Explanation via Machine Arguing

Oana Cocarascu, Antonio Rago, and Francesca Toni[✉]

Imperial College London, London, UK
{oc511,a.rago15,ft}@imperial.ac.uk

Abstract. As AI becomes ever more ubiquitous in our everyday lives, its ability to explain to and interact with humans is evolving into a critical research area. Explainable AI (XAI) has therefore emerged as a popular topic but its research landscape is currently very fragmented. Explanations in the literature have generally been aimed at addressing individual challenges and are often ad-hoc, tailored to specific AIs and/or narrow settings. Further, the extraction of explanations is no simple task; the design of the explanations must be fit for purpose, with considerations including, but not limited to: Is the model or a result being explained? Is the explanation suited to skilled or unskilled explainees? By which means is the information best exhibited? How may users interact with the explanation? As these considerations rise in number, it quickly becomes clear that a systematic way to obtain a variety of explanations for a variety of users and interactions is much needed. In this tutorial we will overview recent approaches showing how these challenges can be addressed by utilising forms of *machine arguing* as the scaffolding underpinning explanations that are delivered to users. Machine arguing amounts to the deployment of methods from *computational argumentation* in AI with suitably mined *argumentation frameworks*, which provide abstractions of "debates". Computational argumentation has been widely used to support applications requiring information exchange between AI systems and users , facilitated by the fact that the capability of arguing is pervasive in human affairs and arguing is core to a multitude of human activities: humans argue to explain, interact and exchange information. Our lecture will focus on how machine arguing can serve as the driving force of explanations in AI in different ways, namely: by building explainable systems with argumentative foundations from linguistic data focusing on reviews), or by extracting argumentative reasoning from existing systems (focusing on a recommender system).

Keywords: Argumentation · Explanation · Explainable AI

1 Introduction

Much of AI researchers' recent efforts are being dedicated towards the extraction of explanations for results by AI tools (e.g. predictions, classifications or recommendations) and the manner in which they are provided to users (e.g. see [19,21] for recent overviews). However, the extraction of explanations is no

M. Manna and A. Pieris (Eds.): Reasoning Web 2020, LNCS 12258, pp. 53–84, 2020.
https://doi.org/10.1007/978-3-030-60067-9_3

simple task; the design of the explanations must be fit for purpose, with considerations including, but not limited to: Is the model or a result being explained? Is the explanation suited to skilled or unskilled explainees? By which means is the information best exhibited? How may users interact with the explanation? As these considerations rise in number, it quickly becomes clear that a systematic way to obtain a variety of explanations for a variety of users and interactions, rather than a multitude of ad-hoc approaches lacking coherence with one another, would simplify the scope of the problem significantly, while also providing a unifying view within the research landscape. At the same time, AI systems are advancing towards providing *interactive* explanations to users. Such systems have become extremely popular in recent years; huge investments from the tech industry leaders have been devoted to developing personal assistant technology, e.g. Microsoft's *Cortana*. However, in order for these products to engage in meaningful explanatory conversations with humans they must be backed by explanatory capabilities to drive interactions.

Computational argumentation (e.g. see [2,5] for recent overviews) is a well-established field in AI focusing on the definition of so-called argumentation frameworks as abstractions of "debates" and the evaluation of the dialectical standing of positions (arguments) within these debates: for example, *abstract argumentation frameworks* [16] represent disagreements (attacks) in debates, whereas *bipolar argumentation frameworks* [9,13] represent agreements (supports) as well as disagreements . Computational argumentation has long been identified as a suitable mechanism to support explanation (e.g. as in [8,22]). Indeed, the capability of arguing is pervasive in human affairs: it is essential in certain professions, e.g. traditionally, the practice of law and politics, but also in evidence-based medicine, where evidence in favour of or against treatments is essential for decision-making. Overall, arguing is core to a multitude of human activities: humans argue to explain, interact and exchange information. As a result, various models of interactive explanation may be supported by argumentation frameworks as the underlying knowledge base. Further, given that argumentation is amenable for human consumption, it can effectively support the human desire to anthropomorphise systems.

Explainable AI can be supported by very many forms of computational argumentation, including the aforementioned abstract (with just attacks) and bipolar (with both attacks and supports) argumentation, several forms of *structured argumentation* and *quantitative bipolar argumentation* (see [7] for an overview). In some (notably abstract, bipolar and quantitative bipolar argumentation) arguments are seen as abstract entities, which can be deemed to be arguments as they are connected by dialectical relations of attack (in all cases) and possibly of support (in bipolar and quantitative bipolar argumentation). We will take this view of arguments as 'abstract' also in this lecture. Other relations are envisaged as possible in some other forms of computational argumentation, e.g. in [17]: in this lecture we will use, in addition to quantitative bipolar argumentation, also *tripolar argumentation* [23], extending bipolar argumentation with a third relation of *neutralisation*.

Computational argumentation is not just about representing arguments and relationships between them: reasoning with the represented arguments to extract semantically justified conclusion is also core to this paradigm, across all its forms. Whereas abstract argumentation, structured argumentation and bipolar argumentation predominantly use *extensions* (i.e. sets of arguments) as their semantics, quantitative bipolar argumentation uses *gradual semantics*: extensions identify sets of arguments that are collectively acceptable, from a dialectical point of view; instead, gradual semantics assign, from within a given set of values, dialectical strengths to individual arguments. In this lecture, we will make use of the second form of semantics.

Can computational argumentation support the vision of explainable AI? In order to do so, argumentation frameworks need to be extracted from the systems that need explaining. If the systems are built from scratch then argumentation for explanation can be injected within them by design, but this still requires a principled extraction from the systems' components. Moreover, the extraction of interactive explanations of various types from the extracted argumentation frameworks needs engineering. We think of the end-to-end process of extraction of argumentation framework and interactive explanations as *machine arguing*, allowing machines to engage with humans in the process of argumentation, and deployable as the scaffolding in which relevant information can be harboured so that a variety of interactive explanatory exchanges for AI systems' outputs can be extracted for various types of users. This is summarised in Fig. 1, where the AI system, in principle, may be any, based on data-centric, symbolic or hybrid methodologies. In this paper, we focus on recommender systems.

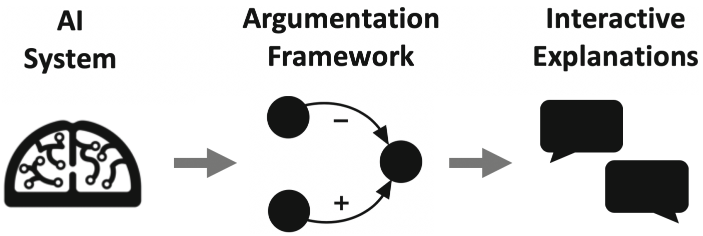


Fig. 1. Machine arguing for interactive explanations.

In this lecture we focus on two examples of this vision of machine arguing for systems' explanation: in Sect. 3 we define an explainable systems from linguistic data (i.e. reviews), so that it has argumentative foundations to support a variety of interactive explanations; and in Sect. 4 we show how interactive explanations can be extracted by argumentative reasoning from a given, non argumentative (recommender) system. In both cases, as in Fig. 1, argumentation frameworks are extracted from the underlying systems: these are quantitative bipolar argumentation frameworks in Sect. 3 and tripolar argumentation fameworks in Sect. 4.

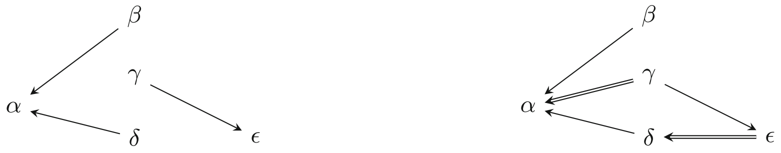


Fig. 2. The argument graphs $\mathcal{G}_F$ for the AF $F = (\{\alpha, \beta, \gamma, \delta, \epsilon\}, \{(\beta, \alpha), (\delta, \alpha), (\gamma, \epsilon)\})$ (left) and $\mathcal{G}_F$ for the BF $F = (\{\alpha, \beta, \gamma, \delta, \epsilon\}, \{(\beta, \alpha), (\delta, \alpha), (\gamma, \epsilon)\}, \{(\gamma, \alpha), (\epsilon, \delta)\})$ (right).

Then, gradual semantics are applied to the frameworks to support the generation of interactive explanations. In Sect. 3 the gradual semantics is chosen by us (by design), and alternative gradual semantics can be applied. Instead, in Sect. 4 the semantics is dictated by the underlying recommender system, so as to match its predicted ratings, and the extraction of the argumentation framework from which explanations are drawn is regulated by the need for this semantics to be dialectically meaningful.

2 Background: Argumentation

Abstract Argumentation frameworks (AFs) are pairs consisting of a set of arguments and a binary (attack) relation between arguments [16]. Formally, an AF is any pair $\langle \mathcal{X}, \mathcal{L}^- \rangle$ where $\mathcal{L}^- \subseteq \mathcal{X} \times \mathcal{X}$. Bipolar Argumentation frameworks (BFs) extend AFs by considering two binary relations: attack and support [9]. Formally, a BF is any triple $(\mathcal{X}, \mathcal{L}^-, \mathcal{L}^+)$ where $\langle \mathcal{X}, \mathcal{L}^- \rangle$ is an AF and $\mathcal{L}^+ \subseteq \mathcal{X} \times \mathcal{X}$. If $\mathcal{L}^+ = \emptyset$, a BF $(\mathcal{X}, \mathcal{L}^-, \mathcal{L}^+)$ can be identified with an AF $\langle \mathcal{X}, \mathcal{L}^- \rangle$, so we can use the term BF to denote BFs as well as AFs.

Any $F = (\mathcal{X}, \mathcal{L}^-, \mathcal{L}^+)$ can be understood and visualised as a directed graph $\mathcal{G}_F$, also called *argument graph*, with nodes $\mathcal{X}$ and two types of edges: $\mathcal{L}^-$ and $\mathcal{L}^+$ (see e.g. [9,14]). In the illustration in Fig. 2, we show $\mathcal{G}_F$ using single ($\rightarrow$) and double ($\Rightarrow$) arrows to denote $\mathcal{L}^-$ and $\mathcal{L}^+$, respectively. In the remainder of the paper, instead, when showing $\mathcal{G}_F$, we will use arrows labelled - to denote $\mathcal{L}^-$ and arrows labelled + to denote $\mathcal{L}^+$, respectively.[1]

Semantics of AFs/BFs amount to "recipes" for determining "winning" sets of arguments or the "dialectical strength" of arguments. These semantics can be respectively defined qualitatively, in terms of *extensions* (e.g. the *grounded* extension [16], defined below), and quantitatively, in terms of a *gradual evaluation* of arguments (e.g. as in [6,7,25] – the former of which, defined below, we will use in this paper).

Given an AF $\langle \mathcal{X}, \mathcal{L}^- \rangle$, let $E \subseteq \mathcal{X}$ *defend* $a \in \mathcal{X}$ iff for all $b \in \mathcal{X}$ attacking a there exists $c \in E$ attacking b. Then, the *grounded extension* of $\langle \mathcal{X}, \mathcal{L}^- \rangle$ is $G = \bigcup_{i \geq 0} G_i$, where G_0 is the set of all *unattacked* arguments (i.e. the set of all arguments $a \in \mathcal{X}$ such that there is no argument $b \in \mathcal{X}$ with $(b, a) \in \mathcal{L}^-$)

[1] These alternative notations are used interchangeably in the literature, as we do here.

and $\forall i \geq 0$, G_{i+1} is the set of all arguments that G_i defends. For any $\langle \mathcal{X}, \mathcal{L}^- \rangle$, the grounded extension G always exists and is unique. As an illustration, in the simple AF in Fig. 2, left, $G = \{\beta, \delta, \gamma\}$.

On the other hand, quantitative semantics allow a *gradual evaluation* of arguments. They can be defined for BFs, as in [4], or for *Quantitative Bipolar Argumentation Frameworks* (QBFs) [6,7], of the form $(\mathcal{X}, \mathcal{L}^-, \mathcal{L}^+, \tau)$ where $(\mathcal{X}, \mathcal{L}^-, \mathcal{L}^+)$ is a BF and $\tau : \mathcal{X} \rightarrow I$ for some interval I (e.g. $I = [0,1]$ or $I = [-1,1]$) gives the intrinsic strength or *base score* of arguments. AFs and BFs are QBFs with special choices of τ [6], so we will sometimes use the term QBF to denote AFs and BFs. The argument graph for $\langle \mathcal{X}, \mathcal{L}^-, \mathcal{L}^+, \tau \rangle$ is the argument graph for $(\mathcal{X}, \mathcal{L}^-, \mathcal{L}^+)$.

Given a QBF $(\mathcal{X}, \mathcal{L}^-, \mathcal{L}^+, \tau)$, the strength of arguments is given by some $\sigma : \mathcal{X} \rightarrow I$. Several such notions have been defined in the literature (e.g. see [7] for an overview). We will use the notion of [25][2], where $I = [0,1]$ and for $a \in \mathcal{X}$:

$$\sigma(a) = c(\tau(a), \mathcal{F}'(\sigma(\mathcal{L}^-(a))), \mathcal{F}'(\sigma(\mathcal{L}^+(a))))$$

such that:

(i) $\mathcal{L}^-(a)$ is the set of all arguments attacking a and if $(a_1, \ldots, a_n)$ is an arbitrary permutation of the $(n \geq 0)$ elements of $\mathcal{L}^-(a)$, then $\sigma(\mathcal{L}^-(a)) = (\sigma(a_1), \ldots, \sigma(a_n))$ (similarly for supporters);

(ii) for $v_0, v_a, v_s \in [0,1]$,

$$c(v_0, v_a, v_s) = v_0 - v_0 \cdot |v_s - v_a| \quad \text{if } v_a \geq v_s,$$
$$c(v_0, v_a, v_s) = v_0 + (1 - v_0) \cdot |v_s - v_a| \quad \text{if } v_a < v_s; \text{ and}$$

(iii) for $S = (v_1, \ldots, v_n) \in [0,1]^*$ and $f'(x,y) = x + y - x \cdot y$:
if $n = 0$: $\mathcal{F}'(S) = 0$; if $n = 1$: $\mathcal{F}'(S) = v_1$; if $n = 2$: $\mathcal{F}'(S) = f'(v_1, v_2)$;
if $n > 2$: $\mathcal{F}'(S) = f'(\mathcal{F}'(v_1, \ldots, v_{n-1}), v_n)$.

Intuitively, the strength $\sigma(a)$ of argument a results from the combination c of three components: the base score $\tau(a)$ of a, the aggregated strength $\mathcal{F}'(\sigma(\mathcal{L}^-(a)))$ of all arguments attacking a and the aggregated strength $\mathcal{F}'(\sigma(\mathcal{L}^+(a)))$ of all arguments supporting a. The combination c decreases the base score of a if the aggregated strength of the attackers is at least as high as the aggregated strength of the supporters (with the decrement proportional to the base score and to the absolute value of the difference between the aggregated strengths). The combination c increases the base score of a otherwise, if the aggregated strength of the attackers is lower than the aggregated strength of the supporters (with the increment proportional to the distance between 1 and the base score and to the absolute value of the difference between the aggregated strengths). Finally, the aggregated strengths are defined recursively (using the probabilistic sum when there are exactly two terms to aggregate - these are either strengths of attackers or of supporters).[3]

[2] Note that several other notions could be used, as overviewed in [7]. we have chosen this specific notion because it satisfies some desirable properties [7] as well as performing well in practice [11].

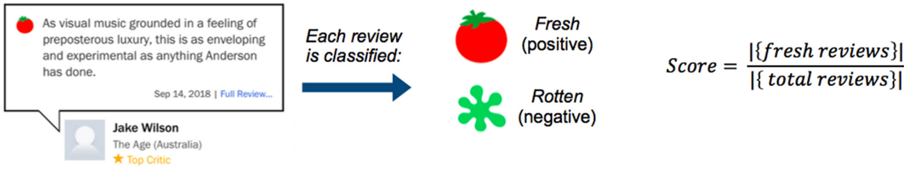
[3] Note that this recursively defined notion treats strengths of attackers and supporters as sets, but needs to consider them in sequence (thus the mention of ‘an arbitrary permutation’).

As an illustration, in the BF in Fig. 2, right, if the base score of all arguments if 0.5, then $\sigma(\gamma) = \tau(\gamma) = 0.5$ and $\sigma(\epsilon) = c(0.5, 0.5, 0) = 0.5 - 0.5 \cdot 0.5 = 0.25$.

3 Building Explainable Systems with Argumentative Foundations

The use of social media has become a regular habit for many and has changed the way people interact with each other. In an age in which e-commerce and audio/video streaming are dominant markets for consumers, products' online reviews are fast becoming the preferred method of quality control for users. The aggregation of these reviews allows users to check the quality of a product while avoiding reviews which may be incoherent and irrelevant.

Within the movie domain, Rotten Tomatoes[4] (RT) is a popular review site that aggregates critics' reviews to obtain an overall percentage of critics who like the movie and critics who do not. The RT score is simplified to a binary classification for the movie of Fresh or Rotten, based on whether it is greater or equal to 60% or not, respectively. This simplification into RT score, Fresh/Rotten classification, gives users a quick way to determine whether a movie is worth watching or not. Figure 3 shows an overview of RT where each review is classified as Fresh/Rotten and the score of the movie is given by the percentage of Fresh reviews.



As visual music grounded in a feeling of preposterous luxury, this is as enveloping and experimental as anything Anderson has done.

Sep 14, 2018 | Full Review...

Jake Wilson
The Age (Australia)
★ Top Critic

$$Score = \frac{|\{fresh\ reviews\}|}{|\{\ total\ reviews\}|}$$

Fig. 3. Rotten Tomatoes summary.

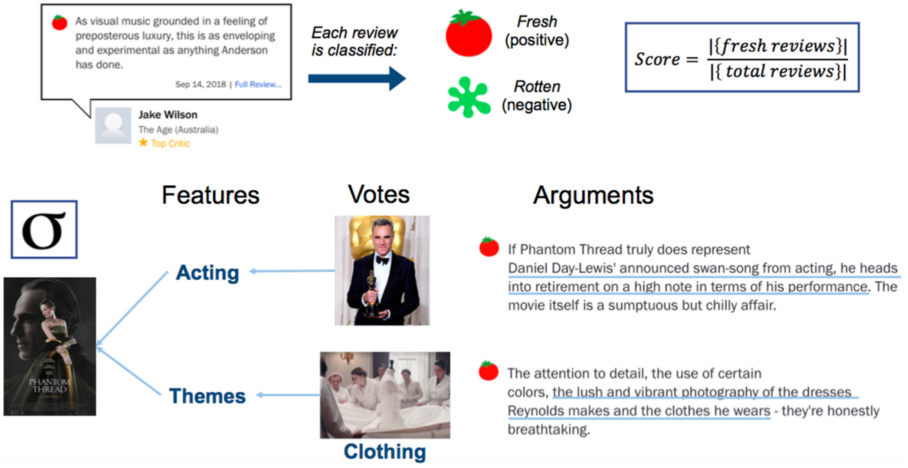
However, the 60% threshold means that a critic's mixed review that is slightly positive overall will have the same weight as a rave review from another critic, leading to the case where a movie with a maximum RT score could be composed of only *generally* positive reviews. Also, the RT score does not take into account user preferences and so factors which decrease the RT score 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. Thus, a method to *explain* the aggregation is needed so that users can decide for themselves.

To address this problem, in this section we present a method [11] that, for any given movie:

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

1. mines arguments from reviews;
2. extracts a QBAF and computes a dialectical strength measure as an alternative to the RT score;
3. supplements the computed strength with dialogical explanations obtained from BAFs extracted from movie reviews that empower users to interact with the system for more information about a movie's aggregated review.

The method relies upon a feature-based characterisation of reviews and mines review aggregations (RAs) from snippets drawn from reviews by critics to obtain votes (on both movies and their (sub-)features). We mine RAs from snippets and determine whether these snippets provide positive or negative votes for ((sub-)features of) movies, by looking for arguments supporting or attacking, respectively, the ((sub-)features of the) movies. The feature-characterisation along with the votes are then used to generate QBAFs, that can then provide a dialectical strength, σ of the movi.e. Our method aims to extract a QBAF for any given movie and provide a dialectical strength, σ for the movie an an alternative to the score that appears on the RT website, as can be seen in Fig. 4.



As visual music grounded in a feeling of preposterous luxury, this is as enveloping and experimental as anything Anderson has done.

Sep 14, 2018 | Full Review...

Jake Wilson
The Age (Australia)
★ Top Critic

$$Score = \frac{|\{fresh\ reviews\}|}{|\{\ total\ reviews\}|}$$

If Phantom Thread truly does represent Daniel Day-Lewis' announced swan-song from acting, he heads into retirement on a high note in terms of his performance. The movie itself is a sumptuous but chilly affair.

The attention to detail, the use of certain colors, the lush and vibrant photography of the dresses Reynolds makes and the clothes he wears - they're honestly breathtaking.

Fig. 4. The method [11] aims to match the RT score with σ computed from the QBAF extracted from the movie reviews.

3.1 Preprocessing

Definition 1. *Let $\mathcal{M}$ be a given set of movies, and $m \in \mathcal{M}$ be any movi.e. A feature-based characterisation of m is a finite set $\mathcal{F}$ of features with sub-features $\mathcal{F}' \subset \mathcal{F}$ such that each $f' \in \mathcal{F}'$ has a unique parent $p(f') \in \mathcal{F}$; for any $f \in \mathcal{F} \backslash \mathcal{F}'$, we define $p(f) = m$.*

Feature-Based Characterisation of Movies: A sub-feature is more specific than its parent feature. For example, for the movie m = *Wonder Wheel*, a feature may be *acting*, which may be the parent of the sub-feature *Kate Winslet*. We will refer to elements of $\mathcal{F} \backslash \mathcal{F}'$ only as features, and to elements of $\mathcal{F}'$ as sub-features. Also, we will refer to a sub-feature with parent f as a sub-feature of f.

Note that this feature-based characterisation may be obtained automatically from metadata and from the top critics' snippets that appear on the RT movie pages. By doing so, for *Wonder Wheel*, we may obtain 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*.

The sub-features in $\mathcal{F}'$ may be of different types, namely *single* (for features f_D and f_W, if we only consider movies with a single director or writer) or *multiple* (for f_A, since movies will generally have more than one actor: *Wonder Wheel* has *Kate Winslet* and *Justin Timberlake* as sub-features of f_A, and f_T, since movies will generally be associated with several themes). In the case of single sub-features, the feature can be equated with the sub-feature (for *Wonder Wheel*, *Woody Allen* is the sole director and so this sub-feature can be represented by f_D itself). Furthermore, sub-features may be *predetermined*, namely obtained from meta-data (as for the sub-features with parents f_A, f_D, f_W in the running example), or *mined* from (snippets of) reviews (for *Wonder Wheel* the sub-feature *amusement park* of f_T may be mined rather than predetermined). For example, for *Wonder Wheel*, we identify the sub-feature *amusement park* (f'_{T1}) as several reviews mention the related terms *Coney Island* and *fairground*, as in *'like the fairground ride for which it's named, Wonder Wheel is entertaining'*.

The movie *Wonder Wheel*, its (sub-)features and how they relate are shown in Fig. 5.

Extracting Phrases Representing Potential Arguments: We analyse each critic's review independently, tokenising each review into sentences and splitting sentences into phrases when specific keywords (*but, although, though, otherwise, however, unless, whereas, despite*) occur. Each phrase may then constitute a potential argument with vote from its critic in the review aggregation.

For illustration, consider the following review for m = *Wonder Wheel* from a critic:
c_1: *Despite a stunning performance by Winslet and some beautiful cinematography by Vittorio Storaro, Wonder Wheel loses its charms quickly and you'll soon be begging to get off this particular ride.*
We extract two phrases:

- p_1: *Despite a stunning performance by Winslet and some beautiful cinematography by Vittorio Storaro*
- p_2: *Wonder Wheel loses its charms quickly and you'll soon be begging to get off this particular ride*

Consider another review from a critic:
c_2: *Like the fairground ride for which it's named, it is entertaining.*

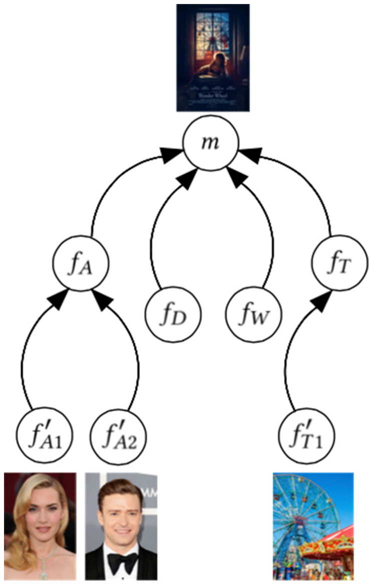


Fig. 5. Extracted (sub-)features for the movie *Wonder Wheel.*

We extract two phrases:

- p_3: *like the fairground ride for which it's named, the film is entertaining*

Finally, consider the following review from a critic:
c_3: *As we watch Allen worry and nitpick over the way women fret over aging, painting Ginny as pathetic, jealous, insecure, and clownish, it's dull, unoriginal, and offensive. Frankly, we've had enough Woody Allen takes on this subject.*

Here we extract two different phrases concerning *Woody Allen*:

- p_4: *As we watch Allen worry and nitpick over the way women fret over aging, painting Ginny as pathetic, jealous, insecure, and clownish, it's dull, unoriginal, and offensive*
- p_5: *Frankly, we've had enough Woody Allen takes on this subject.*

Using this feature-based characterisation of a movie and snippets from the movie reviews by critics, we generate votes on arguments, amounting to the movie in question and its (sub-)features. The result is a *review aggregation* for the movie, which is then transformed into a *QBAF* that we obtain as described next.

Extracting Review Aggregations: Let $m \in \mathcal{M}$ be any movie and $\mathcal{F}$ be a feature-based characterisation of m as given in Definition 1. Let $\mathcal{X}$ denote $\{m\} \cup \mathcal{F}$, referred to as the set of *arguments*. We then define the following:

Definition 2. *A review aggregation for* m *is a triple* $\mathcal{R}(m) = \langle \mathcal{F}, \mathcal{C}, \mathcal{V} \rangle$ *where*

- $\mathcal{C}$ *is a finite, non-empty set of critics;*
- $\mathcal{V} : \mathcal{C} \times \mathcal{X} \rightarrow \{-, +\}$ *is a partial function, with* $\mathcal{V}(c, \alpha)$ *representing the vote of critic* $c \in \mathcal{C}$ *on argument* $\alpha \in \mathcal{X}$.

A positive/negative vote from a critic on a (sub-)feature of the movie signifies positive/negative stance on that (sub-)feature and a positive/negative vote on m signifies positive/negative stance on the overall movi.e.

In order to determine the arguments on which the votes act, we use a *glossary* G using movie-related words for each feature as well as for movies in general. G is as follows (for any $m \in \mathcal{M}$):
$G(m) = \{movie, film, work\}$;
$G(f_D) = \{director\}$;
$G(f_A) = \{acting, cast, portrayal, performance\}$;
$G(f_W) = \{writer, writing, screenplay, screenwriter, screenwriting, script, storyline, character\}$.

We can mine votes using NLP techniques such as sentiment analysis or argument mining. For simplicity, we will use sentiment analysis, which is the process of computationally identifying and categorising opinions expressed in a piece of text to determine whether the polarity towards a particular topic, item, etc. is positive, negative, or neutral. The sentiment polarity of each phrase is translated into a (negative or positive) vote from the corresponding critic. Furthermore, 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 arguments based on occurrences of words from G.

When determining the argument on which a vote acts, sub-features take precedence over features. A mention of "Kate Winslet" (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 sub-feature or a word from the glossary) corresponding to different (sub-)features results in two arguments (and votes), one for each (sub-)feature identified.

For example, from the review from c_1, the system may extract one vote for the sub-feature "Kate Winslet" (f'_{A1}) and one for the movie in general. Thus, p_1 gives $(0.833, f'_{A1})$ therefore $\mathcal{V}(c_1, f'_{A1}) = +$, while p_2 gives $(-0.604, m)$ therefore $\mathcal{V}(c_1, m) = -$. If the neutrality threshold is 0.6 for the absolute value of the polarity, a positive vote corresponding to p_1 is assigned to f'_{A1} and a negative vote corresponding to p_2 is assigned to m. It should be noted that if a feature *cinematography* had been included in our $\mathcal{F}$ then we would have had another vote from c_1. This could be achieved by using more metadata of the movies and hence an occurrence of Storaro would correspond to a vote on cinematography. We determine the votes for the mined f_T in the same way as for the other features. For example, given p_3: *like the fairground ride for which it's named, Wonder Wheel is entertaining* leading to $(0.741, f'_{T1})$, we obtain $\mathcal{V}(c_2, f'_{T1}) = +$. If the review of a single critic results in several phrases associated with an argument with different polarities, we take the one with the highest sentiment magnitude to determine the vote on that argument. For example, given: p_5: $(-0.659,\ f_D)$

and p_6: $(-0.500, f_D)$, then p_5 supersedes and $\mathcal{V}(c_3, f_D) = -$. Figure 6 shows the votes extracted on the movie *Wonder Wheel* and its (sub-)features.

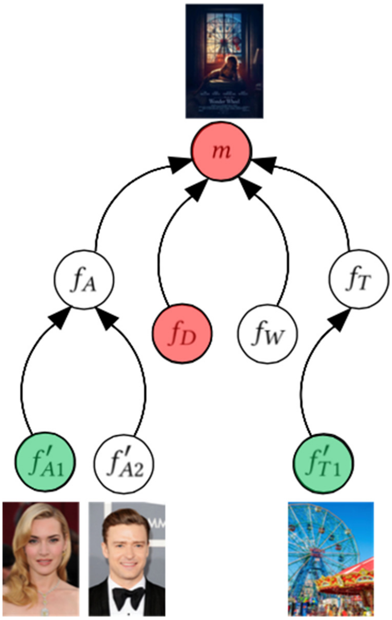


Fig. 6. Votes extracted on the movie and (sub-)features.

Augmenting Review Aggregations: A review aggregation can be *augmented* by exploiting the parent relation: a vote for/against an argument can be seen as a vote for/against the argument's parent.

In the case of f_A, which is a *multiple* feature, we also consider the augmented vote (if any) when determining whether the movie has augmented votes. This is because a movie generally has several actors, whose performances may differ. In the case of f_D and f_W, if we only consider movies with a single director or writer, the features are equated with the sub-feature. If the movie has more than one director or writer, then their augmented votes are not considered when determining the augmented votes of the movie as the contributions in directing and writing cannot be split (i.e. we cannot have director X was better than director Y for the same movie). While being a *multiple* feature, we do not consider the augmented votes of themes for the augmented votes of movies as we mine themes from texts and movies may not have themes. The importance of acting in the augmentation is also due to the fact that it appears in movie metadata, whereas themes do not.

For example, let c's vote on f_A be undefined, c's vote on sub-feature f'_{A1} of f_A be + and there be no − votes from c on any other sub-features of f_A. We then assume that c's overall stance on f_A is positive and therefore set c's vote on f_A to +. This notion of augmented review aggregation combats the brevity of the snippets causing the review aggregation being too sparsely populated.

In our example, given that f'_{A1} is positive and f'_{T1} is positive, then f_A and f_T are augmented each with a positive vote. Figure 7 shows the augmented review aggregations for the movie *Wonder Wheel*.

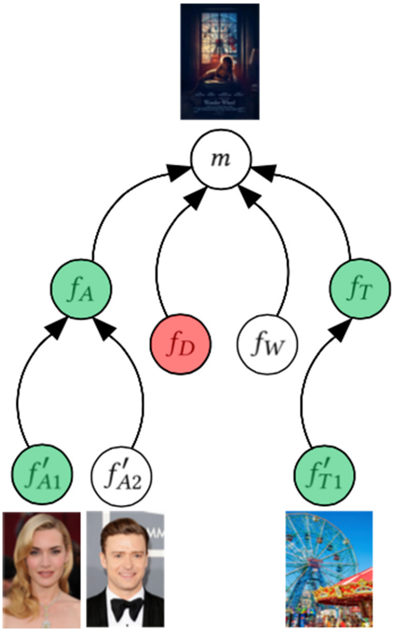


Fig. 7. Augmented review aggregations for the movie *Wonder Wheel*.

3.2 Extracting QBAFs

In order to obtain a QBAF from a review aggregation, we determine: the arguments, the arguments' base scores, and between which arguments attacks and supports are present. Having already identified the arguments, we use an aggregation of critics' votes for base scores, and we impose that a (sub-)feature attacks or supports its parent argument depending on its aggregated stance, as follows:

Definition 3. *Let $\mathcal{R}(m) = \langle \mathcal{F}, \mathcal{C}, \mathcal{V} \rangle$ be any (augmented) review aggregation for $m \in \mathcal{M}$. For any $\gamma \in \mathcal{X} = \mathcal{F} \cup \{m\}$, let $\mathcal{V}^+(\gamma) = |\{c \in \mathcal{C} | \mathcal{V}(c, \gamma) = +\}|$ and*

$\mathcal{V}^-(\gamma) = |\{c \in \mathcal{C} | \mathcal{V}(c,\gamma) = -\}|$. *Then, the* QBAF corresponding to $\mathcal{R}(m)$ *is* $\langle \mathcal{X}, \mathcal{L}^-, \mathcal{L}^+, \tau \rangle$ *such that*

$$\begin{aligned}
\mathcal{L}^- = &\{(\alpha,\beta) \in \mathcal{F}^2 | \beta = p(\alpha) \wedge \mathcal{V}^+(\beta) > \mathcal{V}^-(\beta) \wedge \mathcal{V}^+(\alpha) < \mathcal{V}^-(\alpha)\} \cup \\
&\{(\alpha,\beta) \in \mathcal{F}^2 | \beta = p(\alpha) \wedge \mathcal{V}^+(\beta) < \mathcal{V}^-(\beta) \wedge \mathcal{V}^+(\alpha) > \mathcal{V}^-(\alpha)\} \cup \\
&\{(\alpha,m) | \alpha \in \mathcal{F} \wedge m = p(\alpha) \wedge \mathcal{V}^+(\alpha) < \mathcal{V}^-(\alpha)\}; \\
\mathcal{L}^+ = &\{(\alpha,\beta) \in \mathcal{F}^2 | \beta = p(\alpha) \wedge \mathcal{V}^+(\beta) \geq \mathcal{V}^-(\beta) \wedge \mathcal{V}^+(\alpha) \geq \mathcal{V}^-(\alpha)\} \cup \\
&\{(\alpha,\beta) \in \mathcal{F}^2 | \beta = p(\alpha) \wedge \mathcal{V}^+(\beta) \leq \mathcal{V}^-(\beta) \wedge \mathcal{V}^+(\alpha) \leq \mathcal{V}^-(\alpha)\} \cup \\
&\{(\alpha,m) | \alpha \in \mathcal{F} \wedge m = p(\alpha) \wedge \mathcal{V}^+(\alpha) \geq \mathcal{V}^-(\alpha)\};
\end{aligned}$$

$$\tau(m) = 0.5 + 0.5 \cdot \frac{\mathcal{V}^+(m) - \mathcal{V}^-(m)}{|\mathcal{C}|} \text{ and } \forall f \in \mathcal{F},\ \tau(f) = \frac{|\mathcal{V}^+(f) - \mathcal{V}^-(f)|}{|\mathcal{C}|}.$$

An attack is defined as either from a feature with dominant negative votes (with respect to positive votes) towards the movie itself or from a sub-feature with dominant negative (positive) votes towards a feature with dominant positive (negative, respectively) votes. The latter type of attack can be exemplified by a sub-feature of f_A with positive stance attacking the negative (due to other votes/arguments) feature f_A, which attacks m. Conversely, a support is defined as either from a feature with dominant positive votes towards the movie itself or from a sub-feature with dominant positive (negative) votes towards a feature with dominant positive (negative, respectively) votes. The latter type of support can be exemplified by a sub-feature of f_A with negative stance supporting the negative feature f_A, which attacks m. It should be noted that (sub-)features with equal positive and negative votes are treated as supporters, though we could have assigned no relation.

In our example, we construct the BAF as follows:

- positive argument f'_{A1} supports positive argument f_A;
- neutral argument f'_{A2} neither attacks nor supports positive argument f_A;
- positive argument f_A supports positive argument m;
- negative argument f_D attacks positive argument m;
- neutral argument f_W neither attacks nor supports positive argument m;
- positive argument f'_{T1} supports positive argument f_T;
- positive argument f_T supports positive argument m.

We adapt the base score, $\tau(m) \in [0,1]$, from [24]. Intuitively, $\tau(m) = 1$ represents all critics having a positive stance on the movie while $\tau(m) = 0$ requires universally negative stance. The base score of a (sub-)feature f is again in $[0,1]$ where, differently to movies since a feature already represents positive/negative sentiment towards the argument it supports/attacks, $\tau(f) = 0$ represents no dominant negative/positive stance from the critics on f while $\tau(f) = 1$ represents universally negative/positive stance on f.

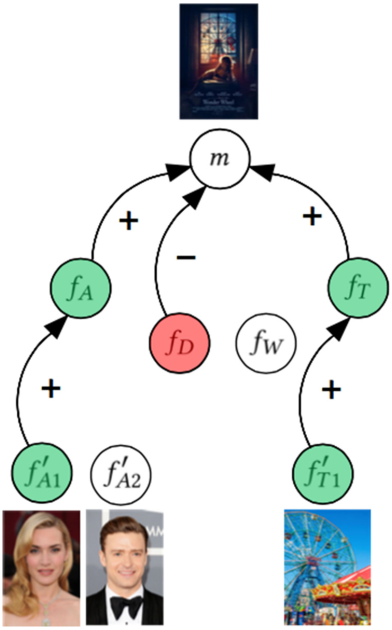


Fig. 8. The BAF obtained for the movie *Wonder Wheel*.

3.3 Explanations

The system relies on a dialogue protocol defined such that a conversation between a user and the system evolves from questions put forward by the user **U** to which the system **S** responds with explanations based on the underlying graph.

Consider the QBAF for *The Post* in Fig. 9. In Fig. 10, we can see that f_W was actually considered to be poor since it attacks m. However, the acting from Tom Hanks f'_{A1} and, particularly, Meryl Streep f'_{A2} contributed to the high strength The argumentation dialogue may then be:

U: *Why was The Post highly rated?*

S: *This movie was highly rated because the acting was great, although the writing was poor.*

U: *Why was the acting considered to be good?*

S: *The acting was considered to be good thanks to Tom Hanks and particularly Meryl Streep.*

U: *What did critics say about Meryl Streep being great?*

S: *"...Streep's hesitations, rue, and ultimate valor are soul-deep..."*

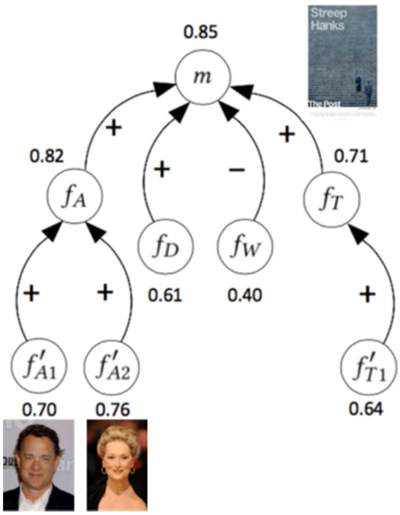


Fig. 9. QBAF for the movie *The Post*.

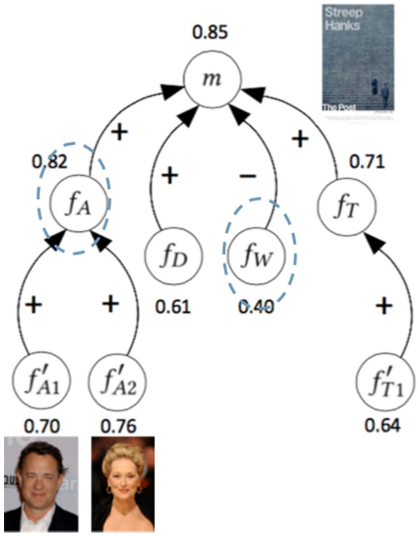


Fig. 10. The strongest supporter and strongest attacker for m.

3.4 Exercise

Given the reviews in Table 1, which represent an excerpt from the top critics reviews of the movie *Inception*:

1. identify a theme from the reviews;
2. extract votes and determine whether they have a positive or negative polarity;
3. construct the BAF from the feature-based characterisation.

Table 1. Reviews for the movie *Inception*.

$\mathbf{R_1}$ A spectacular fantasy thriller based on Nolan's own original screenplay, it is the smartest CGI head-trip since The Matrix
$\mathbf{R_2}$ A heist film of thrilling, almost delirious complexity
$\mathbf{R_3}$ Inception is a boldly constructed wonder with plenty of – as one describes it –"paradoxical architecture"
$\mathbf{R_4}$ Inception is that rare film that can be enjoyed on superficial and progressively deeper levels, a feat that uncannily mimics the mind-bending journey its protagonist takes
$\mathbf{R_5}$ Mr. DiCaprio exercises impressive control in portraying a man on the verge of losing his grip, but Mr. Nolan has not, in the end, given Cobb a rich enough inner life to sustain the performance
$\mathbf{R_6}$ In this smart sci-fi puzzle box, director Christopher Nolan transports the audience to a dreamscape that begins with the familiar and then takes a radical, imaginative leap
$\mathbf{R_7}$ In this wildly ingenious chess game, grandmaster Nolan plants ideas in our heads that disturb and dazzle. The result is a knockout. But be warned: it dreams big
$\mathbf{R_8}$ At first, Inception left me cold, feeling as if I'd just eavesdropped on somebody's bad acid trip. Now I find I can't get the film out of my mind, which is really the whole point of it, isn't it?
$\mathbf{R_9}$ A devilishly complicated, fiendishly enjoyable sci-fi voyage across a dreamscape that is thoroughly compelling

3.5 Solution

The theme extracted: *dream* (Table 2).

From $\mathbf{R_5}$ the system identifies the following potential votes, with polarity:

- *Mr. DiCaprio exercises impressive control in portraying a man on the verge of losing his grip* (0.61)
- *Mr. Nolan has not, in the end, given Cobb a rich enough inner life to sustain the performance* (-0.5)

From $\mathbf{R_7}$ the system identifies the following potential votes, with polarity:

- *In this wildly ingenious chess game, grandmaster Nolan plants ideas in our heads that disturb and dazzle* (0.8)

- *The result is a knockout* (0.5)
- *be warned: it dreams big* (0.4)

From $\mathbf{R_8}$ the system identifies the following potential votes, with polarity:

- *At first, Inception left me cold, feeling as if I'd just eavesdropped on somebody's bad acid trip* (−0.7)
- *now I find I can't get the film out of my mind, which is really the whole point of it, isn't it?* (0.8)

Table 2. Votes (and their polarity) extracted from the reviews in Table 1.

Feature	Vote id	Vote
movie	a_2	A heist film of thrilling, almost delirious complexity (+)
	a_3	Inception is a boldly constructed wonder with plenty of – as one describes it – "paradoxical architecture" (+)
	a_4	Inception is that rare film that can be enjoyed on superficial and progressively deeper levels, a feat that uncannily mimics the mind-bending journey its protagonist takes (+)
	a_8	Now I find I can't get the film out of my mind, which is really the whole point of it, isn't it (+)
	a_1 a_5 a_6 a_7	Augmented (+) Augmented (+) Augmented (+) Augmented (+)
Director Chris Nolan	a_1	A spectacular fantasy thriller based on Nolan's own original screenplay, it is the smartest cgi head-trip since the Matrix (+)
	a_6	In this smart sci-fi puzzle box, director Christopher Nolan transports the audience to a dreamscape that begins with the familiar and then takes a radical, imaginative leap (+)
	a_7	In this wildly ingenious chess game, grandmaster Nolan plants ideas in our heads that disturb and dazzle (+)
writer Chris Nolan	a_1	A spectacular fantasy thriller based on Nolan's own original screenplay, it is the smartest cgi head-trip since the Matrix (+)
acting	a_5	Augmented (+)
Leonardo DiCaprio	a_5	Mr. DiCaprio exercises impressive control in portraying a man on the verge of losing his grip (+)
themes	a_6	Augmented (+)
	a_9	Augmented (+)
dream	a_6	In this smart sci-fi puzzle box, director Christopher Nolan transports the audience to a dreamscape that begins with the familiar and then takes a radical, imaginative leap (+)
	a_9	A devilishly complicated, fiendishly enjoyable sci-fi voyage across a dreamscape that is thoroughly compelling (+)

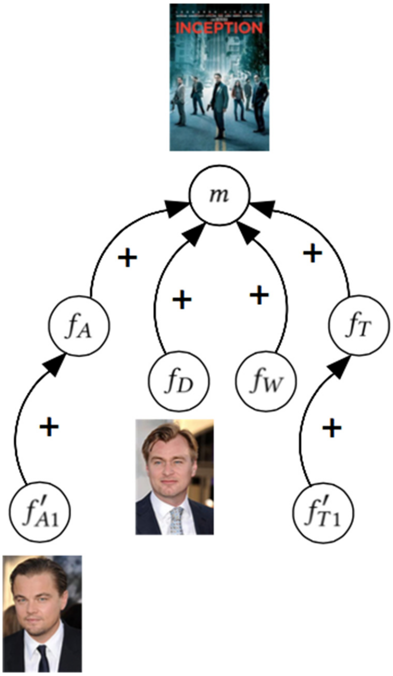


Fig. 11. BAF for movie *Inception*.

4 Extracting Argumentative Explanations from Existing AI Systems

In this section we will demonstrate how argumentative explanations may be extracted from existing AI systems. This is an extremely powerful capability as an argumentative abstraction of a system's output can provide the underlying framework for providing explanations to users in a human-like manner, given that argumentation is amenable to human consumption. This can help to alleviate some of the *black-box* issues with common AI methods since an explainable representation of the system, which is still faithful to its internal mechanisms, may be constructed.

We first introduce the *aspect-item* recommender system (RS) [23] (overviewed in Fig. 12). Here, recommendations are calculated by a hybrid method for calculating predicted ratings from ratings given by the user and by similar users. These predicted ratings (with accuracy which is competitive with the state-of-the-art) are calculated by propagating given ratings from users through an *aspect-item framework* (A-I). This underlying graphical structure comprises *item-aspects* (items and aspects) as nodes and ownership relationships from items to aspects as the edges, e.g. if an item i holds an aspect a there will be an edge (a,i) in the graph. The A-I thus houses the information used in making recommendations, thus it is from this that we define methods for extracting *tripolar argumen-*

tation frameworks (TFs) representing the reasoning for any recommendation. TFs extend classical abstract [16] and bipolar [9] argumentation frameworks by including a 'neutralising' relation (labelled 0) in addition to the standard 'attack' (labelled −) and 'support' (labelled +) relations. We will show how *argumentative explanations* (of various kinds and formats) may then be systematically generated to support interactive recommendations for users, including the opportunity for giving feedback on recommended items and their aspects. Supported formats include, amongst others, conversational and visual explanations. These explanations form the basis for interactions with users to explain recommendations and receive feedback that can be accommodated into the RS to improve its behaviour. Thus, not only are our explanations varied and diverse, but they also account (in a limited sense) for adaptable recommendations over time.

4.1 Preprocessing

Consider an RS where items (e.g. movies) are associated with aspects (e.g. comedy), which in turn have *types* (e.g. genre), and *users* may have provided *ratings* on some of the items and/or aspects. These associations may be seen to form an aspect-item framework underpinning the RS.

Definition 4. *An* aspect-item framework *(A-I for short) is a tuple* $\langle \mathcal{I},\mathcal{A},\mathcal{T},\mathcal{L},\mathcal{U},\mathcal{R} \rangle$ *such that:*

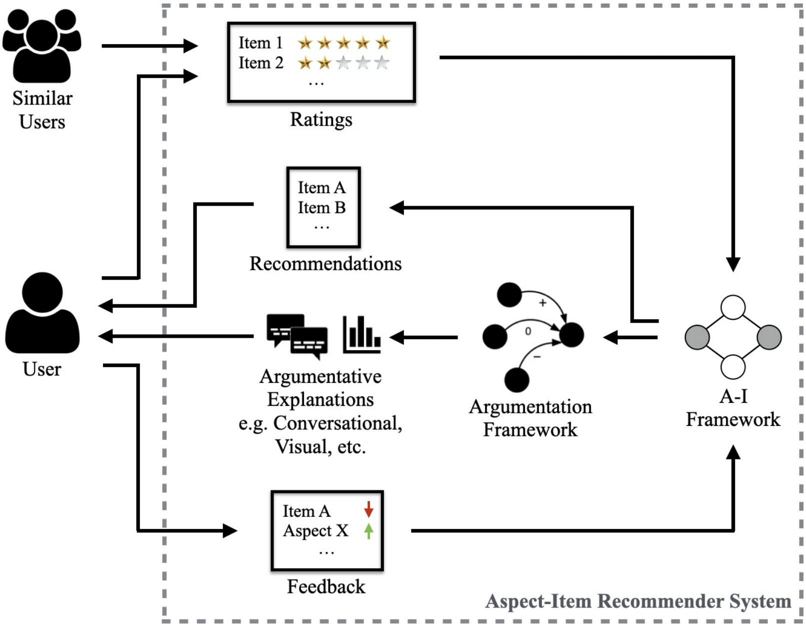


Fig. 12. Overview of the Aspect-Item RS.

- *$\mathcal{I}$ is a finite, non-empty set of* items*;*
- *$\mathcal{A}$ is a finite, non-empty set of* aspects *and $\mathcal{T}$ is a finite, non-empty set of* types *such that for each aspect $a \in \mathcal{A}$ there is a (unique) type $t \in \mathcal{T}$ with t the type of a; for any $t \in \mathcal{T}$, we use $\mathcal{A}_t$ to denote $\{a \in \mathcal{A}|$ the type of a is $t\}$;*
- *the sets $\mathcal{I}$ and $\mathcal{A}$ are disjoint; we use $\mathcal{X}$ to denote $\mathcal{I} \cup \mathcal{A}$, and refer to it as the set of* item-aspects*;*
- *$\mathcal{L} \subseteq (\mathcal{I} \times \mathcal{A})$ is a symmetric binary relation ;*
- *$\mathcal{U}$ is a finite, non-empty set of* users*;*
- *$\mathcal{R} : \mathcal{U} \times \mathcal{X} \rightarrow [-1, 1]$ is a partial function of* ratings*.*

Note that each aspect has a unique type, but of course different aspects may have the same type. Thus, $\mathcal{T}$ implicitly partitions $\mathcal{A}$, by grouping together all aspects with the same type. Note also that we assume that ratings, when defined, are real numbers in the [−1,1] interval. Positive (negative) ratings indicate that the user likes (dislikes, respectively) an item-aspect, with the magnitude indicating the strength of this sentiment. Other types of ratings can be translated into this format, for example a rating $x \in \{1, 2, 3, 4, 5\}$ can be translated into a rating $y \in [-1, 1]$ using $y = ((x - 1)/2) - 1$.

The $\mathcal{I}$, $\mathcal{A}$, $\mathcal{T}$ and $\mathcal{L}$ components of an A-I may be visualised as a graph (thus the term 'graphical chassis' for an A-I), as illustrated in Fig. 13. Here we use the movie domain as an example: items may be movies which are linked to the aspects of type "Genre" if they are of that genre. So the movie "Catch Me If You Can" in $\mathcal{I}$ may be linked to the aspect "Biography" in $\mathcal{A}$, which is of type "Genre" in $\mathcal{T}$, as shown in the figure.

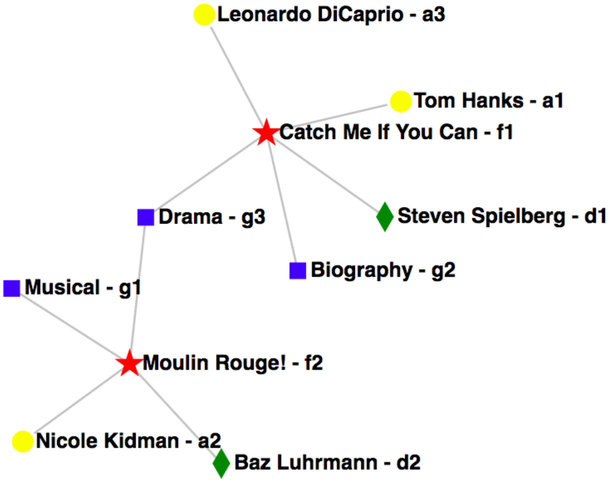


Fig. 13. Example components of an A-I visualised as a graph, with items given by red stars and types: **genres** (whose aspects are blue squares), **actors** (whose aspects are yellow circles) and **directors** (whose aspects are green diamonds). (Color figure online)

Given an A-I, predicted ratings for the items (and aspects, if required) may be calculated. In [23], we propagate the ratings through the A-I based on weightings and a *user profile* containing constants representing how much a wants a particular component to be taken into account. These aspects may be learned from data and then adjusted using feedback mechanisms, as we will show later.

The exact method used for calculating predicted ratings is not pertinent here and so we refer the reader to [23] if they wish to see an example.

4.2 Extracting TFs

We will now show how argumentation frameworks, namely TFs, may be extracted from A-Is in order to harbour the relevant information used in the calculation of a recommendation. Our aim here is to represent entities of the AI system as arguments (e.g. in the RS case, that the user likes an item-aspect) and the way in which they affect each other's evaluations (e.g. predicted ratings in the RS case) as the relations between the arguments.

For example, in the case of the RS of [23], a method for extracting a TF can be summarised as follows:

Definition 5. *Given an A-I and a user u, the corresponding TF $\langle \mathcal{X}, \mathcal{L}^-, \mathcal{L}^+, \mathcal{L}^0 \rangle$ is such that:*

- $\mathcal{X} = \mathcal{I} \cup \mathcal{A}$ *is the set of arguments that u likes each item-aspect;*

and for any $(x, y) \in \mathcal{L}$ such that $\mathcal{R}(u, y)$ is not defined (i.e. y's rating is predicted, not given) :

- $(x, y) \in \mathcal{L}^-$ *iff x had a negative effect on y's predicted rating;*
- $(x, y) \in \mathcal{L}^+$ *iff x had a positive effect on y's predicted rating;*
- $(x, y) \in \mathcal{L}^0$ *iff x had a neutralising effect on y's predicted rating.*

By defining the extraction of the TF as such, the relationships between item-aspects are categorised as argumentative relations based on the way they are used in the predicted rating calculations. In this way, the extraction function is designed to satisfy properties characterising the behaviour required by the user to support informative explanations, e.g. if an argument attacks another, its effect on the latter's predicted rating can be guaranteed to be negative.

Once again, a more explicit version of the extraction function, and the behaviour it exhibits categorised by properties, is shown in [23]. For illustration, a TF obtained from the A-I shown in Fig. 13 for user u, where $f_1 =$ *Catch Me If You Can* and $f_2 =$ *Moulin Rouge*, may be:

$$TF = \langle \{f_1, f_2, a_1, a_2, a_3, d_1, g_1, g_2, g_3\},$$

$$\{(f_2, a_2), (f_2, d_2), (f_2, g_1), (f_2, g_3), (g_3, f_1), (g_2, f_1)\},$$

$$\{(a_1, f_1), (d_1, f_1), (f_1, a_1), (f_1, g_3)\},$$

$$\{(a_3, f_1)\}\rangle,$$

as is visualised in Fig. 14. This could correspond to a situation where there are no arguments affecting f_2 since it is rated by u. Given that this rating is negative and all aspects linked to f_2 are not rated by u, f_2 attacks all such aspects. Conversely, f_1 is not rated by u but may have a positive rating from other users and thus f_1 supports all linked aspects without a rating, i.e. a_1 and g_3. The fact that f_1 is not rated by u means that all aspects linked to f_1 affect it (in an attacking, supporting or neutralising manner).

If all of the ways that item-aspects may affect one another's predicted ratings are shown by the argumentative relations, i.e. if the extraction method is *complete*, in order to explain a predicted rating for a given item-aspect, we may prune the extracted argumentation framework to be the sub-graph of the TF consisting of the item-aspects with a path to the explained item-aspect only. For example, Fig. 15 shows the sub-graph of the TF in Fig. 14, which may be seen as a qualitative explanation for the recommendation f_1 to user u, indicating all of the item-aspects which affected the recommendation.

4.3 Explanations

We will now demonstrate how argumentation frameworks extracted from this RS may be used to generate argumentative explanations to be delivered to users. TFs, like other forms of argumentation framework in different settings, form the basis for a variety of *argumentative explanations* for recommendations dictated by predicted ratings. These explanations use, as their main 'skeleton', sub-graphs

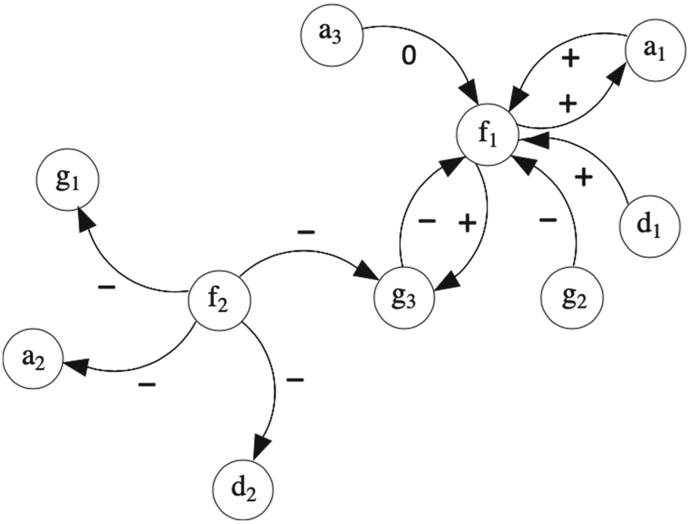


Fig. 14. A graphical representation of a possible TF extracted from the A-I in Fig. 13. Here, '+' indicates 'support' ($\mathcal{L}^+$), '-' indicates 'attack' ($\mathcal{L}^-$) and '0' indicates 'neutralises' ($\mathcal{L}^0$).

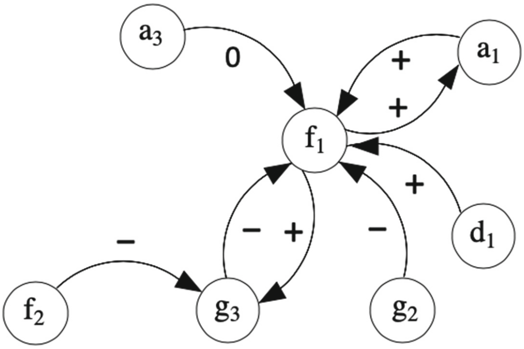


Fig. 15. Pruned version of the extracted TF in Fig. 14 for the recommendation of f_1 to u, with arguments without a path to f_1 removed.

of TFs (e.g. as in Fig. 15) providing content which can then be presented incrementally to users in different formats (e.g. as in Fig. 16) to support different styles of interaction. TFs may also point to controlled forms of feedback from users during interactions, driven by the properties they satisfy (e.g. see [23]). Thus, the use of argumentation frameworks in this manner affords great adaptability in explanations, firstly in the explanations' customisability with regards to content and format and secondly in the way it allows modifications to be made to the system via feedback mechanisms in user interactions.

4.4 Explanation Customisation

We first consider the explanation content, i.e. the information for the rationale behind a recommendation which is delivered to the user: the requirements for identifying this content obviously vary depending on user and context. We posit that the subgraph of TFs in which all nodes lead to the explained argument provides an excellent source for this information, since it represents every item-aspect which may have had an effect on the recommendation. This means that explanations that faithfully represent *how* a recommendation was determined may be drawn from this subgraph (as in Fig. 15).

The content of an explanation may be selected from this subgraph depending on the user's requirements. For example, in a basic case where the user requests information on why an item was recommended, one straightforward way to provide an explanation is for the RS to determine the positive factors which led to this result, which, in our case, would be the supporters in the sub-graph of the TF. In the case of f_1 in the example in Fig. 15, this would correspond to the content column in the first row of Table 3, which in turn could be used to obtain a linguistic explanation as in the rightmost column in Table 3, using the conjunction *because* for the argumentative relation of support and utilising the full *width* of the supporters of f_1 in the TF. If a more balanced explanation for an item being recommended is required, the style of the explanation in the

Table 3. Example variations in explanation content for f_1, with argumentative artefacts in the linguistic explanations highlighted in bold.

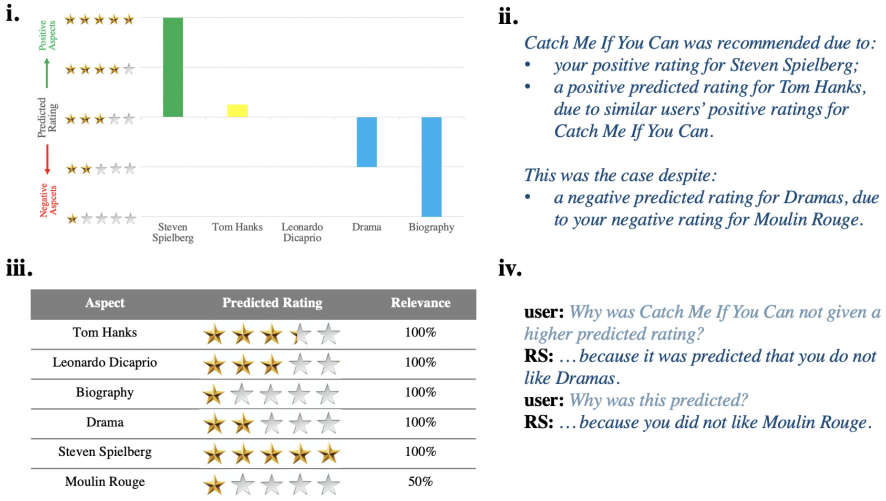
Requirements	Content	Linguistic explanation
All supporters of f_1	a_1, d_1	*Catch Me If You Can was recommended* ***because*** *you like Tom Tom Hanks and Steven Spielberg*
Strongest attacker and strongest supporter of f_1	a_1, g_2	*Catch Me If You Can was recommended* ***because*** *you like Tom Hanks,* ***despite the fact that*** *you dislike Biographies*
An attacker of f_1 and its own attacker	g_3, f_2	*Catch Me If You Can was not recommended* ***because*** *it inferred that you don't like Dramas,* ***since*** *you disliked Moulin Rouge*

rightmost column in the second row of Table 3 may be more appropriate, where the strongest attacker and strongest supporter in (the sub-graph of) the TF are shown, again using appropriate conjunctions for the argumentative relations. However, this still uses only width in the TF and ignores reasons for and against used arguments. In our running example, consider the case where f_1 was not recommended; the third row of Table 3 shows how depth may be used to justify the RS's inference on the user's sentiment on *Dramas*. Here, the language represents the resolutely negative effects along this chain of reasoning.

We have provided a number of examples for selecting the content of explanations from TFs, but note that other methods could be useful, e.g. including neutralisers when the RS is explaining its uncertainty about an inference. For example, [3] use templates to generate inferences of a user's sentiment on aspects in pairwise comparisons, e.g. *Catch Me If You Can was recommended* **because** *you like films by Steven Spielberg,* **especially** *those starring Tom Hanks*. Our argumentation frameworks could support such explanations by comparing the aspects of a movie with their linked items, e.g. in our running example (to use a crude method) if all the items which are linked to both d_1 and a_1 are rated more highly than those linked to d_1 but not a_1, we may construct the same argumentative explanation.

Up to now we have only considered explanations of a linguistic format but other formats are possible, and the choice of the format is an important factor in how receptive a user is towards explanations [18]. The optimal format for an explanation varies significantly based on a range of factors including the application towards which the explanation is targeted and the goals of the explainee [21]. For example, a researcher testing an RS may prefer a graphical format which is true to the TF itself, whereas a user may prefer a linguistic approach which gives the information in a natural, human-like manner. As we have shown in the previous section, argumentation frameworks themselves have been shown to be an effective way of supporting anthropomorphised *conversational* explanations. Other forms of explanations which have been shown to be beneficial in RSs include tabular explanations, e.g. as in [26], where (paraphrased in our setting) the attacking and supporting item-aspects in an explanation may be represented in a table with other attributes shown, e.g. the item-aspect's strength and distance from the recommendation. Visual explanations in the form of charts have also been shown to perform well in studies on user preferences [20].

Figure 16 shows four alternative formats (in addition to the graphical format afforded by sub-graphs of TFs) of user explanation for the example from Fig. 15. Specifically, Fig. 16i shows a visual explanation in the form of charts exploiting the width in the TF, i.e. attacking and supporting aspects coloured by type and organised by their corresponding predicted ratings, thus giving the user a clear indication of each aspect's contribution to the predicted rating of the recommended item. Figure 16ii, meanwhile, targets both depth and width in a linguistic format, which may be textual or spoken, e.g. by an AI assistant, depending on the requirements and preferences of the user. These explanations may be generated by templates or more complicated natural language generation processes, both employing the TF as the underlying knowledge base. Similar information is utilised in Fig. 16iii, which shows a tabular explanation similar to those of [26], where predicted ratings (translated to a 1–5 star scale) are shown alongside a *relevance* parameter, calculated here by inverting the distance from the recommended item. Finally, Fig. 16iv shows a conversational explanation, where the user has requested a counterfactual explanation as to why the item was not rated more highly. As the conversation progresses, the RS may step through the TF to formulate reasoning for its interactions, to which the user may respond with (possibly predetermined, as in [11]) responses. As with the linguistic explanations, conversational explanations may be textual or spoken.



Catch Me If You Can was recommended due to:
- *your positive rating for Steven Spielberg;*
- *a positive predicted rating for Tom Hanks, due to similar users' positive ratings for Catch Me If You Can.*

This was the case despite:
- *a negative predicted rating for Dramas, due to your negative rating for Moulin Rouge.*

Aspect	Predicted Rating	Relevance
Tom Hanks	★★★⯪☆	100%
Leonardo Dicaprio	★★★☆☆	100%
Biography	★☆☆☆☆	100%
Drama	★★☆☆☆	100%
Steven Spielberg	★★★★★	100%
Moulin Rouge	★☆☆☆☆	50%

user: *Why was Catch Me If You Can not given a higher predicted rating?*
RS: *. . . because it was predicted that you do not like Dramas.*
user: *Why was this predicted?*
RS: *. . . because you did not like Moulin Rouge.*

Fig. 16. Possible visual (i), linguistic (ii), tabular (iii) and conversational (iv) explanations for f_1's predicted rating in our running example.

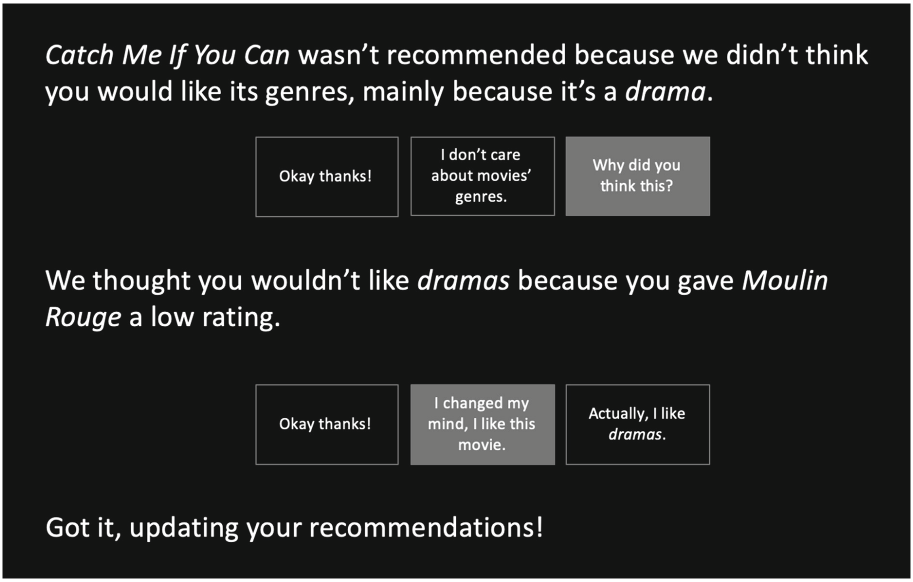
4.5 Feedback

We now consider how argumentative explanations allow for explanation-driven feedback, regarding the way in which a user may interact with the explanation to provide the system with more information. This is an important factor for RSs particularly, as recommendations are highly unlikely to be perfect the first time and, even if they are, user preferences are dynamic and so in the ideal case an RS will adapt to their changes over time [10]. Our consideration here is whether and how the RS is able to elicit more information from the user via feedback mechanisms in these interactions.

Our explanations can leverage the argumentative reading of recommendations afforded by TFs to support feedback. For example, let us focus on explanations for a positive or negative predicted rating consisting of strong supporters or strong attackers, respectively. In both cases, if the user disagrees with the predicted rating of the recommended item being so high or so low, respectively, weakening the supporters or attackers, respectively, may be guaranteed to adjust the predicted rating as desired, depending on the definition of the extracted TF. Likewise, if a user agrees with the contribution of an attacker or supporter, strengthening it may increase the effect it has. In the visual and tabular explanations in Fig. 16, it is easy to see how this intuitive behaviour allows simple indications of potential adjustments to the predicted ratings to be integrated into the explanation format such that their effect on the recommended item's predicted rating is clearly shown to the user. For example, a modifiable bar in the chart or selectable stars in the table for *Steven Spielberg* could be shown along with an indication that any reduction in the predicted rating for *Steven Spielberg* (thus the weakening of a supporter) would in turn reduce the predicted rating of Catch Me If You Can.

Other modifications supported by argumentative explanations depending on the system being explained, e.g. for the RS in [23], adjusting the user profile or selecting a different set of similar users, could also be enacted by the argumentative explanations, e.g. if a user states that they care less/more about a particular type or that they do not consider the similar users' tastes to align with their own, respectively. In the linguistic and conversational explanations, template-based interactions could be structured to include selectable user responses initiating desired modifications. For example, if a user initiates a conversational explanation with an indicated discrepancy, e.g. *I liked Catch Me If You Can, why didn't you recommend it to me?*, then the interaction with the user may be structured to include some of the possible modifications we have mentioned, e.g. as shown in Fig. 17. In the first interaction here, the user is told that the genres, particularly *Drama*, were the main reasons (possibly obtained by determining the type with the strongest attackers) for this movie not being recommended. The user may then state they are satisfied with the explanation, reduce how much genre's are taken into account (which may be guaranteed to increase *Catch Me If You Can*'s predicted rating due to the genres' negative effect on it) or ask for more reasons. In the illustration in the figure, the user does the latter, and in the second interaction the attacker *Moulin Rouge* is highlighted as the negative reasoning. The user may then state that they are satisfied with the explanation or give a higher rating to *Moulin Rouge* or *drama*, both of which may be guaranteed to increase *Catch Me If You Can*'s predicted rating.

Less constrained approaches may also be taken within an iterative feedback process: if some of this (unconstrained) feedback leads to temporary unintended effects on other item-aspects' predicted ratings, further interactions will provide an opportunity for recalibration to adhere to users' preferences.



Catch Me If You Can wasn't recommended because we didn't think you would like its genres, mainly because it's a *drama*.

Okay thanks!
I don't care about movies' genres.
Why did you think this?

We thought you wouldn't like *dramas* because you gave *Moulin Rouge* a low rating.

Okay thanks!
I changed my mind, I like this movie.
Actually, I like *dramas*.

Got it, updating your recommendations!

Fig. 17. An example conversational interaction driven by the argumentative explanations.

4.6 Exercise

1. The example movie RS of Fig. 12 has predicted ratings such that an item-aspect will affect another item-aspect's predicted rating negatively/positively/ neutrally if the former's predicted rating is negative/positive/zero. The ratings are as follows:

	Item-aspect	Actual Rating	Predicted Rating
Items	f_1	0.6	0.6
	f_2	–	0.2
Aspects	a_1	–	0.6
	a_2	0	0
	a_3	1	1
	d_1	-1	-1
	d_2	0.7	0.7
	g_1	-0.1	-0.1
	g_2	–	0.6
	g_3	–	0.6

(a) Sketch a graph of the RS, showing the item-aspects with edges indicating which items hold which aspects.
(b) Extract the attacks, supports and neutralisers for this RS's corresponding TF, assuming there are no ratings from similar users, and add them to the diagram.
(c) Which arguments are contained in the argumentation explanation for f_2?

Solution: This is presented, pictorially, in Fig. 18.

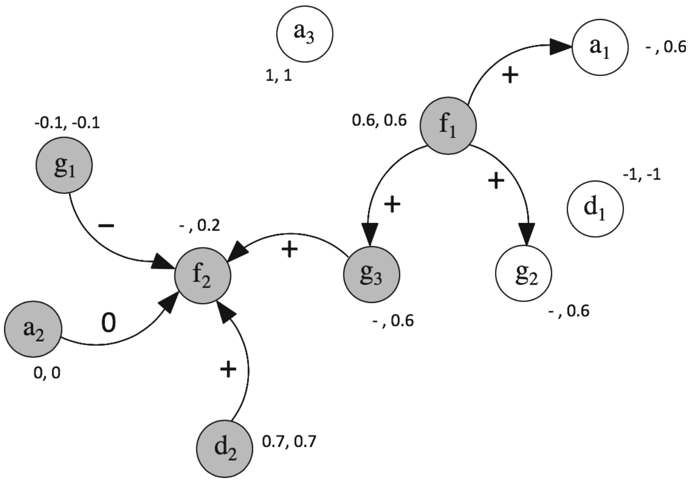


Fig. 18. Labels indicate given and predicted ratings. Arguments in the argumentation explanation for f_2 are highlighted in grey.

5 Conclusions

In this tutorial we have shown how machine arguing can be used in two distinct ways to provide explainability in AI systems. Firstly, we have shown that explainable AI systems can be built from scratch on argumentative foundations by defining a review aggregation system which allows the extraction of quantitative bipolar argumentation frameworks, which themselves support conversational interactions with users. This system is shown to produce results which are comparable with the current techniques and the capability for explainability does not induce privacy or scalability concerns. Secondly, we have shown that argumentative abstractions may be extracted from existing AI systems. To illustrate this we have focused on the domain of recommender systems, demonstrating how those which can be represented as an aspect-item framework allow the extraction of argumentative abstractions of the reasoning process in calculating a recommendation. We also show that once these argumentation frameworks have been derived, various forms of interactive explanation can be generated for users, demonstrating the flexibility of this methodology.

Various other works are based on the same "machine arguing vision" that we have demonstrated in this tutorials. In particular, within our Computational Logic and Argumentation group at Imperial College London we have contributed the following other machine-arguing-based XAI systems:

- in [15] we extract explanations from (and for) the outputs of optimsers for scheduling, and in [27] we apply them to build a system for nurse rostering, where interactive explanations are crucial; here we use abstract argumentation as the underlying form of computational argumentation;
- in [12] we extract explanations for predictions from data, building abstract argumentation frameworks from partial orders over the data; we apply the methodology to tabular data, text classification, and labelled images;
- in [1], we extract graphical explanations, which may be read as argumentative abstractions with two forms of supporting relation, from various forms of Bayesian Network Classifiers to produce counterfactual explanations.

Computational argumentation is uniquely well-placed to support XAI and various avenues for future work exist. These include, but are not limited to, a "theory" of explanations, the extraction of argumentative explanations for a broader variety of AI systems, including opaque ones (such as deep learning), and engineering aspects of machine arguing for other AI systems and applications. We conclude with some considerations relating to these avenues.

From a theoretical view point, it would be interesting to study computational aspects of the methodologies we have described, and in particular the cost of mining argumentation frameworks and explanations from them.

Different types of explanations can be defined to address the expertise of users the system targets. These range from regular users to expert users who require an understanding of the underlying model. From a theoretical point of view, the explanations need to exhibit different properties depending on the type of user the system aims to interact with, which, in turn, will also determine the level of detail included in the explanation. Different argumentation frameworks may be suited to different applications. The type of framework most relevant to the application at hand can be identified either from the properties it fulfills that match with the properties required by the application or in an empirical manner.

Other applications where machine arguing is relevant include: deception detection, fact checking and fake news detection by extracting argumentation frameworks and providing an explanation as to why a piece of text is true or deceptive/false/fake, data summarisation by means of graphical outputs extracted from argumentation frameworks, and explaining outputs of black-box models by means of argumentation frameworks which can help debug and correct the black-box models which learn from labelled data.

The works overviewed in this tutorial and the additional works mentioned here show promising uses of machine-arguing-based XAI. We hope this tutorial will enthuse others to join forces towards the several paths of future work that machine-arguing-based XAI faces, including at the forefront theoretical, empirical and experimental evaluation, as partially outlined here.

References

1. Albini, E., Rago, A., Baroni, P., Toni, F.: Relation-based counterfactual explanations for Bayesian network classifiers. In: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI (2020, To Appear)
2. Atkinson, K., et al.: Towards artificial argumentation. AI Mag. **38**(3), 25–36 (2017)
3. Balog, K., Radlinski, F., Arakelyan, S.: Transparent, scrutable and explainable user models for personalized recommendation. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR, pp. 265–274 (2019)
4. Baroni, P., Comini, G., Rago, A., Toni, F.: Abstract games of argumentation strategy and game-theoretical argument strength. In: An, B., Bazzan, A., Leite, J., Villata, S., van der Torre, L. (eds.) PRIMA 2017. LNCS (LNAI), vol. 10621, pp. 403–419. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-69131-2_24
5. Baroni, P., Gabbay, D., Giacomin, M., van der Torre, L. (eds.): Handbook of Formal Argumentation. College Publications (2018)
6. Baroni, P., Rago, A., Toni, F.: How many properties do we need for gradual argumentation? In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, The 30th innovative Applications of Artificial Intelligence (IAAI), and The 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI), pp. 1736–1743 (2018)
7. Baroni, P., Rago, A., Toni, F.: From fine-grained properties to broad principles for gradual argumentation: a principled spectrum. Int. J. Approximate Reasoning **105**, 252–286 (2019)
8. Briguez, C.E., Budán, M.C., Deagustini, C.A.D., Maguitman, A.G., Capobianco, M., Simari, G.R.: Argument-based mixed recommenders and their application to movie suggestion. Expert Syst. Appl. **41**(14), 6467–6482 (2014)
9. Cayrol, C., Lagasquie-Schiex, M.C.: On the acceptability of arguments in bipolar argumentation frameworks. In: ECSQARU, pp. 378–389 (2005)
10. Chen, X., Zhang, Y., Qin, Z.: Dynamic explainable recommendation based on neural attentive models. In: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI, pp. 53–60 (2019)

11. Cocarascu, O., Rago, A., Toni, F.: Extracting dialogical explanations for review aggregations with argumentative dialogical agents. In: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS, pp. 1261–1269 (2019)
12. Cocarascu, O., Stylianou, A., Cyras, K., Toni, F.: Data-empowered argumentation for dialectically explainable predictions. In: Proceedings of European Conference on Artificial Intelligence, ECAI 2020 (2020)
13. Cohen, A., Gottifredi, S., García, A.J., Simari, G.R.: A survey of different approaches to support in argumentation systems. Knowl. Eng. Rev. **29**(5), 513–550 (2014)
14. Cohen, A., Parsons, S., Sklar, E.I., McBurney, P.: A characterization of types of support between structured arguments and their relationship with support in abstract argumentation. Int. J. Approximate Reasoning **94**, 76–104 (2018)
15. Cyras, K., Letsios, D., Misener, R., Toni, F.: Argumentation for explainable scheduling. In: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI, pp. 2752–2759. AAAI Press (2019)
16. Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artif. Intell. **77**(2), 321–357 (1995)
17. Gabbay, D.M.: Logical foundations for bipolar and tripolar argumentation networks: preliminary results. J. Logic Comput. **26**(1), 247–292 (2016)
18. Gedikli, F., Jannach, D., Ge, M.: How should I explain? A comparison of different explanation types for recommender systems. Int. J. Hum. Comput. Stud. **72**(4), 367–382 (2014)
19. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Comput. Surv. **51**(5), 93:1–93:42 (2019). https://doi.org/10.1145/3236009
20. Herlocker, J.L., Konstan, J.A., Riedl, J.: Explaining collaborative filtering recommendations. In: Proceeding on the ACM Conference on Computer Supported Cooperative Work, CSCW, pp. 241–250 (2000)
21. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artif. Intell. **267**, 1–38 (2019). https://doi.org/10.1016/j.artint.2018.07.007
22. Naveed, S., Donkers, T., Ziegler, J.: Argumentation-based explanations in recommender systems: conceptual framework and empirical results. In: Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization, UMAP, pp. 293–298 (2018)
23. Rago, A., Cocarascu, O., Toni, F.: Argumentation-based recommendations: fantastic explanations and how to find them. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI, pp. 1949–1955 (2018)
24. Rago, A., Toni, F.: Quantitative argumentation debates with votes for opinion polling. In: An, B., Bazzan, A., Leite, J., Villata, S., van der Torre, L. (eds.) PRIMA 2017. LNCS (LNAI), vol. 10621, pp. 369–385. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-69131-2_22
25. Rago, A., Toni, F., Aurisicchio, M., Baroni, P.: Discontinuity-free decision support with quantitative argumentation debates. In: KR, pp. 63–73 (2016)

26. Vig, J., Sen, S., Riedl, J.: Tagsplanations: explaining recommendations using tags. In: Proceedings of the 14th International Conference on Intelligent User Interfaces, IUI, pp. 47–56 (2009)
27. Čyras, K., Karamlou, A., Lee, M., Letsios, D., Misener, R., Toni, F.: AI-assisted schedule explainer for nurse rostering - Demonstration. In: International Conference on Autonomous Agents and Multi-Agent Systems (2020)