reviews are fast becoming the preferred method of quality control for users. Given the sheer quantities of users, items and reviews in these domains, websites providing these services almost universally aggregate reviews so that the information is cognitively manageable by users. Such websites include (among many others): *Trip Advisor*,[1] which aggregates users' reviews of hotels and other travel services; *Rotten Tomatoes*[2] (RT), which aggregates critics' (and, more recently, fans') reviews of movies; and *Amazon*,[3] an e-commerce website which provides users' reviews of its products of many kinds. Indeed, it is becoming more and more difficult to find websites providing

* Corresponding author.
E-mail addresses: a.rago@imperial.ac.uk (A. Rago), oana.cocarascu@kcl.ac.uk (O. Cocarascu), joel.oksanen@icloud.com (J. Oksanen), ft@imperial.ac.uk (F. Toni).

[1] https://www.tripadvisor.com.
[2] https://www.rottentomatoes.com.
[3] https://www.amazon.co.uk.

https://doi.org/10.1016/j.artint.2025.104291
Received 11 January 2024; Received in revised form 7 January 2025; Accepted 9 January 2025

Available online 15 January 2025

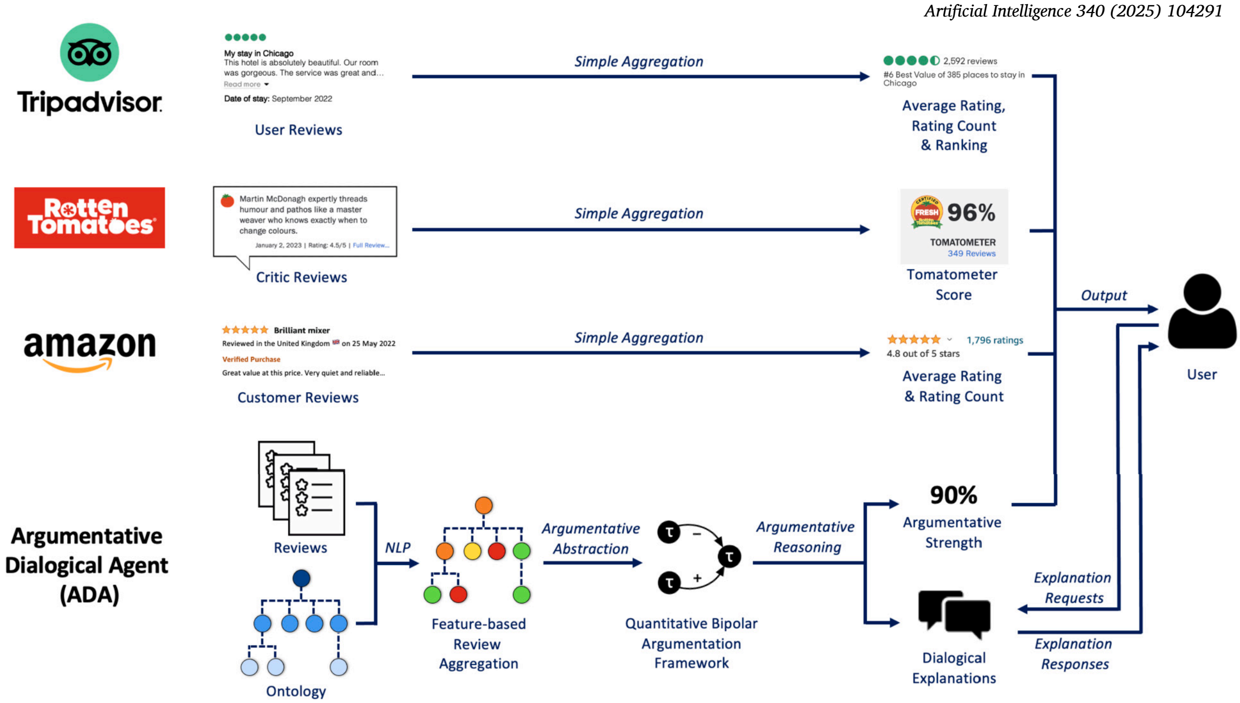


Fig. 1. An overview of ADAs, demonstrating their capacity for extracting aggregated review scores with dialogical explanations.

some catalogue of items to users *without* some form of reviewing capacity and aggregation thereof. These review aggregations are very valuable to these websites, due to increased user interest that they facilitate [15]. A possible explanation for this increase is the fact that they foster trust in users, since they provide a somewhat independent arbitration mechanism for quality, separate to the provider (and any conflict of interests). Such is the importance of this trust that some popular review aggregations have even been integrated in other websites, e.g. RT's *Tomatometer Score* (TS) in streaming sites such as *Apple TV* and search engines such as *Google*.

This phenomenon is not without problems. Consider RT in the movie domain: numerous stakeholders have bemoaned RT's apparently detrimental effect on the movie industry,[4] with one of the more plausible claims being that the ubiquitous TS score (defined in terms of the percentage of critics who believe a movie is *fresh* rather than *rotten*) oversimplifies a movie's aggregated review and "hacks off critical nuance".[5] Another issue is the fact that the TS score represents the percentage of critics who felt anywhere from mildly to wildly positively about a given movie, blurring away differences between critics with mixed, slightly positive views and critics with rave, strongly positive views. Further, the TS score hides away critics' preferences and factors which decrease the TS score but may not have any relevance in a user's personal selection criteria, meaning movies may be overlooked when they may actually be perfectly suited to a user's tastes. These issues are not confined to the movie domain, and result from the widespread opacity of review aggregation sites, which make little effort to explain the reasoning behind the aggregations they provide. That being said, simple numerical measures, such as the TS score, are useful and intuitive indicators for users of all backgrounds. In this paper, we aim to alleviate the issues mentioned above while maintaining the existing advantages, by proposing a novel methodology for review aggregation to complement simple numerical measures with *dialogical explanations*, which are drawn from the reviews underlying the systems and empower users to interact with these systems. Our choice of such dialogical, and thus human-like, explanations is influenced by the findings of [62], notably that explanations should be *social*, *selective* and *contrastive*.

Our methodology leverages techniques from *computational argumentation*, as understood in AI (see [6,9] for overviews). Generally, this comprises formal models for representing and reasoning with knowledge to identify and resolve conflicts therein. While *structured argumentation frameworks* (e.g. see [13]) rely upon a fine-grained representation for arguments and dialectical relations between them based on the arguments' components, we focus on a more abstract view of arguments, as in *abstract argumentation* [36]. Here, *arguments*, linked by a dialectical relation of *attack* in a graph, are evaluated by means of a *semantics*. In our setting, arguments abstractly represent the "goodness" of items and/or their features. Note that our use of unstructured arguments allows us to distill the views expressed in reviews, rather than spelling out the structure of arguments therein, which may be enthymemes [46], i.e. with missing premises. For example, a reviewer may state the view that the acting in a specific movie was great/poor without stating why, e.g. in terms of which actors' performances specifically led to this conclusion: we represent this view as a positive/negative *vote* (respectively) on an unstructured argument about the acting's goodness in the movie.

[4] https://www.theatlantic.com/amp/article/543090/.
[5] https://nyti.ms/2xcXS0Y.

Our methodology makes use of *quantitative bipolar argumentation frameworks* (QBAFs) [10], which include, in addition to the attack relation, a dialectical relation of *support*, as well as *base scores* for arguments, obtained in this paper from positive and negative votes drawn from reviews. Using these QBAFs we define the novel concept of *argumentative dialogical agents (ADAs)*, overviewed in Fig. 1. Rather than simply aggregating reviews on an item using averaging techniques and the like, as is commonly the case on leading websites, an ADA, equipped with an ontology, first employs natural language processing (NLP) to break down the reviews into a feature-based review aggregation. This ontology may be given manually (as is the case for all illustrations until and including Section 7) or automated to various degrees (as we will demonstrate in Section 8), allowing rich, feature-based representation of the reviews to be obtained in a number of settings. Then, the ADA extracts a QBAF corresponding to the feature-based review aggregation, and then applies argumentative semantics to obtain a quantitative strength for arguments/items corresponding to the other measures in the literature. Our theoretical contributions then highlight how certain semantics satisfy properties which are particularly desirable in this domain. Finally, the ADA complements the argumentative strength with dialogical explanations detailing to users how the score, amounting to the quantitative strength of items, was obtained, thus guaranteeing that the ADA is explainable.

The paper is structured as follows. After considering related work (Section 2) and giving the necessary background (Section 3), we begin by defining the extraction process underpinning ADAs from reviews and an ontology to a feature-based review aggregation via NLP techniques, where each feature is weighted by votes extracted from the reviews (Section 4). We then define the process of abstracting QBAFs from review aggregations, proving several properties of the resulting QBAFs which are desirable in this setting (Section 5). Following this, we undertake a formal analysis of three argumentative semantics in this setting specifically, evaluating their suitability here based on properties from the literature in order to select the most appropriate method for generating an argumentative strength (Section 6). We then define the final step, from the QBAF to the dialogical explanations (Section 7). Throughout these sections, we illustrate ADAs' components with a running example as a first case study in the hotel review setting, mimicking information from a website such as Trip Advisor, with a human-authored ontology. Next, we detail two case studies (Section 8). The first demonstrates an ADA being applied to movie reviews (Section 8.1), based on a bespoke ontology, extracted semi-automatically via NLP, and compare it with the TS score from RT, the aforementioned movie review aggregation website. The second case study demonstrates an ADA for product recommendation in e-commerce (Section 8.2). Here, we use a further bespoke ontology, this time extracted fully automatically via NLP using reviews on Amazon's website. Finally, we conclude, looking ahead to future work (Section 9).

This paper builds upon and expands [30] in a number of ways (in addition to fixing typographical errors and simplifying notation):

- We generalise the ADA framework so that it works in any domain where an ontology is available or can be extracted, besides and beyond movies with the application-specific ontology we used in [30].
- This general framework provides a more accurate representation of the argumentation involved, e.g. giving arguments an explicit polarity.
- We undertake a comprehensive analysis of the properties of ADAs in general, as well as of the properties of argumentative semantics which may be used to obtain an argumentative strength here.
- We exemplify ADAs' potential in new contexts, i.e. in the hotel and e-commerce domains. For the latter, we use a novel method for automatically generating an ontology in the e-commerce domain first sketched in [65].
- We adapt the dialogical explanations from [30] to a different context, demonstrating its generality.

2. Related work

2.1. Review aggregation and argumentation

With regards to argumentation-based review aggregation, our method bears some similarities with [23,22,21]. These works define a rating system to represent a collection of reviews as an abstract argumentation framework [36], with each review as an argument and weighted attacks extracted between those which concern a similar topic and contradict one another w.r.t. the rating. The authors show how their method affords explainability, but do not use any support relation or gradual evaluation in their approach and, furthermore, do not envisage that multiple arguments may be drawn from the same review, as we do, driven by ontologies. Another related work, [19], uses a semantic network of topics manually associated to arguments in abstract argumentation frameworks with the help of labels (hashtags), to obtain multi-topic argumentation frameworks analysed in terms of a variant of classical acceptability [36] defined in terms of a notion of distance of arguments in the semantic network. While this work shares with our proposal the use of semantic information (a semantic network in [19] and an ontology in our ADA framework), again they do not use any support relation or gradual evaluation in their approach and, furthermore, they do not envisage or support the mining of arguments and relations from reviews.

Meanwhile, [95] utilise argumentative reasoning in this domain by mining arguments from reviews, but they deploy a different form of argumentation framework to that considered here. Specifically, the authors use *argumentation schemes* [93], i.e. pre-determined templates of argumentative dialogues, to highlight argumentative patterns in reviews for users. Argument schemes were also used in [63] to model the reliability of reviews. While our approach is more constrained than these argument schemes in the types of arguments we allow, we introduce an end-to-end pipeline for explainable review aggregation driven by gradual semantics. Cross-fertilisations between the approaches, which we believe could be fruitful, are left to future work.

Other relevant research on review aggregation concerns the evaluation of the linguistic quality of reviews, e.g. [42,94,48,35], but none of these existing works consider the quality from an argumentative viewpoint. However, evaluating our approach on metrics such as the helpfulness to users, as is considered in all of these approaches, would be interesting future work.

2.2. Argument mining

Argument mining is a well studied field (see [51,50] for recent overviews). It can be seen as an advancement of sentiment analysis and opinion mining (as indicated in [44,51]). Whilst the goal of sentiment analysis and opinion mining is to identify what users think, the goal of argument mining is to understand the reasons why users think the way they do, not limiting to opinions and their sentiment polarity. In our setting, one can see opinions as arguments as to why users should or should not choose an item (e.g. a movie).

Amongst several views on argument mining, the one we use is closest to the one referred to as *relation-based argument mining* in [20] and used, e.g., in [16], where attack and support relations are mined from tweets, i.e. texts which are similar in size to the review snippets we consider in this paper. In our experimental evaluations in Section 8, we use two different techniques for supporting relation-based argument mining: one adapting feature-based sentiment analysis and the other based on deep learning [31].

Several works with applications in the movie setting make use of (feature-based) sentiment analysis and machine learning techniques. Related works include those proposing aggregation methods for recommending movies [97] and document-level sentiment analysis in the movie setting [59,14,84,88,25]. Other works focus on extracting important features from reviews using machine learning techniques [99,66,47].

Our dialogical explanations can be seen as an argumentation-based summary of reviews. [38] propose a general summarisation framework based on abstract argumentation to select sentences from text. This summarisation extracts the most relevant information (arguments) from reviews. We use the mined arguments and relations to generate votes that are used to obtain argumentation frameworks.

2.3. Explainable AI and argumentation

The ability of argumentation to represent and reason with knowledge while transparently resolving conflicts has proven useful to support various forms of explainable AI (see [33,91] for recent surveys), e.g. in computational optimisation, where schedules can be explained by means of argumentation frameworks [32], and in recommender systems, where human-like, interactive explanations can help to increase transparency of the models' internal functionalities [71]. More generally, argumentation can be used to explain AI models' processes in generating outputs and managing conflicts, e.g. between two mutually exclusive options, in a manner which may be seen as being much more amenable to humans than the AI methods. For example, it has been argued that the majority of statements in explanation [4], and in fact all reasoning performed by humans [58], are argumentative. Further, many of the desirable features of explanation, notably that they should be social, selective and contrastive [62], may be supported by argumentation, as we will discuss in Section 7. AI models which have been explained argumentatively include neural systems commonly seen as "black boxes" [43], e.g. [1,83,8] provide argumentation-based explanations for the workings and outputs of neural classifiers. Our focus in this paper is on providing a transparent, argumentation-based method for explainable review aggregations, with the help of black-box techniques to support the mining of argumentation frameworks.

2.4. Conversational AI

Conversational AI systems have become extremely popular as the technologies behind speech recognition and NLP have allowed machines to converse in an ever more human-like manner, as exemplified by the widespread use of virtual assistants, e.g. Amazon's *Alexa*, *Microsoft*'s *Cortana*, *Google Assistant* and *Apple*'s *Siri*. Such conversational AI systems have been shown to be useful in numerous settings: information retrieval and search [98]; preference elicitation in recommender systems [29]; and explanations of predictions or classifications [72]. Our dialogical explanations can be seen as a template-driven form of conversational AI, driven by the argumentation frameworks resulting from review aggregations.

Recently, earlier forms of conversational AI technologies have been somewhat upstaged by the meteoric rise of large language models (LLMs) such as in [18,87,28]. These multi-billion parameter models are trained on enormous datasets to produce incredibly rich and realistic conversational interactions with humans across a wide range of tasks. While these LLMs could in principle be applied to review aggregation, they lack explainability. Indeed, as is often the case with data-driven approaches [79], while the outputs delivered by LLMs may be remarkably convincing, these models are "black-boxes", in that the reasoning for their outputs is often not only difficult to explain but completely uninterpretable to humans. Further, LLMs are known to be prone to "hallucinations" [78], i.e. outputs which may be spoken in the utmost confidence and sound convincing but are, in fact, completely false. These issues are particularly problematic in review aggregation on commercial websites, where the trust between stakeholders, e.g. the reviewers, the producers/purchasers of the items or the owners of the website, is critical but also precarious. Hallucinations, or even just outputs without justification via the model's reasoning, could immediately render a review aggregation useless from a user's perspective since the aggregation would be seen as being unfair. Some attempts have been made to use LLMs to provide explanations for recommendations in the literature [53,76]. However, to the best of our knowledge, they have not yet been proven to be faithful to the method for generating the recommendations: whether the LLM itself is providing the recommendation or not, we have no

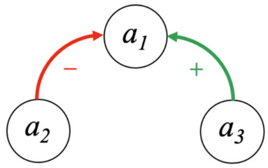


Fig. 2. Graphical representation of a simple QBAF $\langle\mathcal{A},\mathcal{L}^-,\mathcal{L}^+,\tau\rangle$ where $\mathcal{A}=\{a_1,a_2,a_3\}$, $\mathcal{L}^-=\{(a_2,a_1)\}$, $\mathcal{L}^+=\{(a_3,a_1)\}$, $\tau(a_1)=0.5$, $\tau(a_2)=0.8$ and $\tau(a_3)=0.2$. Here, the arguments are represented by nodes in the graph and the attack and support relations are represented by red edges labelled − and green edges labelled +, respectively. (The base scores are not visualised in the graphical representation.) (For interpretation of the colours in the figure(s), the reader is referred to the web version of this article.)

guarantees that it is in fact explaining this process. In fact, it has been shown that the explanations offered by LLMs for their outputs across different settings are not always faithful [89].

There is, therefore, a need for review aggregation systems which utilise the immense power of AI but are, critically, faithfully explainable, in that their reasoning process is completely transparent to users in order to maintain their trust. Our ADA framework aims at filling this gap: while we focus on showing how it can support template-driven dialogues, we believe it could be naturally enriched with the power of LLMs. This enrichment could occur at different points on the pipeline in Fig. 1, e.g. mining our specific form of ontology using fine-tuned LLMs, as overviewed in [96] for different settings, or generating more natural interactions with humans, such as by rephrasing the template-based responses, as we will discuss in Section 7. We leave the exploration of these interesting directions of study for future work.

2.5. Dialogical argumentation

Dialogical argumentation has been studied in various settings, e.g. dialogue games for argumentation [69], persuasion dialogue [70] and strategic argumentation [45]. Similarly, argumentation has also been used to support dialogical agents in games [7], negotiation [57] and decision making with goals [37]. Argumentation-based chatbots have also risen to popularity recently, with a number of goals such as supporting debate [81], persuasion [24] and explanation [80]. We focus on this final use, for supporting the exchange of explanations with users, which has been attempted for a number of tasks other than review aggregation, e.g. in explaining reinforcement learning models [55] or knowledge-based systems [5]. Meanwhile, the advantages of argumentation in driving (dialogical) explanations have also been exploited in the recommender systems domain, which is closely related to that we consider here. The argumentation-based explanations for recommendations in the system of [72] are generated automatically from data without any need for knowledge to be manually incorporated. This system provides argumentation explanations of recommendations that support conversational interactions with a user, both aiding the transparency of the system and guaranteeing that recommendations are improved based on the elicited feedback from the user. Some [27,17,85] use defeasible logic programming (DeLP) [40], another formal argumentation model, to enhance recommendation technologies with argument-based analysis. The movie recommender system of [17] relies on a set of predefined postulates describing the conditions under which a movie should be recommended to a given user and which can be translated into DeLP rules. Our dialogical explanations are extracted from a form of quantitative argumentation mined from reviews with the help of NLP techniques.

3. Background

As in [10], a QBAF is a quadruple $\langle\mathcal{A},\mathcal{L}^-,\mathcal{L}^+,\tau\rangle$ consisting of a set $\mathcal{A}$ of *arguments*, a binary (*attack*) relation $\mathcal{L}^-\subseteq\mathcal{A}\times\mathcal{A}$, a binary (*support*) relation $\mathcal{L}^+\subseteq\mathcal{A}\times\mathcal{A}$ and a total function $\tau:\mathcal{A}\to\mathbb{I}$, for $\mathbb{I}$ a set of values, where, for any $a_i\in\mathcal{A}$, $\tau(a_i)$ is the *base score* of a_i.

For any argument $a_i\in\mathcal{A}$, we use $\mathcal{L}^-(a_i)=\{a_j\in\mathcal{A}|(a_j,a_i)\in\mathcal{L}^-\}$ to refer to a_i's *attackers* and $\mathcal{L}^+(a_i)=\{a_j\in\mathcal{A}|(a_j,a_i)\in\mathcal{L}^+\}$ to refer to a_i's *supporters*. Also, for any arguments $a_i,a_j\in\mathcal{A}$, we let a *path* from a_i to a_j via a relation $\mathcal{L}\subseteq\mathcal{A}\times\mathcal{A}$ be defined as a sequence $(a_0,a_1),\dots,(a_{n-1},a_n)$ for some $n>0$ where $a_0=a_i$, $a_n=a_j$ and, for any $1\le k\le n$, $(a_{k-1},a_k)\in\mathcal{L}$. Then, we say that a QBAF $\langle\mathcal{A},\mathcal{L}^-,\mathcal{L}^+,\tau\rangle$ is a *tree* iff $\exists a_i\in\mathcal{A}$ (called the *root*) such that i) $\nexists(a_i,a_j)\in\mathcal{L}^-\cup\mathcal{L}^+$; ii) $\forall a_k\in\mathcal{A}\setminus\{a_i\}$, there is a path from a_k to a_i via $\mathcal{L}^-\cup\mathcal{L}^+$; and iii) $\nexists a_l\in\mathcal{A}$ with a path from a_l to a_l.

QBAFs can be equipped with a number of *gradual semantics* [10], which are total functions $\sigma:\mathcal{A}\to\mathbb{I}$ and which, for any $a_i\in\mathcal{A}$, assign a *strength* $\sigma(a_i)$ to a_i. For any set of arguments $S\subseteq\mathcal{A}$, we denote $\sigma(S)$ a sequence (in any order) of all elements of the set $\{\sigma(a_i)|a_i\in S\}$ (thus $\sigma(S)\in\mathbb{I}^*$, where $\mathbb{I}^*$ is the set of all sequences of elements of $\mathbb{I}$). When comparing two sequences $\sigma(S_1)$ and $\sigma(S_2)$, for $S_1,S_2\subseteq\mathcal{A}$, we will treat the sequences as multi-sets. Throughout this paper, we will use $\mathbb{I}=[0,1]$.

In our theoretical and experimental evaluation, we will consider three gradual semantics for QBAFs:the *Quantitative Argumentation Debate algorithm* (QuAD, indicated in the remainder with σ_q) [11], the *Discontinuity-Free Quantitative Argumentation Debate algorithm* (DF-QuAD, indicated in the remainder with σ_d) [75] and the *Restricted Euler-based semantics* (REB, indicated in the remainder with σ_r) [3]. We briefly recap them below, for a given QBAF $\langle\mathcal{A},\mathcal{L}^-,\mathcal{L}^+,\tau\rangle$, and exemplify them in the case of the QBAF given in Fig. 2.

In QuAD, the strengths of an argument's attackers and the strengths of its supporters are first combined (separately) with the argument's base score, before these two combined scores are aggregated to give an overall strength for the argument. Formally, for $a_i\in\mathcal{A}$, $\sigma_q(a_i)=g(\tau(a_i),\mathcal{F}_a(\tau(a_i),\sigma_q(\mathcal{L}^-(a_i))),\mathcal{F}_s(\tau(a_i),\sigma_q(\mathcal{L}^+(a_i))))$ where, if $(a_1,\dots,a_n)$ is an arbitrary permutation of the ($n\ge 0$) attackers in $\mathcal{L}^-(a_i)$, $\sigma_q(\mathcal{L}^-(a_i))=(\sigma_q(a_1),\dots,\sigma_q(a_n))$ (similarly for supporters). The operator $g:\mathbb{I}\times\mathbb{I}\cup\{nil\}\times\mathbb{I}\cup\{nil\}\to\mathbb{I}$ is defined, for $v_0,v_a,v_s\in\mathbb{I}$, as: if $v_s=nil$ and $v_a\neq nil$, $g(v_0,v_a,v_s)=v_a$; if $v_a=nil$ and $v_s\neq nil$, $g(v_0,v_a,v_s)=v_s$; if $v_a=v_s=nil$,

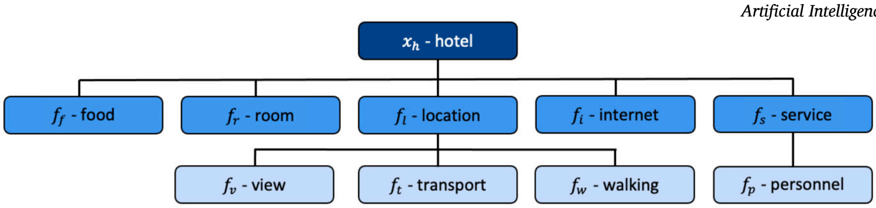


Fig. 3. Graphical representation of the item ontology for the running hotel illustration. A child feature is related to, but more specific than, its parent feature, e.g. f_t is more specific than f_l.

$g(v_0, v_a, v_s) = v_0$; otherwise $g(v_0, v_a, v_s) = \frac{v_a+v_s}{2}$. Letting $\times$ stand for either a or s, the operator $\mathcal{F}_\times$ is defined as $\mathcal{F}_\times : \mathbb{I}^* \to \mathbb{I}$, where for $S = (v_1, \dots, v_m) \in \mathbb{I}^*$, $(w_1, \dots, w_n)$ is an arbitrary permutation of the non-zero elements in S:[6] if $n = 0$: $\mathcal{F}_\times(v_0, S) = nil$; if $n = 1$: $\mathcal{F}_\times(v_0, S) = f_\times(v_0, w)$; if $n > 1$: $\mathcal{F}_\times(v_0, (w_1, \dots, w_n)) = f_\times(\mathcal{F}_\times(v_0, (w_1, \dots, w_{n-1})), w_n)$; with the *base expressions* $f_\times : \mathbb{I} \times \mathbb{I} \to \mathbb{I}$ defined, for $v_0, v \in \mathbb{I}$, as: $f_a(v_0, v) = v_0 \cdot (1 - v)$ and $f_s(v_0, v) = v_0 + v - v_0 \cdot v$. For the QBAF in Fig. 2, for both a_2 and a_3 it holds that $v_a = nil$ and $v_s = nil$, and thus $\sigma_q(a_2) = \tau(a_2) = 0.8$ and $\sigma_q(a_3) = \tau(a_3) = 0.2$. Then, for a_1, we obtain $v_a = 0.5 \cdot (1 - 0.8) = 0.1$ and $v_s = 0.5 + 0.2 - (0.5 \cdot 0.2) = 0.6$, and thus $\sigma_q(a_1) = \frac{0.1+0.6}{2} = 0.35$.

In DF-QuAD, discontinuities in the strengths given by QuAD are avoided by first combining the strengths of an argument's attackers and the strengths of its supporters (separately), before the argument's strength is calculated as either decreasing or increasing the base score depending on which of the attackers or supporters, respectively, are stronger. Formally, for any $a_i \in \mathcal{A}$ with $\tau(a_i) = v_0$ and $n \geq 0$ attackers with strengths $v_1, \dots, v_n$ and $m \geq 0$ supporters with strengths $v'_1, \dots, v'_m$, $\sigma_d(a_i) = \mathcal{C}(v_0, \mathcal{F}(v_1, \dots, v_n), \mathcal{F}(v'_1, \dots, v'_m))$, where the combination function, $\mathcal{C}$, is defined as follows, for $v_a = \mathcal{F}(v_1, \dots, v_n)$ and $v_s = \mathcal{F}(v'_1, \dots, v'_m)$: if $v_a = v_s$ then $\mathcal{C}(v_0, v_a, v_s) = v_0$; else if $v_a > v_s$ then $\mathcal{C}(v_0, v_a, v_s) = v_0 - (v_0 \cdot |v_s - v_a|)$; otherwise $\mathcal{C}(v_0, v_a, v_s) = v_0 + ((1 - v_0) \cdot |v_s - v_a|)$. Given n arguments with strengths $v_1, ..., v_n$, if $n = 0$ then $\mathcal{F}(v_1, ..., v_n) = 0$, otherwise $\mathcal{F}(v_1, ..., v_n) = 1 - \prod_{i=1}^{n}(|1 - v_i|)$. For the QBAF in Fig. 2, for both a_2 and a_3 it holds that $v_a = v_s = 0$, and thus $\sigma_d(a_2) = \tau(a_2) = 0.8$ and $\sigma_d(a_3) = \tau(a_3) = 0.2$. Then, for a_1, we obtain $v_a = 1 - (|1 - 0.8|) = 0.8$ and $v_s = 1 - (|1 - 0.2|) = 0.2$. Thus, $v_a > v_s$ and $\sigma_d(a_1) = 0.5 - (0.5 \cdot |0.2 - 0.8|) = 0.2$.

The REB semantics is such that the difference between the summated strengths of an argument's attackers and the summated strengths of its supporters is calculated, before this is combined with the argument's base score to give an overall strength for the argument. Formally, for $a_i \in \mathcal{A}$, $\sigma_r(a_i) = 1 - \frac{1-\tau(a_i)^2}{1+\tau(a_i)\cdot e^E}$ where $E = \sum_{a_j \in \mathcal{L}^+(a_i)} \sigma_r(a_j) - \sum_{a_k \in \mathcal{L}^-(a_i)} \sigma_r(a_k)$. For the QBAF in Fig. 2, for both a_2 and a_3 it holds that $E = 0$, and so $\sigma_r(a_2) = 1 - \frac{1-0.8^2}{1+0.6} = 0.8$ and $\sigma_r(a_3) = 1 - \frac{1-0.2^2}{1+0.2} = 0.2$. Then, for a_1, $E = 0.2 - 0.8 = -0.6$ and so $\sigma_r(a_1) = 1 - \frac{1-0.5^2}{1+0.5\cdot e^{-0.6}} = 0.411$.

Note that, even in this simple example, the three gradual semantics we consider give different strengths for a_1, and so it is critical to assess theoretically their suitability for review aggregation: we will do so in Section 6.

4. ADAs: from item ontologies to review aggregation frameworks

In this section we introduce the first component in the pipeline of *argumentative dialogical agents (ADAs)*, our formalism for explainable review aggregation (as overviewed in Fig. 1). This component relies upon the association of votes, expressed in reviews, with components of an ontology. The ontology serves as a feature-based characterisation of the items being reviewed, which may be hotels, movies or products, for example.

Definition 1. An *item ontology* for an item x is a tree $\langle \mathcal{N}, \mathcal{E} \rangle$ such that:

- $\mathcal{N} = \{x\} \cup \mathcal{F}$ where $\mathcal{F}$ is a non-empty set of features for the item x;
- $\mathcal{E} \subseteq \mathcal{F} \times (\{x\} \cup \mathcal{F})$ is a directed acyclic relation such that $\forall n_i \in \mathcal{N} \setminus \{x\}$ (equivalently, $\forall n_i \in \mathcal{F}$), $\exists! n_j \in \mathcal{N}$ where $(n_i, n_j) \in \mathcal{E}$; we say that n_j is the (unique) *parent* of n_i and that n_i is a *child* of n_j; with an abuse of notation, we let $\mathcal{E}(n_i) = n_j$.

Note that, given that we do not impose that x has a parent, necessarily it is the root of the tree.

Intuitively, an item ontology represents a reviewed item along with its features, organised in a tree. Typically, any parent feature is more general and less specific than its child features, which is nonetheless related to the parent. For example, in the hotel domain, a feature about transport is more specific than one about location, as location is characterised by transport, amongst other features. Also, in the movie domain, a feature representing the specific actor Ralph Fiennes may be the child of the feature representing acting in general.

Here and up to Section 8, we will exemplify our approach using a running example in the context of hotel reviewing, as an illustration of our proposed methodology. A simple item ontology (for a generic hotel x_h) is graphically represented in Fig. 3.

Item ontologies may be obtained in a number of ways with varying levels of automation: they may be manually constructed based on human judgement, as in the case of our running hotel example; they may be semi-automated, in that a subset of the features

[6] This formulation is a modification of the original formulation of $\mathcal{F}_\times$ in [11], in which $(w_1, \dots, w_n)$ was not used.

are given by hand before the others are obtained from reviews using NLP techniques (as we shall do in Section 8.1); or they may be fully-automated, in that they are entirely extracted from reviews by NLP techniques (as we shall do in Section 8.2). These item ontologies facilitate the extraction of QBAFs from reviews, as we will see.

Once an item ontology has been obtained (from human input and/or NLP techniques), ADAs use votes (on the item or the features) drawn from reviews as a basis for aggregating the reviews in an orderly and convenient argumentative structure.

Definition 2. A *review aggregation framework* for an item x is a triple $\langle \mathcal{O}, \mathcal{R}, \mathcal{V} \rangle$ where:

- $\mathcal{O} = \langle \{x\} \cup \mathcal{F}, \mathcal{E} \rangle$ is an item ontology for x;
- $\mathcal{R}$ is a finite, non-empty set of *reviews*;
- $\mathcal{V} : \mathcal{R} \times (\{x\} \cup \mathcal{F}) \rightarrow \{-, +\}$ is a (possibly partial) function, with $\mathcal{V}(\rho, n_i)$ representing the *vote* of the review $\rho \in \mathcal{R}$ on $n_i \in \{x\} \cup \mathcal{F}$.

Straightforwardly, a positive/negative vote from a review on a feature of the item signifies a positive/negative, respectively, review of that feature and a positive/negative vote on x signifies a positive/negative, respectively, review of the overall item. Note that we assume that each review expresses at most one opinion (vote) about each *node* (item or feature) of the ontology. In doing so, we require that nodes for which there are multiple possible votes, e.g. where multiple contradictory opinions on a node are present, are characterised here by the "predominant" vote, the selection of which will depend on the application. For illustration, in the hotel example, consider a review which contains three comments concerning the hotel's food (f_f), two of which are positive (i.e. that, at breakfast, the eggs are great and, at dinner, the pancakes are spectacular) and one is negative (i.e. that the room service offers very limited choice of food). In this case we opt to extract one positive vote on f_f only from this review, given that this represents the majority and thus the predominant vote. We impose this restriction for simplicity only: accommodating other, more complex voting mechanisms would not be problematic (but may exhibit different behaviour to what we describe in this paper). Note that we assume here that votes are given, as our focus is on illustrating the machinery we are introducing rather than deploying it. We will explore ways to obtain votes automatically from reviews and ontologies, with the support of NLP techniques, in the case studies in Section 8.

Given that reviews may not explicitly contain all of the information useful to review aggregation, we allow review aggregation frameworks to be *augmented* by exploiting the parent relation in item ontologies.

Definition 3. Given a review aggregation framework $\langle \mathcal{O}, \mathcal{R}, \mathcal{V} \rangle$ for an item x, where $\mathcal{O} = \langle \mathcal{N}, \mathcal{E} \rangle$, an *augmented review aggregation framework* for x (corresponding to $\langle \mathcal{O}, \mathcal{R}, \mathcal{V} \rangle$) is a review aggregation framework $\langle \mathcal{O}, \mathcal{R}, \mathcal{V}' \rangle$ for x such that for any $\rho \in \mathcal{R}$ and $n_i \in \mathcal{N}$:

- if n_i is a leaf then:
  - if $\mathcal{V}(\rho, n_i)$ is defined, then $\mathcal{V}'(\rho, n_i) = \mathcal{V}(\rho, n_i)$; else
  - $\mathcal{V}'(\rho, n_i)$ is undefined.
- if n_i is not a leaf then:
  - if $\mathcal{V}(\rho, n_i)$ is defined, then $\mathcal{V}'(\rho, n_i) = \mathcal{V}(\rho, n_i)$; else
  - if $|\{n_j \in \mathcal{N} | \mathcal{E}(n_j) = n_i, \mathcal{V}'(\rho, n_j) = +\}| > |\{n_k \in \mathcal{N} | \mathcal{E}(n_k) = n_i, \mathcal{V}'(\rho, n_k) = -\}|$ then $\mathcal{V}'(\rho, n_i) = +$; else
  - if $|\{n_j \in \mathcal{N} | \mathcal{E}(n_j) = n_i, \mathcal{V}'(\rho, n_j) = +\}| < |\{n_k \in \mathcal{N} | \mathcal{E}(n_k) = n_i, \mathcal{V}'(\rho, n_k) = -\}|$ then $\mathcal{V}'(\rho, n_i) = -$; else
  - $\mathcal{V}'(\rho, n_i)$ is undefined.

Intuitively, in augmented review aggregations, any positive or negative augmented vote can only be the result of positive or negative, respectively, votes for child nodes outnumbering those which are negative or positive, respectively. Thus, in augmented review aggregations, we allow a positive or negative vote for an argument to be seen as a positive or negative vote, respectively, for the argument's parent.

Augmented review aggregation frameworks are well-defined, in the following sense:

Proposition 1. *Any augmented review aggregation framework $\langle \mathcal{O}, \mathcal{R}, \mathcal{V} \rangle$ for x, where $\mathcal{O} = \langle \mathcal{N}, \mathcal{E} \rangle$, is such that for any $\rho \in \mathcal{R}$ and $n_i \in \mathcal{N}$, $\mathcal{V}(\rho, n_i)$ takes exactly one value in $\{-, +\}$ or is undefined.*

Proof. Follows directly from the fact that $\mathcal{O}$ is a tree. □

Proposition 2. *For any review aggregation framework $\langle \mathcal{O}, \mathcal{R}, \mathcal{V} \rangle$ for x, where $\mathcal{O} = \langle \mathcal{N}, \mathcal{E} \rangle$, there exists exactly one augmented review aggregation framework $\langle \mathcal{O}, \mathcal{R}, \mathcal{V}' \rangle$ for x.*

Proof. By Proposition 1, any $n_i \in \mathcal{N}$ takes exactly one value in $\{-, +\}$ or is undefined. Then, it can be seen from Definition 3 that, if n_i is a leaf, whether it takes one value or is undefined is determined by its votes. Further, Definition 3 also shows that if n_i is not a leaf, whether it takes one value or is undefined is determined by its own votes and those on its children. Thus, there can only be one unique augmented review aggregation framework $\langle \mathcal{O}, \mathcal{R}, \mathcal{V}' \rangle$. □

For illustration, consider the reviews in Table 1, and let $\mathcal{R} = \{\rho_1, \rho_2, \rho_3, \rho_4, \rho_5\}$. Then, Table 2 shows possible votes $\mathcal{V}$ in the hotel setting, making use of the nodes in the ontology in Fig. 3, as well as $\mathcal{V}'$ in the augmented review aggregation framework, drawn from

Table 1
Example hotel reviews for the running example.

Review ρ_1	This **hotel** is absolutely beautiful. Our **room** was gorgeous. The **service** was great and the **food** was good as well. I do have one complaint. The hotel is located in a hard to find **location** in Chicago and it takes more than one hour **walking** to reach any tourist attraction.
Review ρ_2	I stayed at this **hotel** when I had to attend a conference in Chicago. I'm truly glad that I will never have to stay here again. The website does a great job of creating an illusion. The hotel is terrible. The **view** is non-existent, all I could see was the bricks from the neighbouring building. The **location** was not great either as there were no close **transport links**. On the bright side, the **internet** was very fast and the **service** was good.
Review ρ_3	I have stayed in **hotels** all over the world, and this is probably the worst that I've ever experienced. The **personnel** were unaccommodating, in particular the guy at the front desk was condescending and not even remotely helpful. If you're here to visit, then the **location** of the hotel is not good, there are no **transport links** nearby.
Review ρ_4	The **room** was not too clean but there are many nice cafes within **walking** distance, the **internet** is fast, and my room had a nice **view**.
Review ρ_5	The **location** is good, **walking** distance to the office I visited. The **internet** was reliable. I found the **room** to be incredibly dusty and the **food** not to my liking.

Table 2
Votes for the ontology in Fig. 3 from the reviews in Table 1 (where + and − indicate a positive and a negative, respectively, vote), as well as augmented votes (where ⊕ and ⊖ indicate a positive and a negative, respectively, augmented vote in the augmented review aggregation framework).

	x_h	f_f	f_r	f_l	f_i	f_s	f_v	f_t	f_w	f_p
ρ_1	+	+	+	−		+			−	
ρ_2	−			−	+	+	−	−		
ρ_3	−			−		⊖		−		−
ρ_4	⊕		−	⊕	+		+		+	
ρ_5		−	−	+	+				+	

$\mathcal{V}$ and from the edges of the ontology. To illustrate the augmentation process here, consider the votes from ρ_2: despite the fact that the review gave more positive than negative votes for the child nodes of x_h, this does not override its negative vote on x_h. However, if we consider ρ_3, we can see that it has not given a vote for f_s but has voted negatively for f_p, and so we augment a negative vote for f_s. Meanwhile, for ρ_4 we can see that a positive vote is augmented for f_l, which then leads to a positive vote being augmented for x_h. After these augmented votes have been added, no more augmentation is possible, e.g. no vote is augmented for x_h from ρ_5 as the number of positive and negative votes for its child nodes are equal, and so the process is complete.

The augmented review aggregation is final in that no further votes can be augmented after the process has been completed once:

Proposition 3. *Let $\langle \mathcal{O}, \mathcal{R}, \mathcal{V} \rangle$ be an augmented review aggregation framework for x, where $\mathcal{O} = \langle \mathcal{N}, \mathcal{E} \rangle$. Then, the augmented review aggregation framework $\langle \mathcal{O}, \mathcal{R}, \mathcal{V}' \rangle$ for x is such that $\mathcal{V}' = \mathcal{V}$.*

Proof. Consider any $n_i \in \mathcal{N}$. If n_i is a leaf, we can see from Definition 3 that if $\mathcal{V}(\rho, n_i)$ is defined, then $\mathcal{V}'(\rho, n_i) = \mathcal{V}(\rho, n_i)$, and if $\mathcal{V}(\rho, n_i)$ is undefined, then $\mathcal{V}'(\rho, n_i)$ is undefined. If n_i is not a leaf, let us assume that $|\{n_j \in \mathcal{N} | \mathcal{E}(n_j) = n_i, \mathcal{V}'(\rho, n_j) = +\}| = |\{n_j \in \mathcal{N} | \mathcal{E}(n_j) = n_i, \mathcal{V}(\rho, n_j) = +\}|$ and $|\{n_k \in \mathcal{N} | \mathcal{E}(n_k) = n_i, \mathcal{V}'(\rho, n_k) = -\}| = |\{n_k \in \mathcal{N} | \mathcal{E}(n_k) = n_i, \mathcal{V}(\rho, n_k) = -\}|$. Then, in all cases in Definition 3, if $\mathcal{V}(\rho, n_i)$ is defined then $\mathcal{V}'(\rho, n_i) = \mathcal{V}(\rho, n_i)$, and if $\mathcal{V}(\rho, n_i)$ is undefined then $\mathcal{V}'(\rho, n_i)$ is undefined. Thus, given the fact that $\mathcal{O}$ is a tree, we can see that the assumption holds and $\mathcal{V}' = \mathcal{V}$. □

In the remainder, unless specified otherwise, we will assume as given a review aggregation framework $\langle \mathcal{O}, \mathcal{R}, \mathcal{V} \rangle$ for an item x, where $\mathcal{O} = \langle \mathcal{N}, \mathcal{E} \rangle$ and $\mathcal{N} = \{x\} \cup \mathcal{F}$, assuming that it is already augmented.

5. ADAs: from review aggregation frameworks to QBAFs

We now define how to obtain QBAFs from (augmented) review aggregation frameworks, with the purpose of representing the information in the reviews for an item argumentatively so to calculate the strength for the item, according to a suitable gradual semantics, and allow explanations to be produced for the "goodness" of the item, in line with its strength, according to the argumentative analysis of the aggregated reviews.

In order to do so, we first introduce our notion of arguments *corresponding to* entities within the review aggregation framework. Here and onward, for any $n_i \in \mathcal{N}$, we let aggregations of the positive and negative votes on n_i be $\mathcal{V}^+(n_i) = |\{\rho \in \mathcal{R} | \mathcal{V}(\rho, n_i) = +\}|$ and $\mathcal{V}^-(n_i) = |\{\rho \in \mathcal{R} | \mathcal{V}(\rho, n_i) = -\}|$, respectively. We note the following, straightforward, result.

Proposition 4. *For any $n_i \in \mathcal{N}$:*

- *if $\mathcal{V}^+(n_i) = |\mathcal{R}|$, then $\mathcal{V}^-(n_i) = 0$;*

- *if* $\mathcal{V}^-(n_i) = |\mathcal{R}|$*, then* $\mathcal{V}^+(n_i) = 0$.

Proof. Follows from the definitions of $\mathcal{V}^+$ and $\mathcal{V}^-$. $\square$

Definition 4. The *set of arguments corresponding to* $\langle \mathcal{O}, \mathcal{R}, \mathcal{V}\rangle$ is $\mathcal{A} = \{\alpha(n_i) | n_i \in \mathcal{N}\}$ for some *argument mapping* α which is a bijection $\alpha : \mathcal{N} \to \mathcal{A}$. We will use *bipolar argument mappings* assigning a (positive or negative) polarity to arguments in $\mathcal{A}$ as follows:

- $\alpha(x)$ has *positive polarity*;
- for any $f_i \in \mathcal{F}$, $\alpha(f_i)$ has *positive polarity* iff $\mathcal{V}^+(f_i) \geq \mathcal{V}^-(f_i)$;
- for any $f_i \in \mathcal{F}$, $\alpha(f_i)$ has *negative polarity* iff $\mathcal{V}^+(f_i) < \mathcal{V}^-(f_i)$.

Note that we see arguments as abstract entities, in the spirit of abstract argumentation [36]. They are matched with elements of the ontology (by α) and can be read succinctly as stating that the matched elements have been positively or negatively reviewed (depending on the polarity), but without explicitly providing evidence from the reviews in the form of premises for the arguments. We leave to future work instantiations of our method to accommodate structured arguments. Note also that the choice of argument mapping is somewhat arbitrary, other than requiring that there is a one-to-one mapping with the item and features in the ontology. Note also that the argument corresponding to the item is always positive so as to align with the many review aggregation methods, e.g. products on e-commerce websites or movies on streaming platforms are reviewed on a single (positive) scale. Furthermore, note that, for the case of $\mathcal{V}^+(f_i) = \mathcal{V}^-(f_i)$, we arbitrarily choose a positive polarity to lean towards positive aggregations.

We will use superscripts + and − to denote positive and negative polarity, respectively, and superscript $*$ for unspecified polarity (amongst + and −). We will also refer to any argument $a_i^+ = \alpha(f_i)$ with positive polarity as a *positive argument* (corresponding to f_i) and to any argument with negative polarity $a_i^- = \alpha(f_i)$ as a *negative argument (corresponding to* f_i*)*. Intuitively, a positive argument represents the reviews being generally positive on the item or one of its features (that to which the argument corresponds), while a negative argument represents that the reviews are generally negative.[7]

For illustration, in the running hotel example, the argument corresponding to x_h is a_h^+, i.e. a positive argument for the hotel (as is always the case for items), the argument corresponding to f_v is a_v^+, i.e. a positive argument for the view since, as per Table 2, $\mathcal{V}^+(f_v) \geq \mathcal{V}^-(f_v)$, and the argument corresponding to f_l is a_l^-, i.e. a negative argument for the location since $\mathcal{V}^+(f_l) < \mathcal{V}^-(f_l)$.

In order to formulate QBAFs, in addition to the arguments we must also determine their base scores and attack and support relations. We use votes to define base scores and argument polarity for defining the relations, as follows:

Definition 5. Let $\mathcal{F} = \{f_1, \ldots, f_n\}$. Then, the *QBAF corresponding to* $\langle \mathcal{O}, \mathcal{R}, \mathcal{V}\rangle$ is $\langle \mathcal{A}, \mathcal{L}^-, \mathcal{L}^+, \tau\rangle$ where:

- $\mathcal{A} = \{\alpha(x), \alpha(f_1), \ldots, \alpha(f_n)\} = \{a_0^+, a_1^*, \ldots, a_n^*\}$ is the set of arguments corresponding to $\langle \mathcal{O}, \mathcal{R}, \mathcal{V}\rangle$ (for some α);
- $\mathcal{L}^- \subset \mathcal{A} \times \mathcal{A}$ is such that $\forall n_i, n_j \in \mathcal{N}$, where $a_i^* = \alpha(n_i)$ and $a_j^* = \alpha(n_j)$, $(a_i^*, a_j^*) \in \mathcal{L}^-$ iff $n_j = \mathcal{E}(n_i)$ and
  - $a_i^* = a_i^+$ and $a_j^* = a_j^-$, or
  - $a_i^* = a_i^-$ and $a_j^* = a_j^+$;
- $\mathcal{L}^+ \subset \mathcal{A} \times \mathcal{A}$ is such that $\forall n_i, n_j \in \mathcal{N}$, where $a_i^* = \alpha(n_i)$ and $a_j^* = \alpha(n_j)$, $(a_i^*, a_j^*) \in \mathcal{L}^+$ iff $n_j = \mathcal{E}(n_i)$ and
  - $a_i^* = a_i^+$ and $a_j^* = a_j^+$, or
  - $a_i^* = a_i^-$ and $a_j^* = a_j^-$;
- $\tau(a_0^+) = 0.5 + 0.5 \cdot \frac{\mathcal{V}^+(a_0^+) - \mathcal{V}^-(a_0^+)}{|\mathcal{R}|}$ and, $\forall i \in \{1, \ldots, n\}$, $\tau(a_i^*) = \frac{|\mathcal{V}^+(f_i) - \mathcal{V}^-(f_i)|}{|\mathcal{R}|}$.

Proposition 5. *Given any QBAF* $\langle \mathcal{A}, \mathcal{L}^-, \mathcal{L}^+, \tau\rangle$ *corresponding to* $\langle \mathcal{O}, \mathcal{R}, \mathcal{V}\rangle$*, for any* $a_i \in \mathcal{A}$*,* $\tau(a_i) \in [0, 1]$.

Proof. Follows directly from Definition 5. $\square$

Note that the base score of the argument a_0^+ corresponding to the item has been adapted from [74], where several useful properties thereof are shown. Intuitively, $\tau(a_0^+) = 1$ represents all reviews voting positively for the item while $\tau(a_0^+) = 0$ requires universally negative votes. Note also that even when all reviews vote negatively towards an item, the corresponding argument is still positive (with a significantly weakened base score), in order to align with industry standards, e.g. on Trip Advisor, Rotten Tomatoes and Amazon. This differs from arguments corresponding to features, which may be positive or negative depending on the majority of the votes. Here, the base score of an argument a_i^* corresponding to a feature f_i is such that $\tau(a_i^*) = 0$ requires equal positive and negative votes from the reviews on f_i while $\tau(a_i^*) = 1$ requires universally positive (negative) votes on f_i if $a_i^* = a_i^+$ ($a_i^* = a_i^-$, respectively).

As regards the definition of the attack and support relations, we impose that there is some form of disagreement between arguments for an attack to be present and only agreement for a support to be present. Thus, an attack is defined as being between a negative argument and a positive argument, while a support is between two positive arguments or two negative arguments. Note also that we exclude that $\mathcal{L}^- = \mathcal{A} \times \mathcal{A}$ and $\mathcal{L}^+ = \mathcal{A} \times \mathcal{A}$ as the QBAF reflects the graphical structure of the ontology and thus, in particular, the

[7] Note that we depart here from the original formulation in [30] in that positive and negative arguments are not defined explicitly in [30].

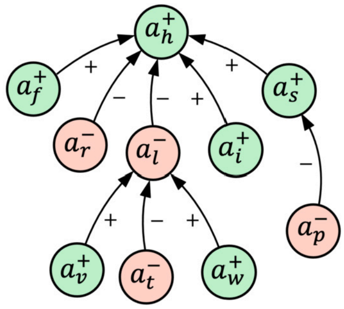


Fig. 4. Graphical representation of the arguments (positive and negative arguments in green and red, respectively), attack relation (labelled −) and support relation (labelled +) for the QBAF in the hotel example.

Table 3
Base scores (τ) and strengths by the QuAD (σ_q), DF-QuAD (σ_d) and REB (σ_r) gradual semantics for the hotel example.

	a_h^+	a_f^+	a_r^-	a_l^-	a_i^+	a_s^+	a_v^+	a_t^-	a_w^+	a_p^-
τ	0.5	0	0.2	0.2	0.8	0.2	0	0.6	0.4	0.2
σ_q	0.578	0	0.2	0.4	0.8	0.16	0	0.6	0.4	0.2
σ_d	0.672	0	0.2	0.36	0.8	0.16	0	0.6	0.4	0.2
σ_r	0.598	0	0.2	0.228	0.8	0.175	0	0.6	0.4	0.2

argument corresponding to the item can never attack or support any arguments. For the arguments' base scores we use an intuitive aggregation of reviews' votes, which differs for arguments corresponding to the item or otherwise, since the former is always a positive argument, whereas the latter may also be negative arguments.

A graphical representation of the QBAF for the running hotel example is shown in Fig. 4 and base scores for the arguments are given in Table 3 (the strengths given by each gradual semantics we consider, shown here, will be discussed in the next section).

In the remainder, unless specified otherwise, we will use $\langle \mathcal{A}, \mathcal{L}^-, \mathcal{L}^+, \tau \rangle$ to indicate the QBAF corresponding to $\langle \mathcal{O}, \mathcal{R}, \mathcal{V} \rangle$ as per Definition 5.

We posit that Lemmas 6–9 below demonstrate the intuitive behaviour of the base scores and the relevance and clarity of the QBAFs in this setting.

The first lemma considers the cases where the maximum, minimum and midpoint are returned as base scores for the argument corresponding to the item. They occur when the votes from reviews are universally positive, universally negative and equally balanced, respectively, which we believe is intuitive here, given that the base score can be considered as a "starting point" for the strength function.

Proposition 6. *The following statements hold:*

- $\tau(a_0^+) = 1$ *iff* $\mathcal{V}^+(x) = |\mathcal{R}|$ *(and thus* $\mathcal{V}^-(x) = 0$*);*
- $\tau(a_0^+) = 0$ *iff* $\mathcal{V}^-(x) = |\mathcal{R}|$ *(and thus* $\mathcal{V}^+(x) = 0$*);*
- $\tau(a_0^+) = 0.5$ *iff* $\mathcal{V}^+(x) = \mathcal{V}^-(x)$.

Proof. Follows directly from Definition 5 and Lemma 4. □

Differently, for the arguments corresponding to the features, the maximum base score occurs when the votes from reviews are universally positive for positive arguments and universally negative for negative arguments, while the minimum base score occurs when they are equally balanced, which, again, we deem intuitive given the role of the base score in the gradual semantics.

Proposition 7. *For any* $a_i^* \in \mathcal{A}$ *with* $a_i^* = \alpha(f_i)$, $f_i \in \mathcal{F}$*:*

- $\tau(a_i^*) = 1$ *iff:*
  - $* = +$ *and* $\mathcal{V}^+(f_i) = |\mathcal{R}|$ *(and thus* $\mathcal{V}^-(f_i) = 0$*), or*
  - $* = -$ *and* $\mathcal{V}^-(f_i) = |\mathcal{R}|$ *(and thus* $\mathcal{V}^+(f_i) = 0$*);*
- $\tau(a_j^*) = 0$ *iff* $\mathcal{V}^+(f_i) = \mathcal{V}^-(f_i)$.

Proof. Follows directly from Definition 5 and Lemma 4. □

The next proposition implies that attacks and supports are mutually exclusive, that all arguments have a path to the argument corresponding to the item and that the QBAF is acyclic. The first means that there is no ambiguity between how one argument affects another, while the second enforces that all arguments are relevant to the item. Meanwhile, the last is due to the fact that the extracted argumentative reasoning follows the parent-child relation $\mathcal{E}$ from Definition 1, a core tenet of our approach, which, we argue, allows for intuitive modes of interaction between the item and its (hierarchical) features, as we will demonstrate in Section 7. We leave to future work the study of different forms of argumentation framework, e.g. QBAFs similar to ours but allowing for relations between sibling arguments, as in [12], or less hierarchical, such as cyclic, representations, as in [23].

Proposition 8. *The following statements hold:*

- $\mathcal{L}^- \cap \mathcal{L}^+ = \emptyset$;
- *for any* $a_i^* \in \mathcal{A}$ *with* $a_i^* = \alpha(f_i)$, $f_i \in \mathcal{F}$, *there exists a path from* a_i^* *to* a_0^+ *via* $\mathcal{L}^- \cup \mathcal{L}^+$;
- *for any* $a_i^* \in \mathcal{A}$, *there does not exist a path from* a_i^* *to* a_i^* *via* $\mathcal{L}^- \cup \mathcal{L}^+$.

Proof. Follows from Definition 5 and the fact that $\langle \mathcal{A}, \mathcal{L}^-, \mathcal{L}^+, \tau \rangle$ is a tree. □

The final lemma considers the case where no votes are present, ensuring that: the arguments are assigned the base scores corresponding to the neutral points (as per Lemmas 6 and 7); and no attacks are present. In this situation, if the base score is considered as a "starting point" for the strength function, arguments will affect one another minimally, as we would hope for with no votes from the reviews.

Proposition 9. *If* $\forall n_i \in \mathcal{N}$, $\forall \rho \in \mathcal{R}$, $\mathcal{V}(\rho, n_i)$ *is undefined, then:*

- $\tau(a_0^+) = 0.5$ *and* $\forall a_j^* \in \mathcal{A} \setminus \{a_0^+\}$, $\tau(a_j^*) = 0$;
- $\mathcal{L}^- = \emptyset$;
- $\forall n_k, n_l \in \mathcal{N}$, *where* $a_k^* = \alpha(n_k)$, $a_l^* = \alpha(n_l)$ *and* $n_l = \mathcal{E}(n_k)$, $(a_k^*, a_l^*) \in \mathcal{L}^+$.

Proof. Follows directly from Definitions 4 and 5. □

Once QBAFs have been obtained, gradual semantics can be applied to calculate the strength for item and features. In an ADA, the item's strength (which depends on that of the features in turn) serves as both a quantitative evaluation of the item (see the next section), resulting from the review aggregation, and to support the generation of dialogical explanations (see Section 7) for the item's evaluation based on this review aggregation.

6. ADAs: from QBAFs to item evaluation

In general, the choice of argumentative semantics for an application is based on the desirable behaviour it should exhibit in the context of that application, which can be defined in the form of *properties* the semantics should satisfy. In this section, we examine how the three different gradual semantics for QBAFs detailed in Section 3 (QuAD, DF-QuAD and REB) behave in the context of review aggregation, assessing theoretically how well the semantics satisfy a number of properties, while arguing in what sense they are desirable in our setting. We select these three semantics as they satisfy different combinations of these properties, but note that we could have used other semantics, e.g. that introduced in [68]; we leave the analysis of other semantics to future work. We use this analysis to select the most appropriate gradual semantics, amongst the three we consider, for the experimental analysis in the case studies (Section 8).

We begin with an illustration, demonstrating how the three gradual semantics can be used in the hotel example to compute the strength of argument a_h^+ in order to evaluate the item x_h (see Table 3).[8]

- For QuAD, we first aggregate the strengths of the attackers, obtaining $\mathcal{F}_a(0.5, 0.4, 0.2) = \mathcal{F}_a(f_a(0.5, 0.4), 0.2) = \mathcal{F}_a(0.5 \cdot (1 - 0.4), 0.2) = \mathcal{F}_a(0.3, 0.2) = f_a(0.3, 0.2) = 0.3 \cdot (1 - 0.2) = 0.24$. Similarly, we then aggregate the strengths of the supporters, giving $\mathcal{F}_s(0.5, 0.8, 0, 0.16) = \mathcal{F}_s(f_s(0.5, 0.8), 0, 0.16) = \mathcal{F}_s(0.5 + 0.8 - 0.5 \cdot 0.8, 0, 0.16) = \mathcal{F}_s(0.9, 0, 0.16) = \mathcal{F}_s(f_s(0.9, 0), 0.16) = \mathcal{F}_s(0.9 + 0 - 0.9 \cdot 0, 0.16) = \mathcal{F}_s(0.9, 0.16) = f_s(0.9, 0.16) = 0.9 + 0.16 - 0.9 \cdot 0.16 = 0.916$. Finally, we combine the aggregated strengths of the attackers and supporters with the base score to obtain the overall strength $\sigma_q(a_h^+) = g(0.5, 0.24, 0.916) = \frac{0.24+0.916}{2} = 0.578$.
- For DF-QuAD, we first calculate the aggregated strengths of a_h^+'s attackers $v_a = \mathcal{F}(0.36, 0.2) = 1 - (|1 - 0.36|) \cdot (|1 - 0.2|) = 1 - 0.64 \cdot 0.8 = 0.488$ and of a_h^+'s supporters $v_s = \mathcal{F}(0.8, 0, 0.16) = 1 - (|1 - 0.8|) \cdot (|1 - 0|) \cdot (|1 - 0.16|) = 1 - 0.2 \cdot 1 \cdot 0.84 = 0.832$. Then, as in QuAD, we combine the aggregated strengths of the attackers and supporters with the base score to obtain the overall strength $\sigma_d(a_h^+) = \mathcal{C}(0.5, 0.488, 0.832) = 0.5 + ((1 - 0.5) \cdot |0.832 - 0.488|) = 0.672$.
- For REB, we first calculate the exponent $E = 0.8 + 0 + 0.175 - 0.228 - 0.2 = 0.547$. Then, this is combined with the base score to obtain the overall strength $\sigma_r(a_h^+) = 1 - \frac{1-0.5^2}{1+0.5 \cdot e^{0.547}} = 0.598$.

[8] Note that the arguments' polarities make no difference to the semantics' calculations.

As this example shows, different gradual semantics may give rise to different item evaluations drawn from exactly the same reviews. Thus it is important, towards deciding which semantics to adopt, to identify desirable properties for the review aggregation context guaranteed by semantics.

In the remainder of this section, we adapt to our context the properties of *(strict) balance* and *(strict) monotonicity* given in [10] and assess how they may or may not be useful for review aggregation. We also give a novel property (Property 3 below) which expresses a desirable requirement for gradual semantics of QBAFs both in the context of review aggregation and elsewhere. We define the properties w.r.t. a generic gradual semantics σ. Then, we use the properties to undertake a theoretical analysis of the suitability of the three aforementioned semantics in our setting. Note that, although strictly speaking we only define one novel property in this section, compared to [10], our instantiation for the existing properties is novel, as is the proof of property satisfaction/violation for the QBAFs specific to our review aggregation context.

In [10], a wide range of properties from the literature are shown to be implied by, or be instances of, a small set of parameterised properties. Thus, for these properties to be deployed here, they need to be suitably instantiated w.r.t. the configurable parameters in [10]. To this end, we will use the notation $\mathcal{L}^-_\emptyset = \{(a^*_i, a^*_j) \in \mathcal{L}^- | \sigma(a^*_i) \neq 0\}$ and $\mathcal{L}^+_\emptyset = \{(a^*_i, a^*_j) \in \mathcal{L}^+ | \sigma(a^*_i) \neq 0\}$ to represent the attackers and supporters, respectively, with a non-zero strength. We call arguments with a zero (non-zero) strength *rejected* (*non-rejected*, respectively) arguments: we make this distinction because we believe that the former should not have an effect on the arguments they attack or support, respectively. For illustration, in the hotel example we posit that it is intuitive that, since the reviews of the feature "view" were neutral overall, its corresponding positive argument a^+_v is rejected (as is the case for all considered semantics, see Table 3) and has no effect on the strength of the argument a^-_l, the negative argument concerning the location.

For the second choice of parameters in [10], i.e. that of $<<$, governing the strictness of the inequalities used to compare arguments' base scores and strengths, we select $<<=<_\top$, where for $m, n \in [0,1]$, $m <_\top n$ denotes $m < n$ or $m = n = 1$ or $m = n = 0$, since this is the strictest setting which differentiates the semantics we consider. It also makes sense in our setting of review aggregation, where there are clear maximum and minimum values. We also make use of the comparison measures used to define the properties in [10], whereby, for $A, B \subseteq \mathcal{A}$: A is as strong as B w.r.t. σ, denoted $A =_\sigma B$, iff $\sigma(A) = \sigma(B)$, where we refer to the multiset of the strengths of arguments in any $X \subseteq \mathcal{A}$, i.e. $\{\sigma(a^*_i) | a^*_i \in X\}$, as $\sigma(X)$; A is at least as strong as B w.r.t. σ, denoted $A \geq_\sigma B$, iff there exists an injective mapping f from $\sigma(B)$ to $\sigma(A)$ such that $\forall a_i \in B$, $\sigma(f(a_i)) \geq \sigma(a_i)$; and A is stronger than B w.r.t. σ, denoted $A >_\sigma B$, iff $A \geq_\sigma B$ and $B \not\geq_\sigma A$.

The first two properties, balance and strict balance [10], state that an imbalance between the strengths of an argument's attackers and supporters must correspond to a difference in its strength and its base score. They can be formulated as follows in our setting:

Property 1. *A strength function σ is:*

- balanced *iff for any $a^*_i \in \mathcal{A}$:*
  1. *If $\mathcal{L}^-_\emptyset(a^*_i) =_\sigma \mathcal{L}^+_\emptyset(a^*_i)$ then $\sigma(a^*_i) = \tau(a^*_i)$.*
  2. *If $\mathcal{L}^-_\emptyset(a^*_i) >_\sigma \mathcal{L}^+_\emptyset(a^*_i)$ then $\sigma(a^*_i) \leq \tau(a^*_i)$.*
  3. *If $\mathcal{L}^-_\emptyset(a^*_i) <_\sigma \mathcal{L}^+_\emptyset(a^*_i)$ then $\sigma(a^*_i) \geq \tau(a^*_i)$.*
- strictly balanced *iff σ is balanced and for any $a^*_i \in \mathcal{A}$:*
  4. *If $\sigma(a^*_i) < \tau(a^*_i)$ then $\mathcal{L}^-_\emptyset(a^*_i) >_\sigma \mathcal{L}^+_\emptyset(a^*_i)$.*
  5. *If $\sigma(a^*_i) > \tau(a^*_i)$ then $\mathcal{L}^-_\emptyset(a^*_i) <_\sigma \mathcal{L}^+_\emptyset(a^*_i)$.*

We believe that balance represents desirable behaviour for review aggregation because the base score gives the intrinsic strength of an argument, and thus the item or feature it corresponds to, *before* it is affected by attackers and supporters, i.e. its child features, so it makes sense to take this as the argument's strength if none are non-rejected. This means that for our evaluation of each item or feature, we start from an aggregation of the votes on that item or feature, respectively, before we consider the argumentative reasoning on its children. We then posit that it is natural to assume that attackers (supporters), i.e. more specific child features which disagree (agree, respectively) on the polarity of the more general parent item or feature do not increase (reduce, respectively) the strength w.r.t. the base score of their parent, given the nature of these relations and how they are obtained as per Definition 5 to represent disagreement (agreement, respectively). For illustration, in the hotel example we would expect that $\sigma(a^+_v) = \tau(a^+_v) = 0$ since the base score is the only information available to us and there is no argumentative reasoning for or against the argument a^+_v. Meanwhile, we believe it is only natural for a user to expect that the strength of a^+_s is not greater than its base score, i.e. $\sigma(a^+_s) \leq \tau(a^+_s)$, since the only argumentative reasoning affecting a^+_s is the attack from a^-_p.

However, strict balance states that any difference in the base score and strength of an argument must correspond to the specific form of *dominance* between the sets of attackers and supporters characterised by $<_\sigma$. For example, given an argument with two attackers, both with strength 0.1, and a single supporter with strength 0.9, strict balance prevents the possibility that the argument has a strength higher than its base score (see [10] for more details). In this setting, we do not necessarily want this to be the case, since an argument concerning a very positive feature supporting an item should be able to "overpower" two only slightly negative features which attack it, for example. (Indeed, in a way strict balance somewhat mimics the discussed issues with the TS score.) For illustration, in the hotel example, suppose we were to keep adding attackers to a^+_0, i.e. features with mostly negative reviews, each of which has a strength lower than the strength of the strongest supporter, $\sigma(a^+_l) = 0.8$ (by all semantics). In this case, strict balance does not allow that a^+_0's strength ever becomes lower than its base score, since one of their strengths must be at least as strong as the strongest supporter for it to be the case that $\mathcal{L}^{-}_{\emptyset}(a^*_i) >_\sigma \mathcal{L}^{+}_{\emptyset}(a^*_i)$. For this reason, we believe that this is not a desirable property in this setting.

The further two properties, of monotonicity and strict monotonicity, require that the strength of an argument depends monotonically on its base score and on the strengths of its attackers and supporters, so that their strengthening/weakening will do likewise to the argument. Before defining these properties, following [10], we let, for any $a^*_i \in \mathcal{A}$, the *shaping triple* of a^*_i be $\langle \tau(a^*_i), \mathcal{L}^+(a^*_i), \mathcal{L}^-(a^*_i)\rangle$, denoted $\mathcal{ST}(a^*_i)$. Then, given $a^*_i, a^*_j \in \mathcal{A}$, $\mathcal{ST}(a^*_j)$ is said to be:

- *as boosting as* $\mathcal{ST}(a^*_i)$, denoted $\mathcal{ST}(a^*_i) \simeq \mathcal{ST}(a^*_j)$, iff $\tau(a^*_i) = \tau(a^*_j)$, $\mathcal{L}^+_{\emptyset}(a^*_i) =_\sigma \mathcal{L}^+_{\emptyset}(a^*_j)$, and $\mathcal{L}^-_{\emptyset}(a^*_i) =_\sigma \mathcal{L}^-_{\emptyset}(a^*_j)$;
- *at least as boosting as* $\mathcal{ST}(a^*_i)$, denoted $\mathcal{ST}(a^*_i) \preceq \mathcal{ST}(a^*_j)$, iff $\tau(a^*_i) \leq \tau(a^*_j)$, $\mathcal{L}^+(a^*_i) \leq_\sigma \mathcal{L}^+(a^*_j)$, and $\mathcal{L}^-(a^*_j) \leq_\sigma \mathcal{L}^-(a^*_i)$.
- *strictly more boosting than* $\mathcal{ST}(a^*_i)$, denoted $\mathcal{ST}(a^*_i) \prec \mathcal{ST}(a^*_j)$, iff $\mathcal{ST}(a^*_i) \preceq \mathcal{ST}(a^*_j)$ and $\mathcal{ST}(a^*_j) \npreceq \mathcal{ST}(a^*_i)$.

Property 2. *A strength function σ is:*

- *monotonic (w.r.t. boosting) iff for any $a^*_i, a^*_j \in \mathcal{A}$, if $\mathcal{ST}(a^*_i) \simeq \mathcal{ST}(a^*_j)$ then $\sigma(a^*_i) = \sigma(a^*_j)$ and if $\mathcal{ST}(a^*_i) \preceq \mathcal{ST}(a^*_j)$ then $\sigma(a^*_i) \leq \sigma(a^*_j)$;*
- *strictly monotonic (w.r.t. boosting) iff σ is monotonic and for any $a^*_i, a^*_j \in \mathcal{A}$, if $\mathcal{ST}(a^*_i) \prec \mathcal{ST}(a^*_j)$ then $\sigma(a^*_i) <_\top \sigma(a^*_j)$.*

Monotonicity makes sense in our setting as one would expect that, all being equal, two arguments (item or features) are evaluated equally, which is an important characteristic for ensuring fairness amongst factors contributing to review aggregations. Further, when one shaping triple is more boosting than the other, we also posit that the prescribed behaviour is intuitive in our setting due to the nature of attackers and supporters as per Definition 5. Specifically, we would like an increase in the number of more specific child features which disagree (agree), in polarity, with their more general parent item or feature, to be restricted from increasing (decreasing, respectively) the strength of the latter. Also, this behaviour requires that, all else being equal, increasing the proportion of positive to negative reviews on an item or feature cannot reduce its overall strength. Note that we do not impose that the strength is necessarily increased or decreased here.

If we consider a^-_r and a^-_p in the hotel example: both have equivalent non-rejected attackers, non-rejected supporters and base scores, i.e. $\mathcal{L}^-_{\emptyset}(a^-_r) = \mathcal{L}^-_{\emptyset}(a^-_p)$ and $\mathcal{L}^+_{\emptyset}(a^-_r) = \mathcal{L}^+_{\emptyset}(a^-_p)$, and $\tau(a^-_r) = \tau(a^-_p)$ and so we would naturally expect that $\sigma(a^-_r) = \sigma(a^-_p)$. However, if we were to increase the number of positive votes on the personnel (represented by a^-_p) with all else remaining equal, we believe it is intuitive that $\sigma(a^-_r) \leq \sigma(a^-_p)$. Further, we would expect that adding the attacking argument a^+_w against a^-_l, i.e. introducing a new child feature to location, on which the reviews were mostly negative, to obtain the QBAF in Fig. 4 from an identical QBAF except for the absence of a^+_w, cannot increase the strength of a^-_l.

Meanwhile, strict monotonicity imposes a strict increase or decrease in the strength in the case of one shaping triple being more boosting than another, so long as the arguments' strengths are not already at the top or bottom values. We believe that this is also an intuitive property in the setting of review aggregation, given that it gives stronger guarantees on the effects of positive or negative votes as they propagate up the tree.

The final property is novel and defines the *attainability* of a semantics, which equates to: all strength values are attainable for any given base score with a certain set of attackers or supporters.

Property 3. *A strength function σ is* attainable *iff for any $a^*_i \in \mathcal{A}$, $\forall v \in [0,1]$, $\exists S, T \in [0,1]^*$ such that if $\sigma(\mathcal{L}^-(a^*_i)) = S$ and $\sigma(\mathcal{L}^+(a^*_i)) = T$, then $\sigma(a^*_i) = v$.*

We believe that this property is intuitive in our context since users would naturally expect that all values in a review scale are attainable in a review aggregation (given sufficient attackers/supporters). For illustration, in the hotel example we would expect that there is some combination of a^+_h's attackers' strengths, $\sigma(\mathcal{L}^-(a^+_h)) \in [0,1]^*$, and supporters' strengths, $\sigma(\mathcal{L}^+(a^+_h)) \in [0,1]^*$, such that any specified $\sigma(a^+_h) \in [0,1]$ can be achieved.

We conclude by assessing the behaviour of the specific semantics we consider as concerns the properties defined earlier. Table 4 summarises the results.

Proposition 10. *σ_q satisfies monotonicity and attainability but violates balance, strict balance and strict monotonicity.*

Proof. The proofs and counterexamples for balance, strict balance, monotonicity, strict monotonicity can be found in [10]. For the proof of attainability, we have three cases: i.) $\tau(a^*_i) = v$, ii.) $\tau(a^*_i) < v$ and iii.) $\tau(a^*_i) > v$. For case i, we let $\mathcal{L}^-(a^*_i) = \mathcal{L}^+(a^*_i) = \emptyset$ and then, by the definition of QuAD (see Section 3), $\sigma_q(a^*_i) = \tau(a^*_i) = v$. For case ii, we let $\mathcal{L}^-(a^*_i) = \emptyset$ and $\mathcal{L}^+(a^*_i) = \{a^*_j\}$ where $\sigma_q(a^*_j) = \frac{v-\tau(a^*_i)}{1-\tau(a^*_i)}$. Then, it can be seen from the definition of QuAD that $\sigma_q(a^*_i) = v$. Analogously for case iii, we let $\mathcal{L}^+(a^*_i) = \emptyset$ and $\mathcal{L}^-(a^*_i) = \{a^*_j\}$ where $\sigma_q(a^*_j) = \frac{\tau(a^*_i)-v}{\tau(a^*_i)}$. Then, it can be seen from the definition of QuAD that $\sigma_q(a^*_i) = v$. Thus, $\sigma_q(a^*_i) = v$ in all cases and σ_q satisfies attainability. $\square$

QuAD does not satisfy balance, indicating that it may not be suitable in some contexts, where it is desirable that attackers do not increase and supporters do not decrease an argument's strength *starting from* the base score. We believe that this is a weakness

Table 4
Gradual semantics and their satisfaction (✓) or violation (·) of the assessed properties.

Semantics	Property		
	(Strict) Balance	(Strict) Monotonicity	Attainability
QuAD	· (·)	✓ (·)	✓
DF-QuAD	✓ (·)	✓ (·)	✓
REB	✓ (·)	✓ (✓)	·

of QuAD in the review aggregation context. The fact that it does not satisfy strict balance is not a concern, however, since we have established that this property is not desirable here.

Proposition 11. *σ_d satisfies balance, monotonicity and attainability but violates strict balance and strict monotonicity.*

Proof. The proofs and counterexamples for balance, strict balance, monotonicity, strict monotonicity can be found in [10]. For the proof of attainability, the proof is the same as that for σ_q. We have three cases: i.) $\tau(a_i^*) = v$, ii.) $\tau(a_i^*) < v$ and iii.) $\tau(a_i^*) > v$. For case i, we let $\mathcal{L}^-(a_i^*) = \mathcal{L}^+(a_i^*) = \emptyset$ and then, by the definition of DF-QuAD (see Section 3), $\sigma_d(a_i^*) = \tau(a_i^*) = v$. For case ii, we let $\mathcal{L}^-(a_i^*) = \emptyset$ and $\mathcal{L}^+(a_i^*) = \{a_j^*\}$ where $\sigma_d(a_j^*) = \frac{v-\tau(a_i^*)}{1-\tau(a_i^*)}$. Then, it can be seen from the definition of DF-QuAD that $\sigma_d(a_i^*) = v$. Analogously for case iii, we let $\mathcal{L}^+(a_i^*) = \emptyset$ and $\mathcal{L}^-(a_i^*) = \{a_j^*\}$ where $\sigma_d(a_j^*) = \frac{\tau(a_i^*)-v}{\tau(a_i^*)}$. Then, it can be seen from the definition of DF-QuAD that $\sigma_d(a_i^*) = v$. Thus, $\sigma_d(a_i^*) = v$ in all cases and σ_d satisfies attainability. □

Thus, DF-QuAD satisfies the same properties as QuAD but adds balance, meaning its behaviour when comparing an argument's strength to its base score may be more intuitive than that of QuAD. It thus seems to be a particularly well-suited semantics for review aggregation. However, both QuAD and DF-QuAD suffer from the disadvantage of not satisfying strict monotonicity, which is desirable here.

Proposition 12. *σ_r satisfies balance, monotonicity, strict monotonicity and attainability but violates strict balance and attainability.*

Proof. The proofs and counterexamples for balance, strict balance, monotonicity, strict monotonicity can be found in [10]. For the counterexample for attainability, let $\tau(a_i^*) = 0.5$ and $v = 0.1$. We then express the definition of REB (see Section 3) in terms of E, i.e. $v = \sigma_r(a_i^*) = 1 - \frac{1-\tau(a_i^*)^2}{1+\tau(a_i^*)\cdot e^E}$ becomes $E = ln(\frac{1.5}{1-v} - 2) = ln(\frac{1.5}{1-0.1} - 2) = ln(-0.33)$. As such, there does not exist a value of $E = \sum_{a_j^* \in \mathcal{L}^+(a_i^*)} \sigma_r(a_j^*) - \sum_{a_k^* \in \mathcal{L}^-(a_i^*)} \sigma_r(a_k^*)$, and thus a set of attackers and supporters, such that $\sigma(a_i^*) = v$. □

Thus, REB adds further guarantees on these effects from modifying the shaping triple, since it satisfies strict monotonicity. However, the fact that REB does not satisfy attainability highlights a problem in that there is a *blind spot* in its strength results. If we express the algorithm in the form $y = 1 - \frac{1-\tau(a_i)^2}{1+\tau(a_i)\cdot e^x}$ we can see that as x approaches ∞ (the balance of reasoning moves in favour of support), it holds that $y = 1$, while as x approaches $-\infty$ (the balance of reasoning moves in favour of attack), it holds that $y = \tau(a_i)^2$, i.e. attackers cannot weaken the argument further than the square of its base score. This implies that as the votes on an item/feature become stronger, the corresponding argument becomes increasingly resistant to the weakening effect of negative reasoning. This also causes an asymmetry between attackers and supporters, inhibiting the attackers' effect compared to the supporters', which is not intuitive in our context.

While we acknowledge that the lack of strict monotonicity in DF-QuAD could be seen as a weakness here, we believe that its satisfaction of attainability makes it better suited than QuAD and REB for review aggregation. We will therefore use DF-QuAD in the remainder, and use σ to stand for σ_d.

7. ADAs: from item evaluation to dialogical explanations

The problem of defining explanations for AI models is widely studied within the explainable AI (XAI) landscape. Many XAI approaches provide static explanations for outputs of AI models in terms of their inputs, e.g. as in feature attribution methods such as LIME [77] or SHAP [54], or as in counterfactual explanations such as in [92]. However, there is a growing trend in XAI towards human-oriented explanations, favouring, in particular, contrastive, selective and social explanations [62] and explanations as model reconciliation between AI models and humans [82]. Dialogical explanations drawn from argumentative structures have recently been advocated as fitting within this trend, including for conflict-resolution [73], for model reconciliation [90], and for explaining image classification [49]. We subscribe to this trend and propose, in this section, dialogical explanations as a natural form of explanation for the review aggregation process underpinning ADAs, drawn from its argumentative backbone. Concretely, we demonstrate some ways in which ADAs can use the QBAFs, extracted from the review aggregation process, and their gradual semantics, to generate dialogical explanations for item recommendations drawn from the reviews, prompted by natural questions by human users. Rather than restricting attention to one particular form of dialogical explanation or another, we explore some options, leaving their empirical analysis in concrete settings to future user studies. We continue to illustrate throughout the section with the hotel example.

First, we loosely define a dialogical explanation as a finite, non-empty sequence of pairs (each containing a request from the user and a response from the ADA) as follows.

Definition 6. A *dialogical explanation for the evaluation of an item* x between an ADA and a user is a finite sequence (for $\delta \in \mathbb{N}$):

$$\langle \mathcal{S}(x), (\mathcal{Q}_1, \mathcal{X}_1), \dots, (\mathcal{Q}_\delta, \mathcal{X}_\delta) \rangle$$

where $\mathcal{S}(x)$ is a statement by the ADA about the evaluation of x and, for all $i \in \{1, \dots, \delta\}$, $\mathcal{Q}_i$ is an *explanation request* from the user *concerning some* $a_k^* \in \mathcal{A}$ and $\mathcal{X}_i$ is an *explanation response* by the ADA *including some* excerpts from $\mathcal{R}$ or some $R \subseteq \mathcal{A}$ such that, for each $a_l^* \in R$, $(a_l^*, a_k^*) \in \mathcal{L}^- \cup \mathcal{L}^+$ or $(a_k^*, a_l^*) \in \mathcal{L}^- \cup \mathcal{L}^+$.

Thus, we see a dialogical explanation as starting with a statement by the ADA, followed by an interleaving of questions by the user (the explanation requests) and answers by the ADA (the explanation responses). These dialogical explanations are naturally *social* [62], as they are conceived as exchanges between the ADA and human users. By focusing on just some arguments in $\mathcal{A}$ (given by R), the dialogical explanations can be *selective* [62], but in such way that the responses are always *relevant* to the requests by leveraging on (sets of) children or parents of the node concerning the requests. By allowing responses to include attacks (in $\mathcal{L}^-$) the dialogical explanations can be *contrastive* [62]. Note that we choose not to impose further restrictions on the notion of dialogical explanation in order to allow for different instances to be defined where different contexts have different dialogical requirements. We also choose not to impose any restrictions on the format of explanation requests and responses. They could be in natural language, as in the following simple illustration of a dialogical explanation for the hotel example:

ADA: *The hotel was highly rated, I would recommend it.*
User: *Why was the hotel highly rated?*
ADA: *The hotel was highly rated because the internet was great, despite the fact that the location was poor.*
User: *Why was the location poorly rated?*
ADA: *The location was poorly rated because the public transport was poor, despite the fact that the walking routes were great.*
User: *Why was the public transport poorly rated?*
ADA: *A reviewer stated that "there were no close transport links".*

This dialogical explanation follows a *top-down* exploration of the QBAF. The following illustration follows a different, random exploration with *downward* reasoning in the responses:

ADA: *The hotel was highly rated, I would recommend it.*
User: *What did reviewers say about the service?*
ADA: *The service was highly rated, with one reviewer stating that "the service was great", despite the fact that the personnel were poorly rated, with one reviewer stating that "the personnel were unaccommodating, in particular the guy at the front desk was condescending and not even remotely helpful".*
User: *What did reviewers say about the internet?*
ADA: *The internet was highly rated, with one reviewer stated that "the internet was very fast" and another stating that "the internet was very reliable".*

The following further illustration adopts a further different, vertical exploration with upward and downward reasoning (allowing more scope for unseen information to be included in the responses):

ADA: *The hotel was highly rated, I would recommend it.*
User: *How was the view rated?*
ADA: *The view was rated highly, which positively affected how the location was rated".*
User: *How was the location rated?*
ADA: *Despite the fact that the walking routes were great, the location was rated poorly due to the transport links, which were poor.*
User: *Why were the transport links poorly rated?*
ADA: *A reviewer stated that "there were no close transport links".*

In the remainder of the paper we will focus on dialogical explanations following the top-down exploration as in the first illustration above. This dialogical explanation can be drawn automatically from the QBAF in Fig. 4 together with the σ_d row in Table 3 as gradual semantics, using the following simple *dialogical templates*, inspired by [30].[9]

[9] To obtain the natural language dialogue above, when deploying these templates, each argument is replaced by its semantic meaning as per the ontology, e.g. a_0^+ is replaced with "hotel" as per Fig. 3. Moreover, the user can select their contributions to the dialogues via pull-down menu or the like, e.g. as done in the case study in Section 8.2, see Fig. 9a.

Definition 7. A dialogical explanation $\langle \mathcal{S}(x), (\mathcal{Q}_1, \mathcal{X}_1), \ldots, (\mathcal{Q}_\delta, \mathcal{X}_\delta) \rangle$ between a user and an ADA is said to use the *strongest attacker and supporter template* iff, for each $i \in \{1, \ldots, \delta\}$, there exists $a_i^* \in \mathcal{A}$ such that $\mathcal{Q}_i$ concerns a_i^*, with $\mathcal{Q}_1$ concerning a_0^+ and $\mathcal{Q}_i$, for $i \neq 1$, concerning arguments other than a_0^+, and $(\mathcal{Q}_i, \mathcal{X}_i)$ is as follows[10]:

if $i = 1$ and $\sigma(a_0^+) \geq 0.5$ or $i \neq 1$ and $a_i^* = a_i^+$:

$\mathcal{S}(x) = \{$(The) a_i^* was highly rated, I would recommend it.$\}$;

$\mathcal{Q}_i = \{$Why was (the) a_i^* highly rated?$\}$;

if $\exists a_j^* \in \mathcal{L}^-(a_i^*) \cup \mathcal{L}^+(a_i^*)$ such that $\sigma(a_j^*) > 0$:

$\mathcal{X}(a_i^*) = \{$(The) a_i^* was highly rated$\} + p_a(max(\mathcal{L}^+(a_i^*))) + p_b(max(\mathcal{L}^-(a_i^*)))$;

if $\nexists a_j^* \in \mathcal{L}^-(a_i^*) \cup \mathcal{L}^+(a_i^*)$ such that $\sigma(a_j^*) > 0$:

$\mathcal{X}(a_i^*) = \{$A reviewer stated that$\} + \{$[phrase from some $\rho \in \mathcal{R}$ where $\mathcal{V}(\rho, n_i) = +$]$\}$

if $i = 1$ and $\sigma(a_0^+) < 0.5$ or $i \neq 1$ and $a_i^* = a_i^-$:

$\mathcal{S}(x) = \{$(The) a_i^* was poorly rated, I would not recommend it.$\}$;

$\mathcal{Q}(a_i^*) = \{$Why was (the) a_i^* poorly rated?$\}$;

if $i = 0$ and $\exists a_j^* \in \mathcal{L}^-(a_i^*) \cup \mathcal{L}^+(a_i^*)$ such that $\sigma(a_j^*) > 0$:

$\mathcal{X}(a_i^*) = \{$(The) a_i^* was poorly rated$\} + p_a(max(\mathcal{L}^-(a_i^*))) + p_b(max(\mathcal{L}^+(a_i^*)))$

if $i \neq 0$ and $\exists a_j^* \in \mathcal{L}^-(a_i^*) \cup \mathcal{L}^+(a_i^*)$ such that $\sigma(a_j^*) > 0$:

$\mathcal{X}(a_i^*) = \{$(The) a_i^* was poorly rated$\} + p_a(max(\mathcal{L}^+(a_i^*))) + p_b(max(\mathcal{L}^-(a_i^*)))$

if $\nexists a_j^* \in \mathcal{L}^-(a_i^*) \cup \mathcal{L}^+(a_i^*)$ such that $\sigma(a_j^*) > 0$:

$\mathcal{X}(a_i^*) = \{$A reviewer stated that$\} + \{$[phrase from some $\rho \in \mathcal{R}$ where $\mathcal{V}(\rho, n_i) = -$]$\}$

where, for any $S \subseteq \mathcal{A}$:

- if $S = \emptyset$ we let $max(S) = \emptyset$ and $p_a(S) = p_b(S) = \{\}$;
- else, $max(S) = \{s\}$ for some $s \in argmax_{s' \in S} \sigma(s')$ and p_a and p_b are functions such that and for any $a_k^* \in \mathcal{A}$:

if $a_k^* = a_k^+$:

$p_a(a_k^*) = \{$because (the) a_k^* was/were considered to be great$\}$

$p_b(a_k^*) = \{$despite the fact that (the) a_k^* was/were considered to be great$\}$

if $a_k^* = a_k^-$:

$p_a(a_k^*) = \{$because (the) a_k^* was/were considered to be poor$\}$

$p_b(a_k^*) = \{$despite the fact that (the) a_k^* was/were considered to be poor$\}$

According to this definition, then, a dialogical explanation is a (restricted) conversation between a user and an ADA in which the user queries the strengths of arguments and the ADA responds with reasoning for their strengths in terms of their attackers' or supporters' strengths or votes (from reviews). Specifically, the explanation of the reasoning for each argument's strength consists of its strongest attacker and its strongest supporter, if any, linked by *because/despite the fact that* connectives, depending on whether attackers or supporters are available. Note that, when an argument is poorly rated, the selection of these arguments differs for the item compared with the other arguments due to the different meaning of their strength evaluation ranges. If an argument has no attackers or supporters, we use phrases from the review (which constitute the votes towards the argument's base score) to explain its strength and to ensure that the explanation response is not empty. Thus, this form of explanation is afforded by all components of the QBAF (the attack and support relations and the base scores) and its gradual semantics.

Definition 7 imposes some restrictions on dialogical explanations, but still allows a high degree of freedom for ADAs, in that different pairs in a dialogical explanation using the strongest attacker and supporter template may be about the same argument and there is no requirement that later pairs follow-up on earlier ones. To limit and guide users' and ADAs' behaviour when deploying these templates, we can impose additional restrictions, e.g. the following:

Definition 8. A dialogical explanation $\langle \mathcal{S}(x), (\mathcal{Q}_1, \mathcal{X}_1), \ldots, (\mathcal{Q}_\delta, \mathcal{X}_\delta) \rangle$ using the strongest attacker and supporter template is said to be:

[10] We use "was" or "were" in the explanation responses here depending on whether the feature is singular or plural, respectively, and "the" where required.

- *non-repetitive* iff, for any $i, j \in \{1, \dots, \delta\}$, if Q_i concerns $a_k^* \in \mathcal{A}$ and Q_j concerns $a_l^* \in \mathcal{A}$, then $a_k^* = a_l^*$ iff $i = j$;
- *linear* iff, for any $j \in \{2, \dots, \delta\}$, there exists $i < j, i \in \{1, \dots, \delta - 1\}$, such that Q_j concerns an argument included in $\mathcal{X}_i$.

Here, non-repetitiveness requires that a dialogical explanation does not allow users to repeat explanation requests, ensuring that dialogical explanations are acyclic. Meanwhile, linearity requires that parents of arguments about which explanation requests are concerned have already been included in explanation responses, ensuring that the reasoning in the review aggregation is stepped through in turn. Note that neither of the restrictions will be suitable for every application, but they illustrate the generality and flexibility of Definition 6. Note also that the earlier illustration of dialogical explanation uses the strongest attacker and supporter template and is both non-repetitive and linear.

Proposition 13. *A linear dialogical explanation $\langle S(x), (Q_1, \mathcal{X}_1), \dots, (Q_\delta, \mathcal{X}_\delta)\rangle$ using the strongest attacker and supporter template is such that, for any $i \in \{2, \dots, \delta\}$, there exists a path (in the QBAF) from the argument that Q_i is concerned with to a_0^+.*

Proof. Follows directly from the fact that the ontology and thus the QBAF is a tree, that each response includes attackers and/or supporters for the argument that the request is concerned with, and by the non-repetitiveness and linearity requirements. □

Proposition 14. *A non-repetitive and linear dialogical explanation using the strongest attacker and supporter template is always guaranteed to exist.*

Proof. By Definition 1, $\mathcal{F} \neq \emptyset$, thus, by Definitions 2, 4 and 5, $\mathcal{A} \supset a_0^+$, and so it must be the case that $\exists a_j^* \in \mathcal{A} \setminus a_0^+$. Then, by Proposition 8, we can see that there must be a path from a_j^* to a_0^+ via $\mathcal{L}^- \cup \mathcal{L}^+$. Thus, $\mathcal{L}^-(a_0^+) \cup \mathcal{L}^+(a_0^+) \neq \emptyset$ and, trivially, an explanation request for a_0^+, Q_1, results in an attacker and/or a supporter of a_0^+ being returned as part of an explanation response, $\mathcal{X}_1$, as per Definition 7. It is easy to see that the dialogical explanation $\langle (Q_1, \mathcal{X}_1)\rangle$ is non-repetitive and linear. □

Note that if we impose further restrictions, e.g. that each explanation request concerns an attacker or supporter in the explanation response in the previous pair, if any, we can also enforce some bounds on δ (amounting to the length of the longest path from the item to the leaves – i.e. nodes without children – of the ontology).

We have focused here on providing illustrations of the flexibility of adopting dialogical explanations drawn from argumentation frameworks, and leave to future work characterising how ADAs and users may decide to engage in these explanations, e.g. by adopting strategies as in [73].

Even when dialogical explanations are as simple as those defined in this section, they allow to delve deep into the reasoning represented by the QBAF as well as the reviewers' comments which are used to generate the base scores. We envisage numerous directions down which future work could be directed to develop ADAs, now that we have introduced the formalism's foundations. More complex dialogical explanations using bigger repertoires of language templates, more fine-grained labelling of strengths (e.g. rather than only highly or poorly rated) and randomisation would improve the naturalness of the dialogue with only slightly more complex definitions (we discuss and illustrate some of these directions in Section 8). Meanwhile, technologies such as generative NLP models could be introduced to increase automation on either receiving requests from users or generating the text for the responses which contain the information provided by the QBAFs.

8. Case studies

In this section, we present two case studies: the first concerns *movie recommendation* (Section 8.1) and uses a bespoke a semi-automated ontology extraction process, while the second concerns *product recommendation* in the retail domain (Section 8.2) and uses a bespoke fully-automated ontology extraction process. The two case studies deploy bespoke methods for mining votes from text, ensuring that each review expresses at most one opinion (vote) about each item or feature of the ontology, as we discussed in Section 4. For each of the case studies, we give examples of dialogical explanations obtained by applying (variants of) the templates in Section 7. Overall, the main aim of this section is to show how to instantiate our abstract framework for review aggregation in practice.

8.1. Case study 1: an ADA in the movie domain

We focus on reviews from Rotten Tomatoes [11] (RT), a popular review aggregation website that aggregates critics' reviews on any movie into the movie's *Tomatometer Score* (TS), amounting to the percentage of critics who like the movie, having reviewed it as *fresh* rather than *rotten*. The TS is simplified to a binary classification for the movie, of fresh or rotten once again, based on whether at least 60% of critics review it as fresh or not, respectively. A short consensus is also written by a moderator to give a linguistic summary of the reviews. The TS, the fresh/rotten classification and the consensus give RT's users a simple way to determine whether a movie is worth watching or not.

[11] https://www.rottentomatoes.com.

In this case study, we assess whether an ADA deployed in this setting is able to match the TS score in an automated and explainable manner. We use a combination of previously defined features related to movies and movie metadata as well as features extracted from the textual reviews to construct an ontology and votes, as discussed below. Then, the ADA framework can be used to obtain QBAFs from which scores (via the DF-QuAD gradual semantics) for movies and dialogical explanations thereof can be generated automatically. These scores can then be directly compared with the TS, with the additional benefit of being faithfully explained by the ADA.

8.1.1. Generating the item ontology

We define the ontology in terms of a feature-based characterisation of movie reviews, obtained from metadata and the top critics' snippets that appear on RT's movie pages (e.g. see the movie *Arrival*'s reviews[12]). This feature-based characterisation gives rise to $\mathcal{F} = \mathcal{F}' \cup \mathcal{F}''$ where all elements of $\mathcal{F}'$ are children of the item (movie) in the ontology, and all elements of $\mathcal{F}''$ are children of elements of $\mathcal{F}'$. $\mathcal{F}'$ comprises features f_A, f_D, f_W, f_T, where f_A is *acting*, f_D is *directing*, f_W is *writing* and f_T is *themes*: other features, e.g. *cinematography*, could be included but we opted for f_A, f_D, f_W and f_T only as they are the most frequently occurring in RT. The elements of $\mathcal{F}''$ may be of different types, namely *single* (e.g. in the case of children of features f_D or f_W, if we only consider movies with a single director or writer: for *Arrival*, *Denis Villeneuve* is the sole director) or *multiple* (e.g. for the children of f_T, since movies will generally be associated with several themes, and for the children of f_A, as movies will generally have more than one actor: *Arrival* has *Amy Adams*, *Jeremy Renner* and *Forest Whitaker* as children of f_A). Furthermore, the elements of $\mathcal{F}''$ may be classified as *predetermined*, namely obtained from meta-data (as for the children of f_A, f_D, f_W), or *mined* from (snippets of) reviews (e.g. for *Arrival* the child *sci-fi* of f_T may be mined rather than predetermined). To determine the mined features in $\mathcal{F}''$ we can use semantic information, e.g. in our experiments we use the semantic network ConceptNet[13] to identify related terms for f_T. For example, for *Arrival*, we identify, as children of f_T, *sci-fi* (f_{T1}) as several reviews mention the related terms *sci-fi* and *alien*, as in *'It doesn't pack the punch you'd expect from an alien invasion movie; it aims higher, and in the end, it cuts deeper'*.

8.1.2. Generating votes

Using this feature-based characterisation of a movie, we use sentiment analysis to generate votes from snippets from the movie reviews written by critics. In order to guarantee that each node has a maximum of one vote from each review, we apply Algorithm 1. Basically, we analyse each critic's review independently, tokenising it into sentences, which are then split into phrases when specific

Algorithm 1: Pseudocode to determine votes.

```
Input:
        Review ρ, Glossary G, Keywords K
        neutrality_threshold ← 0.6
        SA;        /* Sentiment analysis model that, given a text, returns the text's polarity, i.e. a number in ℝ  */
sentences ← extract_sentences(ρ)
phrases ← split_by_keywords(sentences, K)
for (p1, p2) ∈ phrases × phrases do
    /* n is a term from the glossary G                                                                   */
    if SA(p1) > 0 and n ⊂ p1 and SA(p2) < 0 and n ⊂ p2 then
        V(ρ, n) = max(abs(SA(p1), SA(p2)))
end
for p ∈ phrases do
    if |SA(p)| > neutrality_threshold then
        if n1 ⊂ p and n2 ⊂ p and E(n1) = n2 then
            V(ρ, n1) = SA(p).
        if n ⊂ p or ∃G(n) ⊂ p then
            V(ρ, n) = SA(p).
end
```

keywords (in K ={*but, although, though, otherwise, however, unless, whereas, despite*}) occur. Each phrase may then constitute an argument with a vote for the feature(s) mentioned. For illustration, consider the following review for *Arrival*:

> ρ_1: *[Adams] delivers a heart-wrenchingly beautiful performance using her ability to communicate a half-dozen emotions just standing still... But director Denis Villeneuve sometimes gets lost in repetition and blind alleys.*

Here, Algorithm 1 extracts two phrases: p_1: *[Adams] delivers a heart-wrenchingly beautiful performance...* and p_2: *But director Denis Villeneuve sometimes gets lost in repetition and blind alleys.* A review comprising a single sentence, e.g.

> ρ_2: *What's remarkable about Arrival is its contemplative core - and, of course, Ms. Adams's star performance, which is no less impassioned for being self-effacing*

[12] https://www.rottentomatoes.com/m/arrival_2016/reviews?type=top_critics.

[13] http://conceptnet.io/.

Table 5
Glossary used for the movie domain.

n	$G(n)$
m	movie, film, work
f_D	director
f_A	acting, cast, portrayal, performance
f_W	writer, writing, screenplay, screenwriter, screenwriting, storyline, script, character

may be split into: p_3: *What's remarkable about Arrival is its contemplative core* and p_4: *Ms. Adams's star performance, which is no less impassioned for being self-effacing.* Finally, consider the review

ρ_3: *Arrival is not a linear adventure of the mind, and it is a film probably best seen twice.*

Here, Algorithm 1 extracts two phrases concerning *Arrival*: p_5: *Arrival is not a linear adventure of the mind* and p_6: *a film probably best seen twice*.

Next, Algorithm 1 uses a *glossary G* of movie-related words to determine the (phrases on which the) votes are extracted. Table 5 shows the glossary for movies in general as well as for some of the features. When determining votes, children features take precedence over parent features. E.g. a mention of "Amy Adams" (f_{A1}) (with or without a word from $G(f_A)$) connects with f_{A1}, whereas a sole mention of any word from $G(f_A)$ connects with f_A. A text that contains two entities (a child feature and a word from the glossary for the parent feature) results in two (phrases and) votes, one for each feature identified.

Once phrases have been identified, Algorithm 1 uses sentiment analysis to extract votes on them. For each phrase extracted, we determine the sentiment polarity which translates into a (negative or positive) vote from the corresponding review. We impose a threshold on the sentiment polarity to filter out phrases that can be deemed to be "neutral" and therefore cannot be considered to be votes. Votes are then assigned to phrases based on occurrences of words from G. From our *Arrival* example, the ADA may extract one vote for the feature *Amy Adams* (f_{A1}) from p_1 with polarity 0.7, therefore giving $\mathcal{V}(\rho_1, f_{A1}) = +$. If phrase p_2 is deemed to have polarity -0.7 and to be assigned to f_D, then $\mathcal{V}(\rho_1, f_D) = -$. From review ρ_2, Algorithm 1 may extract one vote for the movie in general and one for the feature *Amy Adams* (f_{A1}), e.g. p_3 gives sentiment 0.8 to m, and thus therefore $\mathcal{V}(\rho_2, m) = +$, while p_4 gives sentiment 0.6 to f_{A1}) therefore $\mathcal{V}(\rho_2, f_{A1}) = +$. If the review of a single critic results in several phrases associated with a node in the ontology with different polarities, Algorithm 1 takes that with the highest sentiment magnitude to determine the vote. We make this design choice with the motivation that a stronger sentiment may override a weaker one, though this could lead to the loss of useful information. Other options could have been taken, e.g. an average of the sentiments, which we leave to future work. For example, given: p_5: $(-0.57, m)$ and p_6: $(0.63, m)$, p_6 supersedes and $\mathcal{V}(\rho_3, m) = +$. Here, the sentiment polarity for p_5 is incorrect but, if the neutrality threshold is ± 0.6 as in our experiments, then the negative vote is ignored. Algorithm 1 can be used to determine the votes for the mined feature in the same way as for the other features. For example, in the ase of f_T, given

ρ_4: *It doesn't pack the punch you'd expect from an alien invasion movie; it aims higher, and in the end, it cuts deeper*

leading to $(0.6, f_{T1})$, we obtain $\mathcal{V}(\rho_4, f_{T1}) = +$.

It can be seen that this method for extracting sentiment is not infallible, e.g. when identifying the general sentiment of a sentence which contains multiple phrases of different polarity, but we believe that it presents a reasonable approximation of the sentiment expressed a set of reviews distilled over the distinct relevant features, and thus a starting point for future work.

8.1.3. Experiments & discussion

We tested our proposed method on the aforementioned task of matching a movie's argumentative strength with the TS. To do so, we use DF-QuAD on the QBAFs obtained as per Section 5 with (augmented) aggregation frameworks drawn from the ontology and votes given earlier in this section. The dataset we use consists of the box-office movies from January 2015 to August 2018 inclusive and on the top 100 movies of all time on RT,[14] giving a total of 1281 movies after removing those without reviews. We conducted experiments to determine how well the strength of (arguments for) movies compare with their TS.[15] In the experiments, we analysed movies for which we obtain votes from at least 33% of the critics who reviewed the movies, removing those for which few votes were mined. This results in approximately 900 movies and 35219 snippets, from which we analyse 33554 phrases/arguments.

To evaluate how closely our method mirrors the TS in a quantitative assessment, we modelled the problem as a rating prediction task where we categorised the numerical scale 0-100 into groups 0-9 that represent the predicted "grade" of the movie. For example, a movie grade of 7 represented scores in the range 70-79 and a grade of 9 represented scores in the 90-100 range. We then computed the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE) on the predicted grade. Our method achieved MAE of 1.527

[14] https://www.rottentomatoes.com/top/bestofrt/.

[15] Note that in [30] we considered other NLP techniques to mine votes (i.e. argument mining), but this did not yield any performance gains, hence we omit them here. Also, in [30], we run experiments also with other gradual semantics (i.e. QuAD [11] and REB [3]), but we omit them here based on our theoretical analysis in Section 6: we note, in any case, that the use of these other semantics did not yield any performance gains.

Table 6
Examples of mined themes for movies.

Movie	Theme	Keywords
Lady Bird	adolescence	adolescence, coming of age
The Greatest Showman	entertainment	impresario, amusement park
A Quiet Place	terror	fright, terror, dread, taut, suspense
La La Land	relationship	rapport, intimacy
Moana	melody	riff, song
Phantom Thread	clothing	clothe, fashion, dress, designer

and RMSE of 1.927, in line with existing work [52], demonstrating that our method is able to mirror the TS reasonably accurately when using complex, real-world data. We also evaluated our method on the more complex task of predicting the exact numerical rating as opposed to the rating group and observed an expected drop in performance, MAE 15.978 and RMSE 19.329. These results are also encouraging, demonstrating that ADAs could feasibly replace existing methods deployed for review aggregation in the real world.

We also undertook a qualitative assessment in this task for error analysis. We obtain movie scores close to the TS, e.g. for *RBG* (86 vs TS 90), exact scores, e.g. for *ET The Extraterrestrial* (97), and also results in contrast with the TS, e.g. for *The Greatest Showman* (79 vs TS 36).[16] While some discrepancy is expected, especially when the critics' fresh/rotten assessment is not clearly argued for in the review snippets, large discrepancies are undesirable. One reason for these large discrepancies is the errors which may occur in extracting votes, e.g. from the review: *The director, Michael Gracey, delivers quick doses of excitement in splashy scenes but has little feel for the choreographic action* we only obtain *...doses of excitement in splashy scenes*, resulting in a positive vote. Another reason is the failure to identify any votes from reviews such as the following (negative) text: *In a broader sense, the mishmash does recall the real Barnum, who once sewed half a fish to half an ape and called it a mermaid*, as it does not mention any feature or opinion about the movie itself. With further development of our method and of NLP techniques in general, we envisage that the frequency and impact of these errors will be reduced.

In addition, we experimented with the use of mined features, i.e. children of the f_T feature, as given in Table 6. These mined features can modify a movie's score (i.e. strength) significantly, e.g. for *Call Me by Your Name* from 77 to 88 (TS 98) and for *The Godfather* from 78 to 90 (TS 95), demonstrating the importance of these features.

In addition to scoring movies, ADAs can provide dialogical explanations for the scores, based on the underpinning QBAFs, as we discuss next.

8.1.4. Generating dialogical explanations

We provide several illustrations, for different movies in RT, of non-repetitive and relevant dialogical explanations we obtain using the strongest supporter and attacker template given in Section 7.

Consider the QBAF in Fig. 5 for the movie *Phantom Thread* (TS = 98). All four features in $\mathcal{F}'$ support the movie and thus represent positive reasoning for its strength, with *Daniel Day-Lewis* (f_{A1} in $\mathcal{F}''$) supporting the movie's strongest feature, f_A, and *clothing* (f_{T1} in $\mathcal{F}''$) supporting the next strongest, f_T. A dialogical explanation between a user and an ADA may then be as follows:

ADA: *The movie (Phantom Thread) was highly rated, I would recommend it.*
user: *Why was the movie highly rated?*
ADA: *The movie was highly rated because the acting was considered to be great.*
user: *Why was the acting highly rated?*
ADA: *The acting was highly rated because Daniel Day-Lewis was considered to be great.*
user: *Why was Daniel Day-Lewis highly rated?*
ADA: *A reviewer stated that "...Daniel Day-Lewis remains our greatest actor..."*
user: *Why were the themes highly rated?*
ADA: *The themes were highly rated because the clothing (theme) was considered to be great.*
user: *Why was the clothing (theme) being highly rated?*
ADA: *A reviewer stated that "...it's set in an evocative ecosphere of haute couture fashion..."*

This dialogue is fairly repetitive given this movie's almost universally positive reviews (as well as our dialogical explanations' fairly simple nature for illustration) but consider the QBAF for *The Post* (TS = 91) in Fig. 6. This QBAF has a different structure from that for *Phantom Thread* as no themes were mined for this movie (i.e. f_T has no effect and so is not shown), there is an attack and it has two children supporting f_A. We can see that f_D positively contributes very little and f_W is actually considered to be poor since it attacks m. However, the acting from *Tom Hanks* (f_{A1} in $\mathcal{F}''$) and, particularly, *Meryl Streep* (f_{A2} in $\mathcal{F}''$) contributed to the high strength. The dialogical explanation may then be:

ADA: *The movie (The Post) was highly rated, I would recommend it.*
user: *Why was the movie highly rated?*

[16] Here we report TS scores and use reviews available in the period of interest for the experiments, see the description at the beginning of Section 8.1.3.

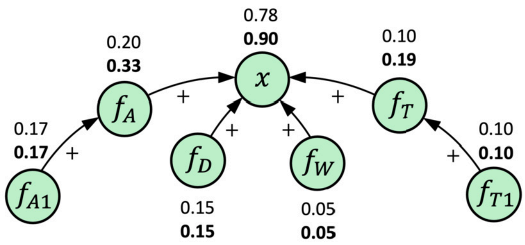


Fig. 5. Graphical representation of the QBAF for the movie *Phantom Thread* with positive arguments in green (no negative arguments were present here), support relations labelled + (no attack relations were present here), base scores in normal font and DF-QuAD strengths in bold font.

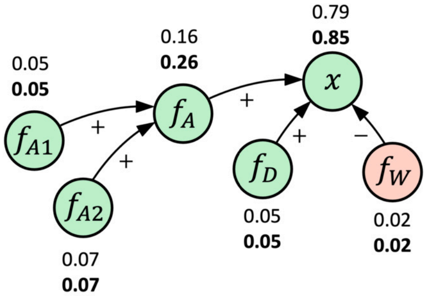


Fig. 6. Graphical representation of the QBAF for the movie *The Post* with positive arguments in green, negative arguments in red, attack relations labelled −, support relations labelled +, base scores in normal font and DF-QuAD strengths in bold font.

ADA: *The movie was highly rated because the acting was considered to be great, despite the fact that the writing was considered to be poor.*
user: *Why was the acting highly rated?*
ADA: *The acting was highly rated because Meryl Streep was considered to be great.*
user: *Why was Meryl Streep highly rated?*
ADA: *A reviewer stated that "...Streep's hesitations, rue, and ultimate valor are soul-deep..."*

There are a number of ways in which more complicated dialogical explanations could be defined to give more interesting and varied interactions in the movie domain in particular. For example, the strength scale could be separated to a greater degree and attackers/supporters other than those with the maximum strength could be considered for different levels of sentiment, rather than simply *great/poor*, e.g. *the writing was considered to be okay but not great*; functions giving phrases could depend upon the type of argument for a more natural sounding phrase, e.g. for an actress: *Meryl Streep put in an excellent performance*; or arguments could be considered in tandem, e.g. *the acting was considered to be exceptional thanks to Tom Hanks and, particularly, Meryl Streep*. We believe that large language models, as in [18,87,28], could play an important role here, generating more natural-sounding dialogues while being constrained to the information contained in ADAs, limiting their associated risks, such as hallucinations. We leave the exploration of these likely fruitful variant of our dialogical explanations to future work.

8.2. Case study 2: an ADA in the retail domain

In the second case study, we focus on reviews from Amazon,[17] an extremely popular site for retail.

To obtain a suitable ontology, we define a method for automatically extracting product ontologies from textual reviews in the presence of little annotated data, using distantly supervised learning with *masked BERT* [34]. We use two BERT models, one to identify *aspects* of products and one to determine *part-of-relations* between these aspects. We then use a further BERT model for aspect-based sentiment analysis for generating votes. Then, the ADA framework can be used to obtain QBAFs from which scores (via the DF-QuAD gradual semantics) for products and dialogical explanations therefor can be generated automatically.

8.2.1. Generating the item ontology

Manually extracting ontologies from reviews that would be representative of an entire set of products (such as those available on Amazon) would be nearly impossible due to the large number of different product categories. Based on the observation that, in review texts, certain grammatical constructs are similar irrespective of the product being reviewed, we manually created a small ontology for a few (randomly) chosen categories of products[18] to annotate data to be used in the ontology extraction. For example, consider the following two sentences, where *lens* is a feature of *camera* and *material* is a feature of *sweater*:

[17] https://www.amazon.co.uk.

[18] In the experiments, we used an ontology with only 5 products, a maximum number of features per product of 54 and a maximum depth of the ontology of 5, which took less than 2 hours to create.

"*I love the **lens** on this **camera***";
"*I love the **material** of this **sweater***".

Masking the entities in bold in the sentences, we obtain:

"*I love the **e1** on this **e2***";
"*I love the **e1** of this **e2***".

Furthermore, given sentences:

"*The **camera housing** is made of shiny **plastic** but it feels nicely weighted*", and
"*It's a lovely warm **jumper** that has a nice **feel** to it*",

when masking the entities in bold, the sentences still contain domain-specific terms such as *shiny*, *weighted*, and *warm*, that are common to wider categories of products, such as *electronics* or *clothing*. Thus, we argue that a varied set of annotated reviews for few selected products should suffice to obtain training data for ontology extraction. To achieve this, we used distantly supervised learning with *masked BERT* [34] as follows. We draw inspiration from [56] and replaced random words with a [MASK] token, letting the model predict these words based on the context provided by the non-masked words in the sequence. Terms that appear in the ontology are automatically identified and masked while also labelling sentences in the review texts to obtain training data for the tasks of aspect and relation extraction. We used a BERT model to identify *aspects* of products and another BERT model to determine *parts-of relations* between aspects. Note that, without masking, the classifiers would simply learn to recognise the terms in the manually created ontology; instead, with masking, they are forced to rely solely on each term's context, which is more varied.

In order to determine the aspects, we first identify the 200 most commonly appearing entities using an out-of-the-box implementation of [61] to join common co-occurrences of tokens into bigrams and trigrams, and select the review sentences that mention exactly one of the entities obtained. Then, we pass these sentences through a BERT classifier to determine whether the entity is a feature aspect, a product aspect, or not an aspect as follows: the input consists of a review sentence (e.g. "The operating system is great"), as well as the entity we aim to classify as a feature aspect, product aspect, or non-aspect (e.g. "operating system"), and we mask the entity with a single [MASK] token. The tokens are then passed through the transformer network and the input for classification is taken from the output position of the masked entity. The linear layer is followed by a softmax operation, which outputs the probabilities of the entity being a non-aspect, feature aspect, or product aspect, respectively.

The same aspect can be referred to by reviewers using different terms (e.g. a *hard drive* can be referred to with the terms *HDD* or *storage*) that need to be grouped under the same aspect node when constructing the ontology tree. However, *hard drive* and *storage* are not strict synonyms even if they are interchangeable within the review texts. To address this, we train a word2vec model on the review texts using the continuous bag-of-words (CBOW) architecture [60] that predicts each word in the text corpus from a window of surrounding words.[19] Once we obtain the word embedding for each aspect, we use the relative cosine similarity measure to construct a weighted graph of aspects to be used in the *Equidistant Nodes Clustering* (ENC) method [26], which was shown to obtain state-of-the-art performance in the synset induction task.

The aspect synsets then form the nodes of the ontology tree as follows. The review sentences that mention a term from exactly two synsets are passed through a BERT classifier to obtain votes for relations between aspects. We aggregate the votes within each synset to obtain a two-way relatedness measure between each synset pair (i.e. if s_i is a feature of s_j, if s_j is a feature of s_i, or if there is no relation between s_i and s_j) that we use to create the ontology. Fig. 7 shows the ontology extracted for *television*.

8.2.2. Generating votes

Given the ontology, votes on aspects of relevant products are drawn automatically from the reviews by automatically identifying the (positive or negative, respectively) sentiment of excerpts of the reviews concerning those aspects. We used TD-BERT [39] trained on SemEval-2014 Task 4 data for aspect-based sentiment analysis [67]. Fig. 8 shows predominantly positive votes in green and predominantly negative votes in red, as well as an excerpt for a review about the "design" aspect. Votes are then used to construct the QBAF whereby (excerpts about) aspects of the ontology form arguments supporting or attacking other aspects or the ("goodness" of the) product, as we outlined in Section 4. For example, for the votes in Fig. 8, "design" supports the product ("mixer").

8.2.3. Experiments & discussion

To evaluate our ontology extraction method, we selected 5 random product categories from [64] (*digital cameras*, *backpacks*, *laptops*, *acoustic guitars*, and *cardigans*); for each, we identified the 200 most commonly appearing nouns in the reviews, and manually created the (input) ontology around them, given that we observed that most of the relevant aspects of the products were included within this (200 nouns) range. This led to over 40 K and 200 K training instances per product and in total (respectively) for aspect extraction and to over 20 K and 100 K for relation extraction. We then trained the two BERT models on the aspect and relation extraction data, respectively, leaving 5% of the data as a held-out testing set, for 3 epochs with a batch size of 32 and 16 (respectively) and Adam

[19] We observed that a small window size of 4 performed best. We hypothesise that this is because a small window size can prevent sibling aspects from being grouped together based on their association with their parent aspect.

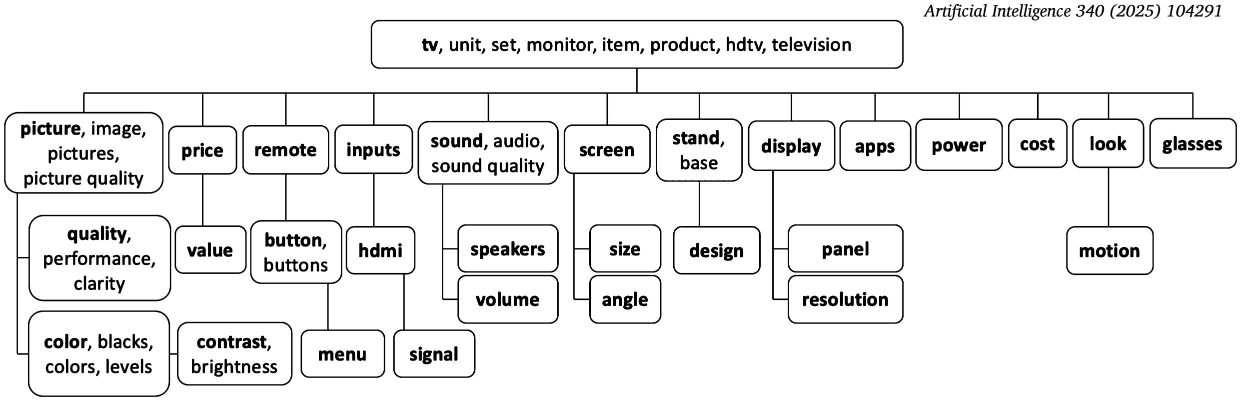


Fig. 7. The ontology we extracted for *television*. Features are highlighted in bold, whereas their synonyms are given in standard font.

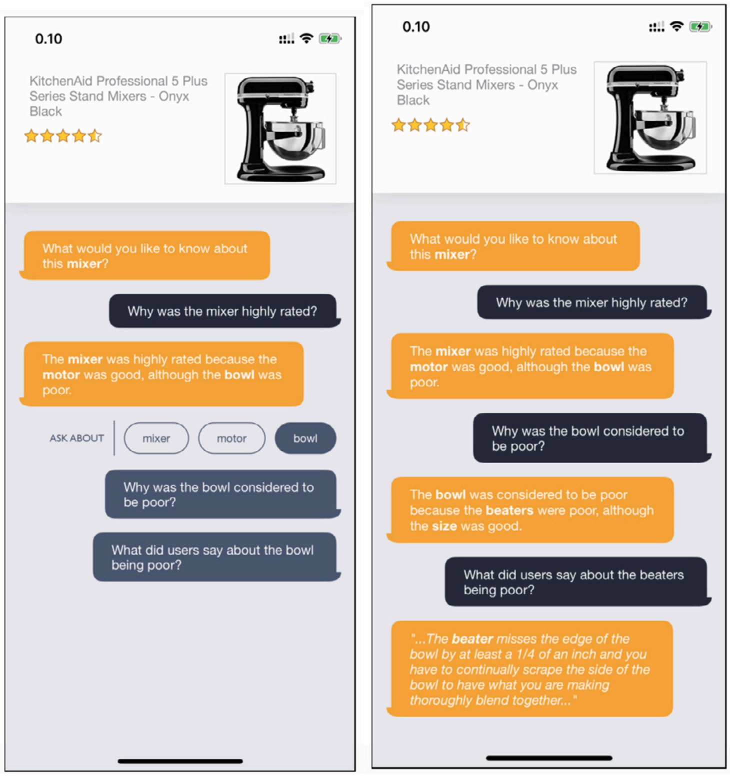
Fig. 8. Votes with predominantly positive/negative votes indicated in green/red, respectively, built from Amazon reviews.

optimiser with standard cross entropy loss. We obtained 82.37% accuracy and 81.43% macro F_1 on the aspect extraction testing set and 79.73% accuracy and 79.84% macro F_1 on the relation extraction testing set. We considered five randomly selected products not included in the training data, representative of 4 out of the 29 distinct Amazon categories in the dataset: *watches*, *necklaces* (both from category Clothing Shoes and Jewelry), *televisions* (Electronics), *stand mixers* (Home and Kitchen), and *video games* (Video Games). This choice includes both popular (*televisions*) and niche (*stand mixers*) products so as to cover a wide spectrum of products. For each product, we used 200,000 review texts as input to the ontology extractor, except for *stand mixer*, for which we could only obtain 28,768 review texts. The QBAFs extracted are evaluated dialectically to obtain a numerical score for the product, using the DF-QuAD gradual semantics as in the previous case study.

We compared the scores given by the gradual semantics for products with ontologies drawn from our ontology extractor to the "ground truth" aggregated score from Amazon, amounting to the average star ratings of the products therein, scaled linearly from 0-5 to 0-100. The intuition is that the score given by the gradual semantics gives a measure of the dialectical support products receive from the opinions expressed in the reviews, along the dimensions identified by the ontologies, and thus this score should be comparable to the aggregated star ratings reviewers give the products. Table 7 shows the mean average error (MAE) and root mean square error (RMSE) for each product category. Here, we compute scores for the 20 most reviewed products in each testing category.

Table 7
Error values we obtain with the proposed ontology extraction method.

	Watch	TV	Necklace	Mixer	Video game
MAE	6.7	3.4	8.8	22.7	12.6
RMSE	8.5	6.7	10.9	23.4	14.6



(a) User controls (b) Conversation

Fig. 9. Illustration of our explainable recommender system built from Amazon reviews.

As seen from the first case study, predicting the exact numerical rating on a scale of 0-100 is a challenging task. Here, the experiments run on Amazon products have the additional complexity of an out-of-distribution test set, with all five product categories tested not seen during training. As expected, the errors are higher than those obtained in the movie experiments, and the results for the niche product, stand mixer, are lower than those of popular products such as the TV.

8.2.4. Generating dialogical explanations

Explanations can be drawn from the QBAF in Fig. 8 as we have shown in Section 7: from (arguments about) the "goodness" of the product to (attacking or supporting arguments for) the "goodness" of aspects thereof, to evidence from reviews from which the QBAF is extracted. Fig. 9 illustrates the dialogical explanations we have embedded into a mobile app prototype. Fig. 9a shows an intermediate step in the generation of the conversation, where the user is given options as to what can be asked from the review aggregation system at that stage, using a similar template as in Section 7, but where users can directly ask about information in reviews. Fig. 9b shows the full conversation generated with a human user. Here, at the second utterance by the review aggregation system, the user is given a positive and a negative aspect (the "motor" and the "bowl", drawn respectively from a supporter and an attacker for the product in the QBAF) and opts for exploring, within the conversation, the reasons for the negative aspects, for which, at the third utterance by the system, a positive and a negative reason (aspects) are provided, supported by evidence at the fourth utterance. Overall, the conversations can deliver explanations fully tailored to the needs of users (in allowing exploring and retrieving information in aspects they care about).

9. Conclusions

We have introduced ADAs, a.k.a. argumentative dialogical agents: a novel, argumentation-based formalism for explainable aggregation of (textual) reviews. We have demonstrated how ADAs can extract feature-based representations as a means for harbouring the information from reviews on items, using an ontology. We have shown that these ontologies may be obtained by different means with varying levels of automation, enabling generality to make ADAs applicable to a wide range of settings where textual reviews are available. ADAs then use the feature-based representation to generate a form of argumentation frameworks, QBAFs, which drive both the evaluation of items via gradual semantics for QBAFs, and the dialogical explanations therefor, which are delivered to users and support interactions with them. We have formally studied (components of) ADAs in terms of a number of theoretical properties. We have also shown how to deploy ADAs, focusing on three diverse application settings for which popular websites exist where review aggregation is paramount, but achieved without any regards for explainability. In doing so, we have provided comprehensive evidence of the adaptability of ADAs to different tasks and settings, as well as how they can support explainability.

We foresee numerous directions for future work, starting with the investigation of the relaxation of the assumptions and simplifications which were made throughout the paper. First, considering the use of different argumentation frameworks could be fruitful, e.g. incorporating structure in arguments could allow for more expressiveness, though more error would likely be induced in the argument mining, while the use of argument schemes, as in [95], while restrictive, could allow for more natural dialogical explanations with users. Also, allowing for cyclic argumentation frameworks, as in [22], could increase the flexibility of our approach, while other base scores or semantics could be experimented with. Moreover, in the case studies, we have demonstrated just a few options of how our approach can be instantiated in real world settings, e.g. from how the ontologies are obtained to the selection the aggregation functions for votes, but there are many more options which could be explored within our methodology. Our ontologies are also rather simple, as are the binary, single votes we extract, where some form of graded vote based on the sentiment analysis, or more complex aggregation functions when multiple votes are present, could improve the accuracy in the empirical analysis. The dialogical explanations which we have introduced also present a starting point from which we expect many developments, e.g. in the richness of the possible conversations or the way in which they are generated.

In addition to these aforementioned investigations, we would also like to expand our study in other directions, from further development of the formalisations of ADAs (e.g. with more sophisticated ontologies), to implementation and evaluation of user-facing systems. It would also be interesting to explore additional methods for the automated extraction of ontologies, in the settings we considered as well as other settings. Further, it would be interesting to see what role LLMs, as in [18,87,28], could play in ADAs for generating more natural dialogical explanations to alleviate some of the reliance on templates (as already mentioned) as well as to automatically classify relations as being attacks or supports, given promising work within argumentation [86] and in e-commerce [2]. We would also like to consider how the argumentation framework can incorporate user preferences, e.g. as is overviewed in general and explored in this context in [12], rather than accommodating them indirectly via explanation of the review aggregation. Finally, we plan to conduct user studies with ADAs to determine the best way to provide the explanations, e.g. graphical, visual or linguistic, for different contexts and users, as in [41]. Specifically, we plan to explore the merits of the different (template-driven) dialogical explanations demonstrated in the paper. These user studies will be pivotal for assessing which explanations to adopt in real world deployments, and will require accounting for the strategic behaviour of ADAs and users engaged in the dialogical explanations.

CRediT authorship contribution statement

Antonio Rago: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Oana Cocarascu:** Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Data curation, Conceptualization. **Joel Oksanen:** Writing – original draft, Software, Methodology, Data curation, Conceptualization. **Francesca Toni:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

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 (grant agreement No. RCSRF2021\11\45). 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). Any views or opinions expressed herein are solely those of the authors listed. Finally, we would like to thank the reviewers and the editor, whose feedback and suggestions helped us to significantly improve the paper.

Data availability

The data used is publicly available as indicated in the paper.

References

[1] E. Albini, P. Lertvittayakumjorn, A. Rago, F. Toni, DAX: deep argumentative explanation for neural networks, CoRR, arXiv:abs/2012.05766, arXiv:2012.05766, 2020.

[2] A.L.F. Alves, C. de Souza Baptista, L. Barbosa, C.B.M. Araujo, Cross-lingual learning strategies for improving product matching quality, in: Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing, SAC 2024, Avila, Spain, April 8–12, 2024, ACM, 2024, pp. 313–320, https://doi.org/10.1145/3605098.3636001.

[3] L. Amgoud, J. Ben-Naim, Evaluation of arguments in weighted bipolar graphs, Int. J. Approx. Reason. 99 (2018) 39–55, https://doi.org/10.1016/j.ijar.2018.05.004.

[4] C. Antaki, I. Leudar, Explaining in conversation: towards an argument model, Eur. J. Soc. Psychol. 22 (1992) 181–194, https://doi.org/10.1002/ejsp.2420220206.

[5] A. Arioua, P. Buche, M. Croitoru, Explanatory dialogues with argumentative faculties over inconsistent knowledge bases, Expert Syst. Appl. 80 (2017) 244–262, https://doi.org/10.1016/j.eswa.2017.03.009.

[6] K. Atkinson, P. Baroni, M. Giacomin, A. Hunter, H. Prakken, C. Reed, G.R. Simari, M. Thimm, S. Villata, Towards artificial argumentation, AI Mag. 38 (2017) 25–36, https://doi.org/10.1609/aimag.v38i3.2704.

[7] K. Atkinson, T.J.M. Bench-Capon, P. McBurney, A dialogue game protocol for multi-agent argument over proposals for action, Auton. Agents Multi-Agent Syst. 11 (2005) 153–171, https://doi.org/10.1007/s10458-005-1166-x.

[8] H. Ayoobi, N. Potyka, F. Toni, SpArX: Sparse Argumentative Explanations for Neural Networks, ECAI 2023 - 26th European Conference on Artificial Intelligence, September 30 - October 4, 2023, Kraków, Poland - Including 12th Conference on Prestigious Applications of Intelligent Systems (PAIS 2023), vol. 372, IOS Press, 2023, pp. 149–156, https://doi.org/10.3233/FAIA230265.

[9] P. Baroni, D. Gabbay, M. Giacomin, L. van der Torre (Eds.), Handbook of Formal Argumentation, College Publications, 2018, https://philpapers.org/rec/BARHOF.

[10] P. Baroni, A. Rago, F. Toni, How many properties do we need for gradual argumentation?, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th Innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2–7, 2018, AAAI Press, 2018, pp. 1736–1743, https://doi.org/10.1609/aaai.v32i1.11544.

[11] P. Baroni, M. Romano, F. Toni, M. Aurisicchio, G. Bertanza, Automatic evaluation of design alternatives with quantitative argumentation, Argum. Comput. 6 (2015) 24–49, https://doi.org/10.1080/19462166.2014.1001791.

[12] E. Battaglia, P. Baroni, A. Rago, F. Toni, Integrating user preferences into gradual bipolar argumentation for personalised decision support, in: Scalable Uncertainty Management - 16th International Conference, SUM 2024, Palermo, Italy, November 27-29, 2024, Proceedings, vol. 15350, Springer, 2024, pp. 14–28, https://doi.org/10.1007/978-3-031-76235-2_2.

[13] P. Besnard, A.J. García, A. Hunter, S. Modgil, H. Prakken, G.R. Simari, F. Toni, Introduction to structured argumentation, Argum. Comput. 5 (2014) 1–4, https://doi.org/10.1080/19462166.2013.869764.

[14] P. Bhatia, Y. Ji, J. Eisenstein, Better document-level sentiment analysis from RST discourse parsing, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015, in: The Association for Computational Linguistics, 2015, pp. 2212–2218, https://doi.org/10.18653/v1/d15-1263.

[15] B. Bickart, R.M. Schindler, Internet forums as influential sources of consumer information, J. Interact. Mark. 15 (2001) 31–40, https://doi.org/10.1002/dir.1014.

[16] T. Bosc, E. Cabrio, S. Villata, Tweeties squabbling: positive and negative results in applying argument mining on social media, in: Computational Models of Argument - Proceedings of COMMA 2016, Potsdam, Germany, 12–16 September, 2016, IOS Press, 2016, pp. 21–32, https://doi.org/10.3233/978-1-61499-686-6-21.

[17] C.E. Briguez, M.C. Budán, C.A.D. Deagustini, A.G. Maguitman, M. Capobianco, G.R. Simari, Argument-based mixed recommenders and their application to movie suggestion, Expert Syst. Appl. 41 (2014) 6467–6482, https://doi.org/10.1016/j.eswa.2014.03.046.

[18] T.B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D.M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, 2020, https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

[19] M.C.D. Budán, M.L. Cobo, D.C. Martínez, G.R. Simari, Proximity semantics for topic-based abstract argumentation, Inf. Sci. 508 (2020) 135–153, https://doi.org/10.1016/j.ins.2019.08.037.

[20] L. Carstens, F. Toni, Towards relation based argumentation mining, in: Proceedings of the 2nd Workshop on Argumentation Mining, ArgMining@HLT-NAACL 2015, June 4, 2015, Denver, Colorado, USA, in: The Association for Computational Linguistics, 2015, pp. 29–34, https://doi.org/10.3115/v1/w15-0504.

[21] D. Ceolin, L.S. Ootes, Review argumentation at scale, in: Computational Models of Argument - Proceedings of COMMA 2024, Hagen, Germany, September 18–20, 2014, vol. 388, IOS Press, 2024, pp. 13–24, https://doi.org/10.3233/FAIA240306.

[22] D. Ceolin, G. Primiero, M. Soprano, J. Wielemaker, Transparent assessment of information quality of online reviews using formal argumentation theory, Inf. Syst. 110 (2022) 102107, https://doi.org/10.1016/j.is.2022.102107.

[23] D. Ceolin, G. Primiero, J. Wielemaker, M. Soprano, Assessing the quality of online reviews using formal argumentation theory, in: Proceedings of the Web Engineering - 21st International Conference, ICWE 2021, Biarritz, France, May 18-21, 2021, 2021, pp. 71–87, https://doi.org/10.1007/978-3-030-74296-6_6.

[24] L.A. Chalaguine, A. Hunter, A persuasive chatbot using a crowd-sourced argument graph and concerns, in: Computational Models of Argument - Proceedings of COMMA 2020, Perugia, Italy, September 4–11, 2020, 2020, pp. 9–20, https://doi.org/10.3233/FAIA200487.

[25] T. Chen, R. Xu, Y. He, X. Wang, Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN, Expert Syst. Appl. 72 (2017) 221–230, https://doi.org/10.1016/j.eswa.2016.10.065.

[26] M. Chernoskutov, D. Ustalov, Equidistant nodes clustering: a soft clustering algorithm applied for synset induction, in: Selected Papers of the XX International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2018), Moscow, Russia, October 9–12, 2018, vol. 2277, 2018, pp. 57–62, https://ceur-ws.org/Vol-2277/paper13.pdf, 2018, CEUR-WS.org.

[27] C.I. Chesñevar, A.G. Maguitman, M.P. González, Empowering recommendation technologies through argumentation, in: Argumentation in Artificial Intelligence, Springer, 2009, pp. 403–422, https://doi.org/10.1007/978-0-387-98197-0_20.

[28] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H.W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A.M. Dai, T.S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, N. Fiedel, Scaling language modeling with pathways, J. Mach. Learn. Res. 24 (2023) 1–113, http://jmlr.org/papers/v24/22-1144.html.

[29] K. Christakopoulou, F. Radlinski, K. Hofmann, Towards conversational recommender systems, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016, 2016, pp. 815–824, https://doi.org/10.1145/2939672.2939746.

[30] O. Cocarascu, A. Rago, F. Toni, Extracting dialogical explanations for review aggregations with argumentative dialogical agents, in: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, Montreal, QC, Canada, May 13–17, 2019, 2019, pp. 1261–1269, International Foundation for Autonomous Agents and Multiagent Systems, http://dl.acm.org/citation.cfm?id=3331830.

[31] O. Cocarascu, F. Toni, Identifying attack and support argumentative relations using deep learning, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017, 2017, pp. 1374–1379, Association for Computational Linguistics https://doi.org/10.18653/v1/d17-1144.

[32] K. Cyras, D. Letsios, R. Misener, F. Toni, Argumentation for explainable scheduling, in: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, the Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, the Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, AAAI Press, 2019, pp. 2752–2759, https://doi.org/10.1609/aaai.v33i01.33012752.

[33] K. Cyras, A. Rago, E. Albini, P. Baroni, F. Toni, Argumentative XAI: a survey, in: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event, Montreal, Canada, 19–27 August 2021, 2021, pp. 4392–4399, https://doi.org/10.24963/ijcai.2021/600.

[34] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2–7, 2019, vol. 1 (Long and Short Papers), 2019, pp. 4171–4186, Association for Computational Linguistics https://doi.org/10.18653/v1/n19-1423.

[35] G.O. Diaz, V. Ng, Modeling and prediction of online product review helpfulness: a survey, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, vol. 1: Long Papers, 2018, pp. 698–708, https://doi.org/10.18653/v1/P18-1065.

[36] P.M. Dung, On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games, Artif. Intell. 77 (1995) 321–358, https://doi.org/10.1016/0004-3702(94)00041-X.

[37] X. Fan, F. Toni, A. Mocanu, M. Williams, Dialogical two-agent decision making with assumption-based argumentation, in: International Conference on Autonomous Agents and Multi-Agent Systems, AAMAS '14, Paris, France, May 5-9, 2014, 2014, pp. 533–540, IFAAMAS/ACM, http://dl.acm.org/citation.cfm?id=2615818.

[38] S. Ferilli, A. Pazienza, S. Angelastro, A. Suglia, A similarity-based abstract argumentation approach to extractive text summarization, in: AI*IA 2017 Advances in Artificial Intelligence - XVIth International Conference of the Italian Association for Artificial Intelligence, Bari, Italy, November 14-17, 2017, Proceedings, vol. 10640, Springer, 2017, pp. 87–100, https://doi.org/10.1007/978-3-319-70169-1_7.

[39] Z. Gao, A. Feng, X. Song, X. Wu, Target-dependent sentiment classification with BERT, IEEE Access 7 (2019) 154290–154299, https://doi.org/10.1109/ACCESS.2019.2946594.

[40] A.J. García, G.R. Simari, Defeasible logic programming: an argumentative approach, Theory Pract. Log. Program. 4 (2004) 95–138, https://doi.org/10.1017/S1471068403001674.

[41] F. Gedikli, D. Jannach, M. Ge, How should I explain? A comparison of different explanation types for recommender systems, Int. J. Hum.-Comput. Stud. 72 (2014) 367–382, https://doi.org/10.1016/j.ijhcs.2013.12.007.

[42] A. Ghose, P.G. Ipeirotis, Estimating the helpfulness and economic impact of product reviews: mining text and reviewer characteristics, IEEE Trans. Knowl. Data Eng. 23 (2011) 1498–1512, https://doi.org/10.1109/TKDE.2010.188.

[43] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Comput. Surv. 51 (2019) 93:1–93:42, https://doi.org/10.1145/3236009.

[44] I. Habernal, J. Eckle-Kohler, I. Gurevych, Argumentation mining on the web from information seeking perspective, in: Proceedings of the Workshop on Frontiers and Connections Between Argumentation Theory and Natural Language Processing, Forlì-Cesena, Italy, July 21-25, 2014, 2014, CEUR-WS.org volume 1341, https://ceur-ws.org/Vol-1341/paper4.pdf.

[45] E. Hadoux, A. Hunter, J. Corrégé, Strategic dialogical argumentation using multi-criteria decision making with application to epistemic and emotional aspects of arguments, in: Foundations of Information and Knowledge Systems - 10th International Symposium, FoIKS 2018, Budapest, Hungary, May 14-18, 2018, Proceedings, vol. 10833, Springer, 2018, pp. 207–224, https://doi.org/10.1007/978-3-319-90050-6_12.

[46] A. Hunter, Understanding enthymemes in deductive argumentation using semantic distance measures, in: Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, the Twelfth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022, AAAI Press, 2022, pp. 5729–5736, https://doi.org/10.1609/aaai.v36i5.20515.

[47] G. Jain, M. Sharma, B. Agarwal, Spam detection on social media using semantic convolutional neural network, Int. J. Knowl. Discov. Bioinform. 8 (2018) 12–26, https://doi.org/10.4018/IJKDB.2018010102.

[48] N. Korfiatis, E. García-Barriocanal, S. Sánchez-Alonso, Evaluating content quality and helpfulness of online product reviews: the interplay of review helpfulness vs. review content, Electron. Commer. Res. Appl. 11 (2012) 205–217, https://doi.org/10.1016/j.elerap.2011.10.003.

[49] A. Kori, B. Glocker, F. Toni, Explaining image classifiers with visual debates, in: Discovery Science - 27th International Conference, DS 2024, Pisa, Italy, October 14-16, 2024, Proceedings, Springer, 2024, To Appear.

[50] J. Lawrence, C. Reed, Argument mining: a survey, Comput. Linguist. 45 (2019) 765–818, https://doi.org/10.1162/coli_a_00364.

[51] M. Lippi, P. Torroni, Argumentation mining: state of the art and emerging trends, ACM Trans. Internet Technol. 16 (2016) 10:1–10:25, https://doi.org/10.1145/2850417.

[52] H. Lu, W. Ma, M. Zhang, M. de Rijke, Y. Liu, S. Ma, Standing in your shoes: external assessments for personalized recommender systems, in: SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11–15, 2021, ACM, 2021, pp. 1523–1533, https://doi.org/10.1145/3404835.3462916.

[53] S. Lubos, T.N.T. Tran, A. Felfernig, S.P. Erdeniz, V. Le, LLM-generated explanations for recommender systems, in: Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization, UMAP Adjunct 2024, Cagliari, Italy, July 1-4, 2024, ACM, 2024, https://doi.org/10.1145/3631700.3665185.

[54] S.M. Lundberg, S. Lee, A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, 2017, pp. 4765–4774, https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

[55] P. Madumal, T. Miller, L. Sonenberg, F. Vetere, A grounded interaction protocol for explainable artificial intelligence, in: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, Montreal, QC, Canada, May 13–17, 2019, in: International Foundation for Autonomous Agents and Multiagent Systems, 2019, pp. 1033–1041, http://dl.acm.org/citation.cfm?id=3331801.

[56] C. Malaviya, C. Bhagavatula, A. Bosselut, Y. Choi, Commonsense knowledge base completion with structural and semantic context, in: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, the Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, the Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, AAAI Press, 2020, pp. 2925–2933, https://doi.org/10.1609/AAAI. V34I03.5684.

[57] P. McBurney, R.M. van Eijk, S. Parsons, L. Amgoud, A dialogue game protocol for agent purchase negotiations, Auton. Agents Multi-Agent Syst. 7 (2003) 235–273, https://doi.org/10.1023/A:1024787301515.

[58] H. Mercier, D. Sperber, Why do humans reason? Arguments for an argumentative theory, Behav. Brain Sci. 34 (2011) 57–74, https://doi.org/10.1017/s0140525x10000968.

[59] G. Mesnil, T. Mikolov, M. Ranzato, Y. Bengio, Ensemble of generative and discriminative techniques for sentiment analysis of movie reviews, in: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, in: Workshop Track Proceedings, 2015, http://arxiv.org/abs/1412.5335.

[60] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: Y. Bengio, Y. LeCun (Eds.), 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, in: Workshop Track Proceedings, 2013, http://arxiv.org/abs/1301.3781.

[61] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013, in: Proceedings of a Meeting Held December

5-8, 2013, Lake Tahoe, Nevada, United States, 2013, pp. 3111–3119, https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html.
[62] T. Miller, Explanation in artificial intelligence: insights from the social sciences, Artif. Intell. 267 (2019) 1–38, https://doi.org/10.1016/j.artint.2018.07.007.
[63] J. Mumford, S. Sarkadi, K. Atkinson, T.J.M. Bench-Capon, Applying argument schemes for simulating online review platforms, in: Computational Models of Argument - Proceedings of COMMA 2024, Hagen, Germany, September 18–20, 2014, vol. 388, IOS Press, 2024, pp. 169–180, https://doi.org/10.3233/FAIA240319.
[64] J. Ni, J. Li, J.J. McAuley, Justifying recommendations using distantly-labeled reviews and fine-grained aspects, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3–7, 2019, Association for Computational Linguistics, 2019, pp. 188–197, https://doi.org/10.18653/V1/D19-1018.
[65] J. Oksanen, O. Cocarascu, F. Toni, Ontology extraction and argumentation for explainable recommendations from textual reviews, in: The 1st International Workshop on Knowledge Representation for Hybrid Intelligence (KR4HI), Amsterdam, the Netherlands, June 14th, 2022, 2022, https://drive.google.com/file/d/1pYx_pEuMHcYJAGfobuo75Cd9JsEQhB5b/view.
[66] V. Parkhe, B. Biswas, Sentiment analysis of movie reviews: finding most important movie aspects using driving factors, Soft Comput. 20 (2016) 3373–3379, https://doi.org/10.1007/s00500-015-1779-1.
[67] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, Semeval-2014 task 4: aspect based sentiment analysis, in: Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014, Dublin, Ireland, August 23–24, 2014, The Association for Computer Linguistics, 2014, pp. 27–35, https://doi.org/10.3115/v1/s14-2004.
[68] N. Potyka, Continuous dynamical systems for weighted bipolar argumentation, in: Principles of Knowledge Representation and Reasoning: Proceedings of the Sixteenth International Conference, KR 2018, Tempe, Arizona, 30 October - 2 November 2018, AAAI Press, 2018, pp. 148–157, https://aaai.org/ocs/index.php/KR/KR18/paper/view/17985.
[69] H. Prakken, Coherence and flexibility in dialogue games for argumentation, J. Log. Comput. 15 (2005) 1009–1040, https://doi.org/10.1093/logcom/exi046.
[70] H. Prakken, Formal systems for persuasion dialogue, Knowl. Eng. Rev. 21 (2006) 163–188, https://doi.org/10.1017/S0269888906000865.
[71] A. Rago, O. Cocarascu, C. Bechlivanidis, D.A. Lagnado, F. Toni, Argumentative explanations for interactive recommendations, Artif. Intell. 296 (2021) 103506, https://doi.org/10.1016/j.artint.2021.103506.
[72] A. Rago, O. Cocarascu, F. Toni, Argumentation-based recommendations: fantastic explanations and how to find them, in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI, 2018, July 13-19, 2018, Stockholm, Sweden, 2018, pp. 1949–1955, ijcai.org.
[73] A. Rago, H. Li, F. Toni, Interactive explanations by conflict resolution via argumentative exchanges, in: Proceedings of the 20th International Conference on Principles of Knowledge Representation and Reasoning, KR 2023, Rhodes, Greece, September 2-8, 2023, 2023, pp. 582–592.
[74] A. Rago, F. Toni, Quantitative argumentation debates with votes for opinion polling, in: PRIMA 2017: Principles and Practice of Multi-Agent Systems - 20th International Conference, Nice, France, October 30 - November 3, 2017, Proceedings, vol. 10621, 2017, pp. 369–385, https://doi.org/10.1007/978-3-319-69131-2_22.
[75] A. Rago, F. Toni, M. Aurisicchio, P. Baroni, Discontinuity-free decision support with quantitative argumentation debates, in: Principles of Knowledge Representation and Reasoning: Proceedings of the Fifteenth International Conference, KR 2016, Cape Town, South Africa, April 25-29, 2016, AAAI Press, 2016, pp. 63–73, http://www.aaai.org/ocs/index.php/KR/KR16/paper/view/12874.
[76] B. Rahdari, H. Ding, Z. Fan, Y. Ma, Z. Chen, A. Deoras, B. Kveton, Logic-scaffolding: personalized aspect-instructed recommendation explanation generation using LLMs, in: Proceedings of the 17th ACM International Conference on Web Search and Data Mining, WSDM 2024, Merida, Mexico, March 4-8, 2024, ACM, 2024, pp. 1078–1081, https://doi.org/10.1145/3616855.3635689.
[77] M.T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?": explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016, ACM, 2016, pp. 1135–1144, https://doi.org/10.1145/2939672.2939778.
[78] S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, E.M. Smith, Y. Boureau, J. Weston, Recipes for building an open-domain chatbot, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021, 2021, pp. 300–325, Association for Computational Linguistics, https://doi.org/10.18653/v1/2021.eacl-main.24.
[79] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nat. Mach. Intell. 1 (2019) 206–215, https://doi.org/10.1038/s42256-019-0048-x.
[80] I. Sassoon, N. Kökciyan, E. Sklar, S. Parsons, Explainable argumentation for wellness consultation, in: Explainable, Transparent Autonomous Agents and Multi-Agent Systems - First International, Workshop, EXTRAAMAS 2019, Montreal, QC, Canada, May 13–14, 2019, Revised Selected Papers, vol. 11763, 2019, pp. 186–202.
[81] N. Slonim, Y. Bilu, C. Alzate, R. Bar-Haim, B. Bogin, F. Bonin, L. Choshen, E. Cohen-Karlik, L. Dankin, L. Edelstein, L. Ein-Dor, R. Friedman-Melamed, A. Gavron, A. Gera, M. Gleize, S. Gretz, D. Gutfreund, A. Halfon, D. Hershcovich, R. Hoory, Y. Hou, S. Hummel, M. Jacovi, C. Jochim, Y. Kantor, Y. Katz, D. Konopnicki, Z. Kons, L. Kotlerman, D. Krieger, D. Lahav, T. Lavee, R. Levy, N. Liberman, Y. Mass, A. Menczel, S. Mirkin, G. Moshkowich, S. Ofek-Koifman, M. Orbach, E. Rabinovich, R. Rinott, S. Shechtman, D. Sheinwald, E. Shnarch, I. Shnayderman, A. Soffer, A. Spector, B. Sznajder, A. Toledo, O. Toledo-Ronen, E. Venezian, R. Aharonov, An autonomous debating system, Nature 591 (2021) 379–384, https://doi.org/10.1038/s41586-021-03215-w.
[82] S. Sreedharan, T. Chakraborti, S. Kambhampati, Foundations of explanations as model reconciliation, Artif. Intell. 301 (2021) 103558, https://doi.org/10.1016/j.artint.2021.103558.
[83] P. Sukpanichnant, A. Rago, P. Lertvittayakumjorn, F. Toni, Neural qbafs: Explaining Neural Networks Under lrp-Based Argumentation Frameworks, AIxIA 2021 - Advances in Artificial Intelligence - 20th International Conference of the Italian Association for Artificial Intelligence, Virtual Event, December 1-3, 2021, Revised Selected Papers, vol. 13196, Springer, 2021, pp. 429–444, https://doi.org/10.1007/978-3-031-08421-8_30.
[84] D. Tang, B. Qin, T. Liu, Document modeling with gated recurrent neural network for sentiment classification, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015, in: The Association for Computational Linguistics, 2015, pp. 1422–1432, https://doi.org/10.18653/v1/d15-1167.
[85] J.C. Teze, S. Gottifredi, A.J. García, G.R. Simari, Improving argumentation-based recommender systems through context-adaptable selection criteria, Expert Syst. Appl. 42 (2015) 8243–8258, https://doi.org/10.1016/j.eswa.2015.06.048.
[86] L. Thorburn, A. Kruger, Optimizing language models for argumentative reasoning, in: Proceedings of the 1st Workshop on Argumentation & Machine Learning Co-Located with 9th International Conference on Computational Models of Argument, COMMA 2022, Cardiff, Wales, September 13th, 2022, vol. 3208, 2022, pp. 27–44, CEUR-WS.org, https://ceur-ws.org/Vol-3208/paper3.pdf.
[87] K. Tirumala, A.H. Markosyan, L. Zettlemoyer, A. Aghajanyan, Memorization without overfitting: analyzing the training dynamics of large language models, in: Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, 2022, http://papers.nips.cc/paper_files/paper/2022/hash/fa0509f4dab6807e2cb465715bf2d249-Abstract-Conference.html, 2022.
[88] A. Tripathy, A. Agrawal, S.K. Rath, Classification of sentiment reviews using n-gram machine learning approach, Expert Syst. Appl. 57 (2016) 117–126, https://doi.org/10.1016/j.eswa.2016.03.028.
[89] M. Turpin, J. Michael, E. Perez, S.R. Bowman, Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting, in: Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023, http://papers.nips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract-Conference.html.

[90] S.L. Vasileiou, A. Kumar, W. Yeoh, T.C. Son, F. Toni, Dialectical reconciliation via structured argumentative dialogues, in: Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning, KR 2024, Hanoi, Vietnam, November 2–8, 2024, 2024.

[91] A. Vassiliades, N. Bassiliades, T. Patkos, Argumentation and explainable artificial intelligence: a survey, Knowl. Eng. Rev. 36 (2021) e5, https://doi.org/10.1017/S0269888921000011.

[92] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: automated decisions and the gdpr, Harv. J. Law Technol. 31 (2017) 841, https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/hjlt31&section=29.

[93] D. Walton, C. Reed, F. Macagno, Argumentation Schemes, Cambridge University Press, 2008, http://www.cambridge.org/us/academic/subjects/philosophy/logic/argumentation-schemes.

[94] P.F. Wu, H. van der Heijden, N. Korfiatis, The influences of negativity and review quality on the helpfulness of online reviews, in: Proceedings of the International Conference on Information Systems, ICIS 2011, Shanghai, China, December 4-7, 2011, Association for Information Systems, 2011, http://aisel.aisnet.org/icis2011/proceedings/ebusiness/1.

[95] A.Z. Wyner, J. Schneider, K. Atkinson, T.J.M. Bench-Capon, Semi-automated argumentative analysis of online product reviews, in: Computational Models of Argument - Proceedings of COMMA 2012, Vienna, Austria, September 10–12, 2012, vol. 245, IOS Press, 2012, pp. 43–50, https://doi.org/10.3233/978-1-61499-111-3-43.

[96] H. Ye, N. Zhang, H. Chen, H. Chen, Generative knowledge graph construction: a review, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022, Association for Computational Linguistics, 2022, pp. 1–17, https://doi.org/10.18653/v1/2022.emnlp-main.1.

[97] A. Zenebe, A.F. Norcio, Representation, similarity measures and aggregation methods using fuzzy sets for content-based recommender systems, Fuzzy Sets Syst. 160 (2009) 76–94, https://doi.org/10.1016/j.fss.2008.03.017.

[98] Y. Zhang, X. Chen, Q. Ai, L. Yang, W.B. Croft, Towards conversational search and recommendation: system ask, user respond, in: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018, ACM, 2018, pp. 177–186.

[99] Y. Zhang, G. Lai, M. Zhang, Y. Zhang, Y. Liu, S. Ma, Explicit factor models for explainable recommendation based on phrase-level sentiment analysis, in: The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast, QLD, Australia - July 06 - 11, 2014, ACM, 2014, pp. 83–92.