Data-Empowered Argumentation for Dialectically Explainable Predictions

Oana Cocarascu and **Andria Stylianou** and **Kristijonas Čyras** and **Francesca Toni** [1]

Abstract. Today's AI landscape is permeated by plentiful data and dominated by powerful data-centric methods with the potential to impact a wide range of human sectors. Yet, in some settings this potential is hindered by these data-centric AI methods being mostly opaque. Considerable efforts are currently being devoted to defining methods for explaining black-box techniques in some settings, while the use of transparent methods is being advocated in others, especially when high-stake decisions are involved, as in healthcare and the practice of law. In this paper we advocate a novel transparent paradigm of Data-Empowered Argumentation (DEAr in short) for dialectically explainable predictions. DEAr relies upon the extraction of argumentation debates from data, so that the dialectical outcomes of these debates amount to predictions (e.g. classifications) that can be explained dialectically. The argumentation debates consist of (data) arguments which may not be linguistic in general but may nonetheless be deemed to be 'arguments' in that they are dialectically related, for instance by disagreeing on data labels. We illustrate and experiment with the DEAr paradigm in three settings, making use, respectively, of categorical data, (annotated) images and text. We show empirically that DEAr is competitive with another transparent model, namely decision trees (DTs), while also providing naturally dialectical explanations.

1 INTRODUCTION

Data-centric AI is receiving a great deal of attention nowadays, thanks to its ability to produce high-performing predictive models on real-world data. However, data-centric methods and models are often opaque in nature, making their decisions hard to understand. Moreover, they may sometimes leverage on biases in the data to give fatal or unfair decisions, raising inevitably some ethical dilemma especially in critical domains such as healthcare and the practise of law. Overall, explanainability of data-centric methods has been identified as a crucial challenge in AI [19].

Considerable efforts are currently being devoted to defining surrogate (mostly linear) models for explaining black-box models, aiming to overcome their lack of transparency [19, 1, 12], with some of the most popular being model-agnostic [38, 37, 31]. Alternatively, existing transparent models can be coupled with black-box models to give hybrid (high-performing and explainable) systems. For example, Case-Based Reasoning (CBR) and Artificial Neural Networks (ANNs) have recently been combined to give a twin-system providing post-hoc explanations by-example, using the former to explain the predictions of the latter [23]. The most widespread hybrid systems though use Decision Trees (DTs) for explaining other (black-box) methods (e.g. for prediction from tabular data [8, 27, 6, 22]).

At the same time, there is an open discussion on whether making black-box methods explainable can establish a trustworthy relationship between the user and the system [39]. The use of transparent reasoning methods is being advocated in alternative to surrogates and twinning models as is explanation by design, whereby models concurrently produce predictions and naturally induced explanations such as DTs and kNNs.

In AI, as in Psychology and Philosophy, there is no agreed definition for the notion of "explanation" [41]). However, within the social science literature on explanation [34], conversation-based models of explanation have been advocated and proved to be very useful models for explanation in AI, including some where statements in explanations are argumentative [4]. In this paper we propose the Data-Empowered Argumentation (DEAr) paradigm as a transparent method for prediction from which dialectical explanations can be drawn naturally.

DEAr relies upon the fundamental tenet that argumentation debates extracted from data can be the basis for predictions from the data (as well as explanations thereof). The predictions result from analysing the argumentation debates by means of semantics, as conventional in the field of computational argumentation in AI (for an overview of this field see [5]). The explanations are fragments of the argumentation debates tailored to the predictions, including arguments for and against them, in a manner that is leaning towards accepting the prediction. In this paper, we choose as argumentation debates abstract argumentation frameworks [13]. In general, these provide abstractions of reasoning problems of various kinds in terms of directed graphs whose nodes are interpreted as arguments and whose edges represent a binary relation of "attack" between arguments. Reasoning then amounts to identifying attack-free sets of arguments that can self-defend against attacks, e.g. by forming the grounded extension [13].

In DEAr, argumentation debates in the form of abstract argumentation frameworks are mined from data, generalising the approach advocated in [9, 10] as a form of case-based reasoning in legal settings. In this paper we evaluate empirically the usefulness of DEAr, in comparison with DTs, in various settings and applications to support the use of DEAr as a solution (at least in some settings) to the explainable AI problem.

The paper is organised as follows. Section 2 gives essential background on abstract argumentation. Section 3 defines the DEAr paradigm and Section 4 gives the pipeline for deploying DEAr experimentally, paving the way towards the experiments in three data settings: categorical data (Section 5), (annotated) images (Section 6) and text (Section 7). These three sections have a similar structure,

[1] Imperial College London, United Kingdom, email: {oana.cocarascu11,andria.stylianou13,k.cyras,f.toni}@ic.ac.uk

each of them providing a description of datasets used, the choices of DEAr's parameters and the methods deployed to obtain argumentation debates in DEAr followed by an empirical evaluation and a discussion of the output results, in comparison, amongst others, with Decision Trees (DTs) as our benchmark alternative transparent model. Section 8 compares DEAr with the closest related work and Section 9 concludes and considers future directions.

2 BACKGROUND

An *abstract argumentation framework* [13] is a pair $(Args, \rightsquigarrow)$, where $Args$ is a set (of *arguments*) and $\rightsquigarrow$ is a binary relation on $Args$ (where, for $\mathsf{a}, \mathsf{b} \in Args$, if $\mathsf{a} \rightsquigarrow \mathsf{b}$, then we say that a *attacks* b and that a is an *attacker of* b). For a set of arguments $E \subseteq Args$ and an argument $\mathsf{a} \in Args$, E *defends* a if for all $\mathsf{b} \rightsquigarrow \mathsf{a}$ there exists $\mathsf{c} \in E$ such that $\mathsf{c} \rightsquigarrow \mathsf{b}$. Then, the *grounded extension* of $(Args, \rightsquigarrow)$ can be constructed as $\mathbb{G} = \bigcup_{i \geqslant 0} G_i$, where G_0 is the set of all unattacked arguments, and $\forall i \geqslant 0$, G_{i+1} is the set of arguments that G_i defends. For any $(Args, \rightsquigarrow)$, the grounded extension $\mathbb{G}$ always exists and is unique and, if $(Args, \rightsquigarrow)$ is well-founded [13], extensions under other semantics (e.g. under the stable semantics [13]) are equal to $\mathbb{G}$.

Explanations for abstract argumentation outcomes can be defined in the form of *dispute trees* [9, 10], where a *dispute tree* for $\mathsf{a} \in Args$ is a tree $\mathcal{T}$ such that:

1. every node of $\mathcal{T}$ is of the form $[L : \mathsf{x}]$, with $L \in \{\mathsf{P}, \mathsf{O}\}$, $\mathsf{x} \in Args$: the node is *labelled* by argument x and assigned the status of either *proponent* (P) or *opponent* (O);
2. the root of $\mathcal{T}$ is a P node labelled by a;
3. for every P node n, labelled by some $\mathsf{b} \in Args$, and for every $\mathsf{c} \in Args$ such that $\mathsf{c} \rightsquigarrow \mathsf{b}$, there exists a child of n, which is an O node labelled by c;
4. for every O node n, labelled by some $\mathsf{b} \in Args$, there exists at most one child of n which is a P node labelled by some $\mathsf{c} \in Args$ such that $\mathsf{c} \rightsquigarrow \mathsf{b}$;
5. there are no other nodes in $\mathcal{T}$ except those given by 1–4.

A dispute tree $\mathcal{T}$ is an *admissible dispute tree* iff (i) every O node in $\mathcal{T}$ has a child, and (ii) no argument in $\mathcal{T}$ labels both P and O nodes.

A dispute tree $\mathcal{T}$ is a *maximal dispute tree* iff for all opponent nodes $[\mathsf{O} : \mathsf{x}]$ which are leaves in $\mathcal{T}$ there is no $\mathsf{y} \in Args$ such that y attacks x.

3 DATA-EMPOWERED ARGUMENTATION

Formally, consider a *training dataset* $\mathcal{D}$ consisting of a finite (but possibly large) set of datapoints, each labelled with an *outcome* from a set $\mathbb{O}$. In this paper, for simplicity, we assume that $\mathbb{O} = \{\delta, \overline{\delta}\}$, namely datapoints can be labelled by one of two (distinct) outcomes, and the prediction task is binary classification. Specifically, the prediction task amounts to determining which amongst δ or $\overline{\delta}$ should be the legitimate outcome for an unlabelled datapoint dp_U.

DEAr relies upon the assumption that one of the outcomes is identified as the *default outcome*, which is intuitively the outcome drawn in the absence of any useful information. In the remainder of the paper we will assume without loss of generality that δ indicates the default outcome in $\mathbb{O}$. We will see that the choice of default outcome δ is context-dependent and heuristic: this choice amounts to instantiating a core parameter in the deployment of DEAr.

DEAr makes the further assumption that $\mathcal{D} \cup \{dp_U\}$ is equipped with a partial order $\succcurlyeq$ (namely a reflexive, antisymmetric, and transitive relation) , so that, for datapoints $dp_X, dp_Y \in \mathcal{D} \cup \{dp_U\}$, $dp_X \succcurlyeq dp_Y$ means intuitively that dp_X is *more informative than or as informative as* dp_Y; we will use $dp_X \succ dp_Y$ to indicate that $dp_X \succcurlyeq dp_Y$ and $dp_X \neq dp_Y$, namely that dp_X is *strictly more informative than* dp_Y.

Consider, for illustration, the specific setting where (labelled and unlabelled) datapoints are characterised by binary features from a given set $\mathbb{F}$. Then, $\succcurlyeq$ may amount to the $\subseteq$ relation between sets. With this choice of $\succcurlyeq$, $dp_X \succ dp_Y$ if the set of features of dp_X is a strict superset of the set of features of dp_Y. In general, datapoints may be formulated in other terms: in the reminder of this section we will assume that (labelled and unlabelled) datapoints are given in terms of generic *characterisations*, and that a labelled datapoint is of the form (C, o) for C the characterisation and $o \in \mathbb{O}$.

In this paper, argumentation debates amount to abstract argumentation frameworks mined from datasets and labelled datapoints as given above. These argumentation debates are deterministically obtained from the choices detailed earlier as well as an two additional choices, one for each of two additional parameters that need to be instantiated when DEAr is deployed. The first choice amounts of a synthetic datapoint associated with the default outcome (referred to as the *default argument*), whose description expresses conditions under which the default outcome can be argued for. A possible choice for this datapoint is the least element of $\succcurlyeq$, i.e. the least informative possible datapoint. The second choice amounts to a notion of *irrelevance* $\not\sim$ between unlabelled datapoints and labelled datapoints: for $dp_X \in \mathcal{D}$, $dp_U \not\sim dp_X$ stands for "dp_X is *irrelevant to* dp_U". As an illustration, if datapoints are characterised by binary features, a possible choice for $\not\sim$ is $\not\supseteq$, namely $dp_U \not\sim dp_X$ if dp_X has features that dp_U lacks. Whichever the definition of $\not\sim$, we will assume that it satisfies the property that $C \not\sim (C, o)$ never holds, for C any characterisation (thus, $\not\sim$ satisfies a form of anti-reflexivity).

Formally, given choices for all parameters, the argumentation debate mined from data for the purposes of binary classification is as follows:

Definition 3.1. Let $\mathcal{D}$ be a finite dataset, consisting of labelled datapoints dp_i, each of the form (C_i, o_i) with C_i a characterisation of the datapoint and $o_i \in \mathbb{O}$, $\mathbb{O} = \{\delta, \overline{\delta}\}$ with δ the default outcome. Let dp_U be an unlabelled datapoint. Finally, let $\succcurlyeq$ be a partial order over $\mathcal{D} \cup \{dp_U\}$ and $\not\sim$ a notion of irrelevance. Then, an *argumentation debate* mined from $\mathcal{D} \cup \{dp_U\}$ is an abstract argumentation framework $(Args, \rightsquigarrow)$ with

- $Args = \mathcal{D} \cup \{(C_\delta, \delta)\} \cup \{dp_U\}$, for C_δ a characterisation of the *default argument* (C_δ, δ);
- for $(X, o_X), (Y, o_Y) \in \mathcal{D} \cup \{(C_\delta, \delta)\}$, it holds that $(X, o_X) \rightsquigarrow (Y, o_Y)$ iff
  1. $o_X \neq o_Y$, and
  2. either $X \succ Y$ and $\nexists (Z, o_X) \in \mathcal{D} \cup \{(C_\delta, \delta)\}$ with $X \succ Z \succ Y$
  3. or $X = Y$;
- for $(Y, o_Y) \in \mathcal{D} \cup \{(C_\delta, \delta)\}$, it holds that $dp_U \rightsquigarrow (Y, o_Y)$ iff $N \not\sim Y$.

In the second bullet, condition 1 amounts to (X, o_X) and (Y, o_Y) having different outcomes, case 2 amounts to (X, o_X) being strictly more informative than (Y, o_Y) and imposes a form of informational minimality on the attacking argument, whereas case 3 deals with noise (datapoints with the same characterisation but conflicting outcomes attack one another). In the third bullet, the unlabelled datapoint/argument attacks any datapoints/arguments that are irrelevant to it. The mined argumentation debate can be seen as including a "model" of the dataset, identifying conflicts between datapoints that need to be resolved every time a prediction is to be made.

Note that if $\mathcal{D}$ is *coherent*, namely $\nexists dp_X, dp_Y \in \mathcal{D}$ such that $dp_X = (C, o_X)$ and $dp_Y = (C, o_Y)$ for $o_X \neq o_Y$, then case 3 in the definition of attack never applies.

The following properties hold of mined argumentation debates:

Theorem 3.1. *Let $(Args, \leadsto)$ be an argumentation debate mined from $\mathcal{D} \cup \{dp_U\}$, and let G be its grounded extension.*
i) $\mathbb{G}$ is non-empty and contains dp_U.
ii) If $\mathcal{D}$ is coherent then $(Args, \leadsto)$ is acyclic.
iii) If $\mathcal{D}$ is coherent then either (C_δ, δ) or some dp such that dp attacks (C_δ, δ) belongs to $\mathbb{G}$.

DEAr relies upon membership of the default argument in the grounded extension of the argumentation debate mined from a dataset/unlabelled datapoint combination to determine a prediction for the unlabelled datapoint:

Definition 3.2. Let $(Args, \leadsto)$ be an argumentation debate mined from $\mathcal{D} \cup \{dp_U\}$, and let $\mathbb{G}$ be the grounded extension of $(Args, \leadsto)$. The *DEAr prediction* for dp_U is δ if $(C_\delta, \delta) \in \mathbb{G}$, and $\overline{\delta}$ otherwise.

As standard for abstract argumentation, the argumentation debate mined from a dataset and an unlabelled datapoint can be visualised as a graph (the arguments being the nodes, and the attack relation the edges of the graph). Then DEAr predictions can be naturally explained in terms of sub-graphs of the argumentation debate including the default argument and all its descendants. As a refinement, disputes between fictional *proponent* and *opponent* players, in the form of admissible dispute trees or, if none exists, maximal dispute trees can be extracted from these sub-graphs to explain the predictions, generalising the approach of [9] to the setting of DEAr.

4 THE DEAr PIPELINE

In the remainder of the paper we will present a number of experiments with DEAr. These require specific choices of parameters for a variety of datasets and characterisations, within the pipeline depicted in Figure 1.

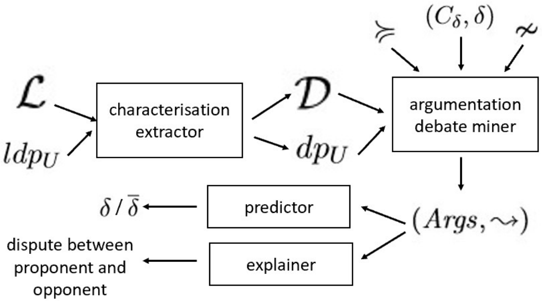
$\succsim$
(C_δ, δ)
$\nsim$
$\mathcal{L}$
ldp_U
characterisation extractor
$\mathcal{D}$
dp_U
argumentation debate miner
$\delta\,/\,\overline{\delta}$
predictor
$(Args, \leadsto)$
dispute between proponent and opponent
explainer

Figure 1: The DEAr pipeline. Each experiment with DEAr requires instantiating the parameters $((C_\delta, \delta), \succsim, \nsim)$, as well as engineering suitable characterisations by a 'characterisation extractor' that obtains $\mathcal{D}$ from a possibly much larger dataset $\mathcal{L}$. The 'argumentation debate miner implements Definition 3.1, and the 'predictor' implements Definition 3.2. The 'explainer'returns dialectical explanations in the form of disputes between a proponent and an opponent.

The characterisation extractor may be designed to obtain a coherent dataset or not. In either case, it may identify and/or select features in datapoints. These features may take continuous values, and the set of features may be very large in general. All deployments of DEAr in this paper makes use of a relatively small set $\mathbb{F}$ of binary features (these are the salient features for classification) and of a variety of methods for obtaining these features from larger sets $\mathbb{F}_\mathcal{L}$ of (possibly continuous) features, for structured and unstructured data.

In the reminder of this section we give a toy illustration of the DEAr pipeline, for specific choices of parameters, corresponding to the choices in the AA-CBR paradigm [9, 10].

Let examples in $\mathcal{L}$ be characterised by 5 attributes $a_1, \dots, a_5$ and an outcome in $\mathbb{O} = \{o_1, o_2\}$. Suppose each attribute may take one of 4 distinct, discrete values, say $v^1_{a_i}, \dots, v^4_{a_i}$ for a_i. Then $\mathbb{F}_\mathcal{L} = \{a_1 = v^1_{a_1}, \dots, a_5 = v^4_{a_5}\}$ consists of 20 features (binary attribute-value pairs). Suppose that we select 4 features, namely $\mathbb{F}$ consists of 4 elements. Concretely, say $\mathbb{F} = \{a_1 = v^1_{a_1}, a_2 = v^4_{a_2}, a_3 = v^2_{a_3}, a_4 = v^4_{a_4}\} = \{\mathsf{a}, \mathsf{b}, \mathsf{c}, \mathsf{d}\}$ If $\mathcal{L} = \{(\{\mathsf{a}, a_1 = v^2_{a_1}, a_5 = v^1_{a_5}\}, o_2), (\{\mathsf{a}, a_1 = v^3_{a_1}\}, o_2), (\{\mathsf{a}, \mathsf{b}\}, o_1)\}$ then $\mathcal{D} = \{(\{\mathsf{a}\}, o_2), (\{\mathsf{a}, \mathsf{b}\}, o_1)\}$ is the result of the 'characterisation extractor' [2]. Also, if ldp_U is $\{\mathsf{a}, \mathsf{d}\}$, then the 'characterisation extractor' would return $dp_U = ldp_U$.

Consider the following choice of parameters:

- $\succsim = \supseteq$ (and thus $\succ = \supset$);
- $(C_\delta, \delta) = (\{\}, \delta)$, for $\delta = o_1$ (and $\overline{\delta} = o_2$);
- $\nsim = \not\succsim = \not\supseteq$.

Then the 'argumentation debate miner' gives $(Args, \leadsto)$ with $\leadsto = \{((\{\mathsf{a}\}, o_2), (\{\}, o_1)), ((\{\mathsf{a}, \mathsf{b}\}, o_1), (\{\mathsf{a}\}, o_2)), (dp_U = \{\mathsf{a}, \mathsf{d}\}, (\{\mathsf{a}, \mathsf{b}\}, o_1))\}$. Since $\mathbb{G} = \{(\{\mathsf{a}\}, o_2)$,
$unlab = \mathsf{a}, \mathsf{d}\}\}$ and $(\{\}, o_1) \notin \mathbb{G}$, the 'predictor' returns o_2 (i.e. the *non-default*) as the prediction for dp_U.

A possible explanation for the prediction o_2 for dp_U is given by the following (maximal) dispute tree:

$$\begin{array}{c} [P : (\{\}, o_1)] \\ | \\ [O : (\{\mathsf{a}\}, o_2)] \\ | \\ [P : (\{\mathsf{a}, \mathsf{b}\}, o_1)] \\ | \\ [O : dp_U = \{\mathsf{a}, \mathsf{d}\}] \end{array}$$

Presented dialectically, this dispute between proponent and opponent unfolds with the following arguments:

P: 'o_1';
O: $(\{\mathsf{a}\}, o_2)$ attacks the outcome 'o_1';
P: $(\{\mathsf{a}, \mathsf{b}\}, o_1)$ attacks the previous argument;
O: b is irrelevant (as absent from dp_U).

Basically, P starts by arguing for outcome o_1 saying that, in the absence of any information, the outcome should be o_1. O then argues against this outcome by putting forward an argument with outcome o_2. P then puts forward a more informative example that gives outcome o_1 with some features that are in dp_U (i.e. a). However, the unlabelled datapoint attacks this argument, thus this argument is discarded from having an influence on the prediction. Hence, the prediction is o_2.

Here, the explanation for the prediction being o_2 relies upon the opponent using dp_U to defeat the proponent who is trying to defend

[2] Note that if $\mathcal{L}$ had also included $(\{\mathsf{a}, a_5 = v^1_{a_5}\}, o_2)$, then restricting attention to $\mathbb{F}$ would have resulted in an incoherent $\mathcal{D} = \{(\{\mathsf{a}\}, o_1), (\{\mathsf{a}\}, o_2)\}$ as output of the 'characterisation extractor'.

the default argument. The 'explainer' could produce this dispute as a dialectical explanation, in any of the formats above, but it could also use the dispute to unearth *counterfactual explanations* such as "If dp_U contained b instead of d, then the outcome would have been o_1". Indeed, consider now $dp'_U = \{\mathsf{a}, \mathsf{b}\}$ (thus dp'_U is dp_U with b instead of d) then $(\{\}, o_1) \in \mathbb{G}$ where $\mathbb{G} = \{(\{\}, o_1), (\{\mathsf{a}, \mathsf{b}\}, o_1), dp'_U = \{\mathsf{a}, \mathsf{b}\}\}$ and the outcome for dp'_U is o_1. In this case, the dialectical explanation may be drawn from the admissible dispute tree:

$$\begin{array}{c} [P : (\{\}, o_1)] \\ | \\ [O : (\{\mathsf{a}\}, o_2)] \\ | \\ [P : (\{\mathsf{a}, \mathsf{b}\}, o_1)] \end{array}$$

We leave the fleshing out of the 'explainer' including in particular the definition of counterfactual explanation in the general case as future work, and focus instead in he remainder of the paper on predictive performances of DEAr in several empirical settings.

5 DEAr FOR CATEGORICAL DATA

The first empirical evaluation uses the pipeline in Figure 1 to provide predictions (and explanations thereof) for a categorical dataset. We choose the mushroom dataset[3] from the UCI Machine Learning Repository [11]. This dataset contains 8124 examples of gilled mushrooms classified as edible or poisonous. Each example is characterised by 22 categorical attributes that can take a number of different values, leading to 126 binary features. Our starting point $\mathcal{L}$ consists of (subsets of) the 8124 examples as datapoints, each characterised by a subset of the 126 binary features ($\mathbb{F}_{\mathcal{L}}$).

5.1 Characterisation Extractor

As our characterisations we use sets of features from a reduced set of features $\mathbb{F} \subseteq \mathbb{F}_{\mathcal{L}}$. Thus, in this empirical setting, the first stage in the DEAr pipeline is concerned with dimensionality reduction. To obtain $\mathbb{F}$ we use an autoencoder, a type of Artificial Neural Network (ANN), as our characterisation extractor. ANNs have been widely applied both in classification tasks and in dimensionality reduction, e.g. as in [42]. ANN-based feature selection methods use multilayer perceptrons to determine which features are redundant [18] as well as autoencoders [20, 21, 43]. These are unsupervised learning models based on ANNs which take a set of features as input and aim, through training, to reconstruct the inputs [21, 14].

Our proposed autoencoder is shown in Figure 2. The autoencoder has one hidden layer with (for $X \subseteq \mathbb{F}_{\mathcal{L}}$):

1. an encoder function $f(X) = \sigma(XW^{(1)})$
2. a decoder function $\sigma(f(X)W^{(2)})$

where $W^{(1)}, W^{(2)}$ are the weight parameters in the encoder and decoder, respectively.

In order to select $\mathbb{F}$, we average the weights in $W^{(1)}$ for each input and select the top F factors. F can be chosen in many alternative ways, either iteratively (starting from a small number of features, until a coherent $\mathcal{D}$ is obtained) or empirically (as in the experiments in Section 5.3, where by construction $\mathcal{D}$ is guaranteed to be coherent).

Once selected $\mathbb{F}$, the inputs $\mathcal{D}$ and dp_U for DEAr are automatically obtained from $\mathcal{L}$ and ldp_U by restriction to the features in $\mathbb{F}$. Formally, $\mathcal{D} = \{(Y, o) | (X, o) \in \mathcal{L}, Y = X \cap \mathbb{F}\}$.

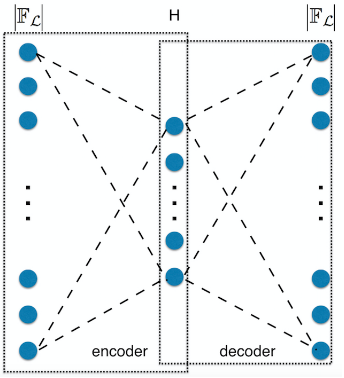

Figure 2: Autoencoder architecture: $|\mathbb{F}_{\mathcal{L}}|$ binary features are used to train the autoencoder to obtain a **code** (hidden layer **h** of size $H < |\mathbb{F}_{\mathcal{L}}|$) that best captures the input $|\mathbb{F}_{\mathcal{L}}|$ features.

5.2 Choices of Parameters

For this empirical evaluation, we choose

- $\succcurlyeq = \supseteq$ (and thus $\succ = \supset$);
- $(C_\delta, \delta) = (\{\}, \delta)$, for $\mathbb{O} = \{$edible, poisonous$\}$ and δ=edible, $\bar{\delta}$=poisonous;
- $\not\sim = \not\succcurlyeq = \not\supseteq$.

Thus, the default argument is characterised by the empty set of features. The default is chosen empirically (to obtain best performances, as given in Section 5.3). The choice means that in the absence of any information (i.e. features) about a mushroom, it can be deemed edible, as represented by (C_δ, δ). However, within the argumentation debate mined by DEAr, as soon as a mushroom with any features is encountered, it being edible has to be justified by countering all the relevant examples of poisonous mushrooms.

5.3 Empirical Evaluation

We deploy DEAr's miner and predictor as given in Section 3 and verify that it predicts well against Decision Tree (DT) learning (the explainable method we have selected as a measure for comparison with DEAr). Given that this first experiment uses ANNs for characterisation extraction (i.e. feature selection), we first compare the DEAr predictor with ANNs. Overall, we show that our method significantly outperforms DTs as well as ANNs, while being less sensitive to the size of the training dataset.

In Table 1 we report 5-fold cross-validation results, using weighted averages for each metric, for a stand-alone ANN (with a single hidder layer of size 22 or 30), for a combination of Autoencoder+ANN, and for DEAr.

Table 1: 5-fold cross-validation results for the mushroom dataset.

Hidden layer size 22	**Precision**	**Recall**	$\mathbf{F_1}$
DEAr	**0.97**	**0.96**	**0.958**
Autoencoder + ANN	0.938	0.894	0.878
ANN	0.934	0.888	0.87
Hidden layer size 30			
DEAr	**0.97**	**0.962**	**0.962**
Autoencoder + ANN	0.932	0.886	0.86
ANN	0.936	0.896	0.88

In the autoencoder we use sigmoid as activation function and binary cross entropy as loss function. We experiment with different

[3] archive.ics.uci.edu/ml/datasets/Mushroom

sizes (10, 22, 30, 50, as these were less than the original number of features, i.e. 126) for the hidden layer in the autoencoder/number of features in $\mathbb{F}$ but report results for $H \in \{22, 30\}$ (see Figure 2). We also experimented with tanh and ReLU as activations functions and with various optimizers, but report results for the best performing combinations of choices obtained using grid search.

The chosen ANN has one hidden layer and uses sigmoid as activation function and softmax to make predictions. The hyper-parameters were optimised using the Adam method [24] with learning rate 0.001. For Autoencoder+ANN, we use the learnt weights $W^{(1)}$ from the encoder, which we do not optimise during training, and softmax for classification. In both cases, we trained for 50 epochs or until the performance on the development set stopped improving.

As shown in Table 1, our method performs better than the two ANN approaches with differences in F_1 up to 8% when using a hidden layer size of 22, and up to 10% when using a hidden layer size of 30.

We also conducted experiments to test whether our method can better cope with smaller datasets than the Autoencoder+ANN method (arguably the better performing of the two end-to-end ANN methods). Hence we run experiments on 6000 randomly drawn examples and 5000 randomly drawn examples, respectively, from the original mushroom dataset, and tested on the remaining examples in the starting dataset. We repeated the experiments 5 times and report the average performances in Table 2. Here as well we use the learnt weights $W^{(1)}$ from the encoder in Autoencoder+ANN and softmax for prediction.

Table 2 finally compares our method with DTs. For DTs and DEAr alike, we use the learnt weights from the encoder to select the top 22 features as $\mathbb{F}$ and give $\mathcal{D}$ as discussed in Section 5.1. We used information gain for DTs.

The experiments on reduced datasets show that Autoencoder+ANN is less performing than our method and DTs. Our approach performs better than DTs throughout all experiments, with improvements in F_1 up to 20% with training set size of 6000 examples, and up to 12% with training set size of 5000 examples.

Table 2: Average of 5 runs of training on a reduced dataset and testing on the remaining examples.

Training set size: 6000, Testing set size: 2124			
Hidden layer size 22	**Precision**	**Recall**	**$\mathbf{F_1}$**
DEAr	**0.978**	**0.976**	**0.976**
Autoencoder + ANN	0.802	0.642	0.61
Decision Trees	0.858	0.774	0.762
Hidden layer size 30			
DEAr	**0.966**	**0.964**	**0.966**
Autoencoder + ANN	0.802	0.638	0.604
Decision Trees	0.852	0.772	0.766
Training set size: 5000, Testing set size: 3124			
Hidden layer size 22	**Precision**	**Recall**	**$\mathbf{F_1}$**
DEAr	**0.954**	**0.954**	**0.954**
Autoencoder + ANN	0.84	0.76	0.75
Decision Trees	0.876	0.828	0.826
Hidden layer size 30			
DEAr	**0.97**	**0.97**	**0.97**
Autoencoder + ANN	0.84	0.756	0.748
Decision Trees	0.886	0.844	0.844

6 DEAr FOR (ANNOTATED) IMAGES

The second empirical evaluation uses the pipeline in Figure 1 to provide predictions (and explanations thereof) for a two publicly available datasets of images, namely the CelebFaces Attributes (celebA)[30] and the Objects with Attributes (OwA)[15]. Both datasets are manually annotated with semantic features. CelebA is a large-scale collection of more than 200K celebrity images, providing 40 binary attributes for each image. OwA consists of the aPascal dataset (from PASCAL VOC2008 challenge[15])and amounts to 6340 training images with 64 attribute labels for each image. The images are divided in 20 classes. Figure 3 shows some examples of attribute annotated images from each dataset.

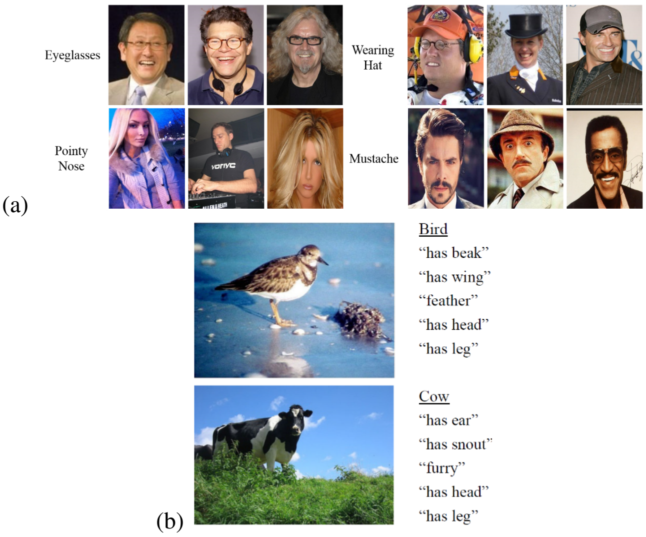

Figure 3: Examples of semantically annotated images from (a) celebA and (b) OwA datasets.

6.1 Characterisation Extractor

We use all attribute annotations in the two datasets as characterisations. Due to replications, this reduces the size of CelebA to 110K and the OwA dataset to 2.2K. The resulting datasets are both coherent. Since we have defined DEAr for binary classification only, we follow a one-vs-all strategy, labelling the images from celebA and OwA as Male/Female and person/non-person, respectively.

6.2 Choices of Parameters

We use the same choices of $\succcurlyeq$ and $\not\sim$ as in Section 5.2. However, the choice of the default argument (C_δ, δ) requires a careful selection as it is crucial in the model's predictive performance. The experiments show that in this setting the best performing choice is a synthetic datapoint which is sufficiently informative while containing the minimum number of annotated features. We thus choose as characterisation of the default argument a singleton set consisting of a singular attribute with the highest importance. We used univariate chi-squared statistic as our feature selection algorithm, returning the attributes "Wearing Lipstick" and "Skin" to be the features with the most significant impact on the predictive label for the two datasets. Hence, the default arguments are:

- $(C_\delta, \delta) = (\{$"Wearing Lipstick"$\}$,female)
- $(C_\delta, \delta) = (\{$"Skin"$\}$, person)

for celebA and OwA, respectively.

6.3 Empirical Evaluation

A comparison of the outcomes provided by DEAr and DTs is shown in Table 3, alongside a comparison with kNN (k=3). Here we choose kNN as a further measure of comparison as it naturally lends itself to prediction with these two datasets. For celebA, we split the 110K dataset into 5 batches and then perform 5-fold cross-validation for each batch, taking then the average over the outputs across the 5 batches. For OWa, we perform 5-fold cross-validation on the entire 2.2K dataset. Table 3 shows the results of the experiments for both datasets. Here DEAr has respectable but lower precision than both DTs and kNNs for both datasets, but higher or same recall than both methods, more so for celebA. Furthermore, DEAr has comparable prediction accuracy as DTs and kNNs for celebA and OwA.

Table 3: 5-fold cross validation results using DEAr, DTs and kNNs (k=3) on celebA and OwA.

celebA	**Precision**	**Recall**	$\mathbf{F_1}$	**Accuracy**
DEAr	0.79	1.00	0.88	0.89
DTs	0.93	0.87	0.90	0.90
kNNs	0.87	0.83	0.85	0.86
OwA				
DEAr	0.85	1.0	0.92	0.94
DTs	0.98	1.0	0.99	0.99
kNNs	0.97	0.99	0.98	0.99

7 DEAr FOR TEXT

The final empirical evaluation uses the pipeline in Figure 1 in a Natural Language Processing (NLP) task, namely Sentiment Analysis. This aims to analyse people's opinions or sentiments towards topics, items, etc. and their attributes [29, 28, 45].

We select the texts[4] used in evaluation in [26] that extracted and manually labeled 1000 sentences from Amazon reviews about cell phones and accessories category used in [33] and 1000 sentences from the IMDb movie review sentiment dataset used in [32]. These datasets consist of short sentences that can be clearly distinguished by humans as being positive or negative.

For each of the two websites, there are 500 positive and 500 negative sentences, selected so that they have a clear positive or negative polarity. Below are some examples from the dataset:
Amazon negative: I advise EVERYONE DO NOT BE FOOLED!
Amazon positive: So Far So Good!
Amazon negative: The commercials are the most misleading.
Amazon positive: I have yet to run this new battery below two bars and that's three days without charging.

IMDb negative: It had some average acting from the main person, and it was a low budget as you clearly can see.
IMDb positive: It's practically perfect in all of them a true masterpiece in a sea of faux "masterpieces".
IMDb negative: All in all, a great disappointment.
IMDb positive: This is definitely a cult classic well worth viewing and sharing with others.

7.1 Characterisation Extractor

As we now operate with text, instead of having categorical features as in the two previous empirical settings,we represent $\mathbb{F}_{\mathcal{L}}$, i.e. the vocabulary obtained from the lemmas in each of the datasets $\mathcal{L}$, by means of a one-hot-encoding, where 1 indicates the presence of words in the text, and 0 the absence. Since each dataset $\mathcal{L}$ is much smaller than the dataset from the previous experiments, an autoencoder is not fit to be used to select the most relevant features: we use instead an ensemble technique, namely Random Forests (RFs). We experiment with RFs directly on the datapoints as well as after a pre-processing stage o cluster related terms. For example, "nice" and "beautiful" can be clustered and replaced by a single word, e.g. "nice". We use semantic similarity for clustering as follows: for two words to be deemed semantically similar they need to have the same sentiment polarity (e.g. "good" will be tested against "great" which has a positive polarity but not against "bad" which has a negative polarity) and the same POS tag (e.g. "good" will be tested against "amazing" which is an adjective but not against "work" which can be a verb or a noun depending on context). We use the semantic network ConceptNet[5] and select the pairs that have relatedness score above 0.3. Table 4 shows examples of pairs of words and their similarity. Using our imposed threshold, we obtain a single cluster {great, amazing, terrific} and replace each of these words encountered in texts with the word *great*. Thus, we reduce the number of features from $|\mathbb{F}_{\mathcal{L}}| = 9$ to $|\mathbb{F}| = 7$.

Table 4: Examples of word similarity.

Word 1	**Word 2**	**Similarity score**
great	funny	-0.01
great	fine	0.007
great	interesting	0.115
great	good	0.184
great	important	0.283
great	superb	0.487
great	amazing	0.572
great	wonderful	0.577
great	awesome	0.598
great	terrific	0.653
art	work	0.132
art	fan	0.041
art	treasure	0.084

Whether we have clusters or not, we use RFs (with 150 random trees on various sub-samples of the dataset and averaging to improve the predictive accuracy and to control overfitting) with TF-IDF features of the training dataset. In the remainer, TF-IDF$_i$ stands for word i in the given set of "documents" (i.e. texts).

As an alternative, we also experiment with the following score for word i which represents a combination of TF-IDF and the polarity of synsets based on the POS tag (for S_i = synsets(i)):

$$\text{TF-IDF-SYN}_i = \text{TF-IDF}_i * \max(\text{pos}, \text{neg})$$

$$\text{pos} = \frac{\sum_{j \in S_i} \text{pos\_val}_j}{|S_i|}; \text{neg} = \frac{\sum_{j \in S_i} \text{neg\_val}_j}{|S_i|}$$

In the remainder, we will refer to the weight of word i computed using TF-IDF as w_{1i} = TF-IDF$_i$ and to the weight of word i computed using TF-IDF and synsets as w_{2i} = TF-IDF-SYN$_i$. Also, w will stand for any of w_{1i} and w_{2i} when either can be used.

7.2 Choices of Parameters

We experimented with several alternative choices of parameters:

[4] archive.ics.uci.edu/ml/datasets/Sentiment+Labelled+Sentences

[5] conceptnet.io

- $\succcurlyeq = \supseteq$ as in the previous two empirical settings as well as a new partial order (referred to as *upgraded* later) defined as $(X, o) \succcurlyeq (Y, o')$ iff $\sum_{x_i \in X} w(x_i) > \sum_{y_i \in Y} w(y_i) \wedge |X| > |Y|$;
- $(C_\delta, \delta) = (\{\}, \delta)$ for $\delta \in \mathbb{O} = \{+, -\}$ (we will see that different choices are best for the different datasets);
- $\not\sim = \not\supseteq$.

Thus, our choice of $\succcurlyeq$ replaces $\supseteq$ in the two previous empirical settings. We use this new notion of $\succcurlyeq$ on the basis that features are equipped with a weight (w) indicating their importance in classification. Intuitively, a datapoint is more informative than another with fewer and weaker features as it provides more information describing the sentiment polarity. For example, assuming the selected features are adjectives, the (positive) sentence *The movie was amazing, with a very good story and a great cast* is more informative than the (negative) sentence *Awful film* as the former has more information.

7.3 Empirical Evaluation

We experiment with various configurations and report the best results using 5 fold cross-validation in Table 5. For DTs, we used gini impurity criterion and TF-IDF features. For DEAr, we experimented with various numbers of selected features as well as other parameters such as the default used (0 for negative polarity and 1 for positive polarity), whether we used synsets when selecting features, clusters to group similar features, and the upgraded version of DEAr.

The average F_1 results with DEar using all features for IMDB is 72% and for Amazon 76.2%. For IMDB, DT and DEar trained on the selected features do not generally perform better, but the best results using DEar are better by 1.8%. For Amazon, DT with selected features does not improve the results when using all features, but DEar performs better than DT with all features, with improvements of up to 3.6%. In all cases when using selected features, DEar performs better than DT with improvements of 3% for IMDB (500 features) and 4.3% for Amazon (500 features). The best results are obtained with the upgraded version of DEar, thus considering the weight of feature importance when constructing the framework, no synsets when selecting features, and using clusters. The difference with respect to parameters is given by the default selected: negative in the case of IMDB and positive in the case of Amazon. Using the upgraded DEar generally performs better than the original version for both datasets. Synsets are an important aspect for Amazon, and a similar pattern can be seen for clusters.

Table 5: F_1 results for DTs and DEAr using 5-fold cross validation, varying $|\mathbb{F}|$, choice of default δ, use of **S**(ynsets), **C**(lusters), **U**(pgraded) $\succcurlyeq$.

	$\lvert\mathbb{F}\rvert$	**S**	δ	**C**	**U**	**DTs**	**DEAr**
IMDB	200	Y	1	N	N	69.4	70.7
	200	Y	0	N	Y	69.4	71.9
	200	N	0	N	Y	70.7	71.9
	200	N	0	Y	Y	72.5	73.8
	500	Y	0	N	N	67.1	69.4
	500	Y	0	N	Y	67.1	70.1
	800	Y	0	N	Y	67	69.8
Amazon	200	N	1	Y	N	75.8	77.1
	200	N	0	N	Y	74.1	77
	200	N	1	N	Y	74.1	77.6
	200	N	1	Y	Y	77	79.8
	500	N	1	Y	N	75.3	76.9
	500	N	1	Y	Y	75.4	79.7
	800	N	1	Y	Y	76	79.7

8 RELATED WORK

Recently, several works have suggested a hybrid approach aiming to provide human-interpretable explanations for a complex (black-box) model predictions using a transparent (white-box) counterpart model. A hybrid classifier is proposed in [35, 40] that combines KNN with CNNs, while DTs [16, 44], decision forests [25] are used as the transparent counterpart. Closest to DEAr is the proposal of [23], suggesting a "twin" system that uses CBR to provide plausible explanations to ANN's predictions. Rather than using CBR, we use an argumentative abstraction thereof.

Argumentation has been used extensively to generate explanations in AI, e.g. for explaining decisions [2, 47, 46] and recommendations [7, 36], in some cases using dispute trees as we envisage [17, 46]. We use argumentation to explain predictions that are also generated by argumentation. Two works are closest to ours: [3], using argumentation to perform concept learning and AA-CBR [9, 10], mining argumentation frameworks for case-based reasoning. Differently from us, [3] uses both datapoints and hypotheses as arguments, and bases attack only on classification disagreement (with the use of preferences over hypotheses). AA-CBR is a special case of our approach. Neither approach has been validated experimentally as we do.

Dispute trees have been used in explanations for example in [17, 46]. In particular, [17] formalised dialectical explanations for argument-based reasoning whereas [46] used argumentation to explain multi-criteria decision making obtained from dispute trees in a legal setting and [47] used dispute trees for explaining human-generated decisions (over the outcome of bills through the UK parliament). We suggested the use of dispute trees as a step towards a variety of explanations, to be developed as part of future work.

9 CONCLUSION

We have presented DEAr, a method inspired by AA-CBR [9, 10], to obtain argumentation debates as abstractions of the prediction problem from labelled datasets, including categorical, annotated images and text. Reasoning with these abstractions, as standard in argumentation in AI, gives competitive predictions which are also naturally explainable dialectically. We have shown experimentally that our method is competitive with (and sometimes outperforms) Decision Trees which are also explainable, using a variety of configurations for DEAr (that go well beyond AA-CBR).

The deployment of DEAr requires the combination of methods from argumentation in (symbolic) AI with components of standard data-centric approaches (e.g. feature selection and dimensionality reduction using autoencoders, statistical methods and random forests). We have focused on DEAr's predictive ability, and future work is needed to explore its explainability in full. The argumentation debates that it produces can be used as an explainable model per se, but can also serve as the starting point for various forms of explanation. We have illustrated preliminary notions of dialectical explanations based on dispute trees. We plan to study other types of explanations extracted from dispute trees, e.g. (as mentioned) counterfactual explanations.

We plan to conduct further experiments, e.g. with raw images and continuous features, as well as experimental evaluations with human users as to the amenability of our explanations in general and comparatively with Decision Trees. Indeed, in the presence of tens of arguments in our case or features in rules in the case of Decision Trees, the explanations may become complicated and hence affect interpretability.

REFERENCES

[1] Amina Adadi and Mohammed Berrada, 'Peeking inside the black-box: A survey on explainable artificial intelligence (xai)', *IEEE Access*, **6**, 52138–52160, (2018).

[2] Leila Amgoud and Henri Prade, 'Using Arguments for Making and Explaining Decisions', *Artificial Intelligence*, **173**(3-4), 413–436, (2009).

[3] Leila Amgoud and Mathieu Serrurier, 'Agents that argue and explain classifications', *Autonomous Agents and Multi-Agent Systems*, **16**(2), 187–209, (2007).

[4] Charles Antaki and Ivan Leudar, 'Explaining in conversation: Towards an argument model', *European Journal of Social Psychology*, **22**(2), 181–194, (1992).

[5] Katie Atkinson, Pietro Baroni, Massimiliano Giacomin, Anthony Hunter, Henry Prakken, Chris Reed, Guillermo Ricardo Simari, Matthias Thimm, and Serena Villata, 'Towards artificial argumentation', *AI Magazine*, **38**(3), 25–36, (2017).

[6] Olcay Boz, 'Extracting decision trees from trained neural networks', in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and Data Mining*, pp. 456–461. ACM, (2002).

[7] C. E. Briguez, M. C. Budán, C. A. D. Deagustini, A. G. Maguitman, M. Capobianco, and G. R. Simari, 'Argument-based mixed recommenders and their application to movie suggestion', *Expert Systems with Applications*, **41**(14), 6467–6482, (2014).

[8] Mark Craven and Jude W Shavlik, 'Extracting tree-structured representations of trained networks', in *Advances in neural information processing systems*, pp. 24–30, (1996).

[9] Kristijonas Čyras, Ken Satoh, and Francesca Toni, 'Abstract Argumentation for Case-Based Reasoning', in *Principles of Knowledge Representation and Reasoning, 15th International Conference*, pp. 549–552, (2016).

[10] Kristijonas Čyras, Ken Satoh, and Francesca Toni, 'Explanation for Case-Based Reasoning via Abstract Argumentation', in *6th International Conference on Computational Models of Argument*, pp. 243–254, (2016).

[11] Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository, 2017.

[12] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić, 'Explainable artificial intelligence: A survey', in *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, pp. 0210–0215. IEEE, (2018).

[13] Phan Minh Dung, 'On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games', *Artificial Intelligence*, **77**(2), 321 – 357, (1995).

[14] Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio, 'Why does unsupervised pretraining help deep learning?', *Journal of Machine Learning Research*, **11**, 625–660, (2010).

[15] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman, 'The pascal visual object classes (voc) challenge', *International journal of computer vision*, **88**(2), 303–338, (2010).

[16] Nicholas Frosst and Geoffrey Hinton, 'Distilling a neural network into a soft decision tree', *arXiv preprint arXiv:1711.09784*, (2017).

[17] Alejandro Javier García, Carlos Iván Chesñevar, Nicolás D. Rotstein, and Guillermo Ricardo Simari, 'Formalizing dialectical explanation support for argument-based reasoning in knowledge-based systems', *Expert Systems with Applications*, **40**(8), 3233–3247, (2013).

[18] Eduardo Gasca, José Salvador Sánchez, and R. Alonso, 'Eliminating redundancy and irrelevance using a new MLP-based feature selection method.', *Pattern Recognition*, **39**(2), 313–315, (2006).

[19] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi, 'A survey of methods for explaining black box models', *ACM Computing Surveys*, **51**(5), 93:1–93:42, (2019).

[20] Kai Han, Chao Li, and Xin Shi, 'Autoencoder feature selector', *CoRR*, **abs/1710.08310**, (2017).

[21] Geoffrey Hinton and Ruslan Salakhutdinov, 'Reducing the dimensionality of data with neural networks', *Science*, **313**(5786), 504 – 507, (2006).

[22] Ulf Johansson and Lars Niklasson, 'Evolving decision trees using oracle guides', in *2009 IEEE Symposium on Computational Intelligence and Data Mining*, pp. 238–244. IEEE, (2009).

[23] Eoin M Kenny and Mark T Keane, 'Twin-systems to explain artificial neural networks using case-based reasoning: comparative tests of feature-weighting methods in ANN-CBR twins for XAI', in *Twenty-Eighth International Joint Conferences on Artifical Intelligence (IJCAI)*, pp. 2708–2715, (2019).

[24] Diederik P. Kingma and Jimmy Ba, 'Adam: A method for stochastic optimization', in *3rd International Conference on Learning Representations, ICLR*, (2015).

[25] Peter Kontschieder, Madalina Fiterau, Antonio Criminisi, and Samuel Rota Bulo, 'Deep neural decision forests', in *Proceedings of the IEEE international conference on computer vision*, pp. 1467–1475, (2015).

[26] Dimitrios Kotzias, Misha Denil, Nando de Freitas, and Padhraic Smyth, 'From group to individual labels using deep features', in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 597–606, (2015).

[27] R Krishnan, G Sivakumar, and P Bhattacharya, 'Extracting decision trees from trained neural networks', *Pattern recognition*, **32**(12), (1999).

[28] Bing Liu, *Sentiment Analysis and Opinion Mining*, Synthesis Lectures on Human Language Technologies, 2012.

[29] Bing Liu, *Sentiment Analysis - Mining Opinions, Sentiments, and Emotions*, Cambridge University Press, 2015.

[30] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang, 'Deep learning face attributes in the wild', in *Proceedings of International Conference on Computer Vision (ICCV)*, (2015).

[31] Scott M Lundberg and Su-In Lee, 'A unified approach to interpreting model predictions', in *Advances in Neural Information Processing Systems*, pp. 4765–4774, (2017).

[32] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts, 'Learning word vectors for sentiment analysis', in *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, (2011).

[33] Julian J. McAuley and Jure Leskovec, 'Hidden factors and hidden topics: understanding rating dimensions with review text', in *Seventh ACM Conference on Recommender Systems, RecSys*, pp. 165–172, (2013).

[34] Tim Miller, 'Explanation in artificial intelligence: Insights from the social sciences', *Artificial Intelligence*, **267**, 1–38, (2019).

[35] Nicolas Papernot and Patrick D. McDaniel, 'Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning', *CoRR*, **abs/1803.04765**, (2018).

[36] Antonio Rago, Oana Cocarascu, and Francesca Toni, '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).

[37] Marco Túlio Ribeiro, Sameer Singh, and Carlos 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*, pp. 1135–1144, (2016).

[38] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin, 'Anchors: High-precision model-agnostic explanations', in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, (2018).

[39] Cynthia Rudin, 'Please stop explaining black box models for high stakes decisions', *arXiv preprint arXiv:1811.10154*, (2018).

[40] Sadiq Sani, Nirmalie Wiratunga, and Stewart Massie, 'Learning deep features for knn-based human activity recognition.', (2017).

[41] Frode Sørmo, Jörg Cassens, and Agnar Aamodt, 'Explanation in Case-Based Reasoning–Perspectives and Goals', *Artificial Intelligence Review*, **24**(2), 109–143, (2005).

[42] A. Verikas and M. Bacauskiene, 'Feature selection with neural networks', *Pattern Recognition Letters*, **23**(11), 1323–1335, (2002).

[43] Shuyang Wang, Zhengming Ding, and Yun Fu, 'Feature selection guided auto-encoder', in *Proceedings of AAAI*, pp. 2725–2731, (2017).

[44] Yongxin Yang, Irene Garcia Morillo, and Timothy M Hospedales, 'Deep neural decision trees', *arXiv preprint arXiv:1806.06988*, (2018).

[45] Jun Zhao, Kang Liu, and Liheng Xu, 'Sentiment analysis: Mining opinions, sentiments, and emotions', *Computational Linguistics*, **42**(3), 595–598, (2016).

[46] Qiaoting Zhong, Xiuyi Fan, Xudong Luo, and Francesca Toni, 'An explainable multi-attribute decision model based on argumentation', *Expert Systems with Applications*, **117**, 42–61, (2019).

[47] Kristijonas Čyras, David Birch, Yike Guo, Francesca Toni, Rajvinder Dulay, Sally Turvey, Daniel Greenberg, and Tharindi Hapuarachchi, 'Explanations by arbitrated argumentative dispute', *Expert Systems with Applications*, **127**, 141 – 156, (2019).