Towards a Framework for Evaluating Explanations in Automated Fact Verification

Neema Kotonya and Francesca Toni
Department of Computing
Imperial College London
{n.kotonya18,f.toni}@imperial.ac.uk

Abstract

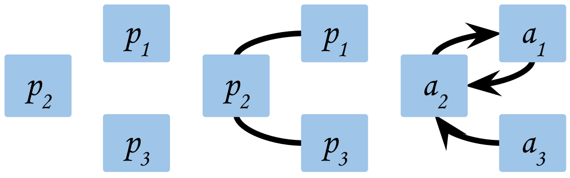
As deep neural models in NLP become more complex, and as a consequence opaque, the necessity to interpret them becomes greater. A burgeoning interest has emerged in rationalizing explanations to provide short and coherent justifications for predictions. In this position paper, we advocate for a formal framework for key concepts and properties about *rationalizing* explanations to support their evaluation systematically. We also outline one such formal framework, tailored to rationalizing explanations of increasingly complex structures, from *free-form* explanations to *deductive* explanations, to *argumentative* explanations (with the richest structure). Focusing on the *automated fact verification* task, we provide illustrations of the use and usefulness of our formalization for evaluating explanations, tailored to their varying structures.

Keywords: Automated Fact Verification, Explainable AI, Natural Language Explanations, Evaluation of Explanations, Properties of Explanations.

1. Introduction

In recent years, we have seen great performance success in natural language generation (NLG) and understanding (NLU), facilitated primarily by the use of sophisticated large language models (LLMs), e.g. LLaMA (Touvron et al., 2023). Despite these accomplishments, the complexity of these models calls for a greater need to interpret their computations. Interpretability of this kind would be desirable in numerous settings, e.g. some models are employed in safety and privacy critical applications (Deza et al., 2021), where it is important to understand whether these models are making the correct predictions for the right reasons (McCoy et al., 2019).

An increased focus on model interpretability has given way to several insightful works (Belinkov et al., 2020; Madsen et al., 2021), exploring several angles including examining model robustness through the use of perturbations (e.g. adversarial attacks) (Song et al., 2021), and generating natural language explanations and evaluating their faithfulness (Jacovi and Goldberg, 2020). The latter is the focus of this position paper. Explanations extracted for deep neural models' predictions take a variety of forms. Earlier work employed explanations in the form of attention heat-maps and highlighted tokens (Li et al., 2016). More recent work focuses on generating richer explanations, e.g. in graphical form (Thayaparan et al., 2021; Saha et al., 2021; Lampinen et al., 2022b), alongside techniques, similar to work in text generation and summarization, to obtain model faithful or label-consistent explanations (Kumar and Talukdar, 2020; Chrysostomou and Aletras, 2021).



(a) Free-form (b) Deductive (c) Argumentative

Figure 1: Abstract illustrations of the three classes of explanations explored in this paper (where the p_i are propositions and the a_j are arguments).

Here, we take the view that explanations are best understood as rationales for predictions, where these rationales are expressed in natural language, and typically extracted from the input text(s). The idea of rationalizing texts as explanations is discussed by (Lei et al., 2016) who deem rationales parts of the input text that are the most relevant for the predictions. This technique is similar to the processes humans undertake when delivering explanations to each other. We take the view that, depending on context and human preferences, rationales should be in one of three different formats, as illustrated in Figure 1.[1] These have already been proposed individually in some of the literature. For example Atanasova et al. (2020b) construct free-form explanations in the form of summaries; deductive explanations are employed by Krishna et al. (2022); and contrastive explanations which can be modeled as argumentative frameworks are employed by Derczynski et al. (2017); Gorrell et al. (2019); Schuster et al. (2021). We systematize

[1] Here, and in the remainder of this paper, illustrations are constructed by hand for space reasons, unless a source is explicitly given.

this line of work, by providing a paradigm in which we formally define three explanation formats, state their properties, and devise metrics for evaluating explanations of each format.

We focus on per-prediction *rationalizing explanations*, which are specific to and offer reasons for individual predictions. We ground our analysis and paradigm on automated fact verification, for two main reasons. First, evidence-based automated fact verification is a knowledge-intensive task, thus it is important to provide explanations for predictions providing a summary of this evidence (and counter-evidence documents, if any). Second, the nature of misinformation claims is often polarizing, emotive, and hyper-partisan (Potthast et al., 2018), and thus explanations that are rationalizing can aid in gaining the end user's trust in the system delivering the predictions. Rationales may amount to free text (e.g. as in (Camburu et al., 2018)), or structured descriptions (e.g. as in (Tafjord et al., 2021)), possibly with an argumentative flavor (e.g. as in Kotonya and Toni (2019), Schuster et al. (2021), Chen et al. (2021), and Dougrez-Lewis et al. (2022)), present explanations as debates including competing (and possibly contrasting) arguments.

Concretely, we make the following contributions.

- First, we offer formal definitions of terms frequently employed in the literature on explanations for neural NLP. We present a paradigm for conceptualizing rationale-based explanations, expanding on (Wiegreffe and Marasovic, 2021), and viewing explanations by degree of structure (see Figure 1).
- Second, we argue the case for concrete metrics for evaluating such explanations based on properties. To this end, we propose and define several desirable properties for evaluating *free-form*, *deductive*, and *argumentative* explanations. We then offer means for employing said properties for evaluation in empirical settings.

Note that, while we restrict attention to (the evaluation of) explanations for predictions for the task of automated fact verification, the formalism, properties and metrics that we introduce can in principle be employed for evaluating explanations irrespective of the underlying task.

2. Related Work

The need for an explanation of NLP prediction tools is well established (Doshi-Velez and Kim, 2017; Ribeiro et al., 2018; Gohel et al., 2021) and rationalizations, as explanations are advocated by several (Rajani et al., 2019; DeYoung et al., 2020). In the context of automated fact verification, several bespoke forms of explanations have been proposed (see (Kotonya and Toni, 2020a) for an overview), including rationalizations as explanations (Rana et al., 2022; Si et al., 2023). Despite the many advances in rationalizing explainable NLP, we still observe the following shortcomings in the existing literature.

First, there are no agreed-upon definitions of what constitutes an explanation for an NLP prediction or the preferred methods for generating explanations. Many approaches are taken when characterizing and generating rationales for explanations. For example, DeYoung et al. (2020) describes rationales as *snippets that support the outputs* of a model. Wadden et al. (2020) consider rationales to be *a minimal collection of sentences* the sum of which implies the veracity of a claim. Schuster et al. (2021) extend this idea to include both supporting and contrastive evidence, i.e., rationales can favor one verdict or support an alternative verdict. Ross et al. (2021) explore explanations as edits, contrastive explanations in this case amount to the edits to the inputs which an alternative output. Contrastive explanations are closely related to counterfactual explanations (Guidotti, 2022).

Second, there is considerable work on identifying and evaluating properties of explanations in NLP (Jacovi and Goldberg, 2020; Atanasova et al., 2023) but the focus of existing works is not on the evaluation of rationalizing explanations and prioritize properties can be interpreted as user requirements and are related to the relationship between the explanation and model prediction, e.g. faithfulness (Jacovi and Goldberg, 2020; Atanasova et al., 2023), robustness (Datta et al., 2021) and sufficiency (Chrysostomou and Aletras, 2022). Instead, the focus of our work is on properties related to explanation form, i.e. what new properties emerge as we enrich the structure of an explanation? As there has been keen interest in the evaluation of deep NLP models (Ribeiro et al., 2020), it would be valuable for this to extend to the evaluation of rationalizing explanations.

Third, there has been little effort to define and formalize a rigorous set of desirable criteria specific to rationalizing explanations. Some recent examples exist (Nauta et al., 2023), but they do not focus on rationale-based explanations for fact-checking as we do and take a high-level approach to discuss properties, whereas we offer concrete definitions. An example of some effort in this direction for a free-form explanation for fact-checking is given in (Kotonya and Toni, 2020b). Atanasova et al. (2020a) perform a diagnostic study of explainability for text classification concerning several properties. However, their focus is not specifically on rationalizing explanations; also they do not consider explainability in the automated fact verification context. This paper aims to outline a direction for addressing these issues.

3. Definitions

We define three classes of NLP (rationale-based) explanations, as abstractions of explanations found in the literature. The three classes amount to *free-form* explanations (§ 3.1); *deductive* explanations (§ 3.2), e.g. chains of facts as in Yang et al. (2018); and *argumentative* explanations (§ 3.3), e.g. providing reasons for supporting or refuting a claim as in Wadden et al. (2020) and Schuster et al. (2021). We assume that the end users of our proposed framework are humans, and this motivates us to consider explanation formats and properties that align closely with human explanations, following established views in Explainable AI (XAI), most notably by Miller (2019, 2021).

3.1. Free-form Explanations

Free-form explanations are the most common explanatory outputs in the rationale-based landscape and reflect a large part of earlier literature on explaining deep neural models (Wiegreffe and Marasovic, 2021). We define them abstractly below, using a generic notion of *proposition*. Note that, since we want to provide as general guidance as possible to designers of explanations, we do not place any stipulations on the nature of propositions but, in practice, the decision of what to admit as propositions needs to be taken before free-form explanations are drawn from models. Concretely, propositions could amount, for example, to words or phrases (e.g. occurring in the input text being explained), model predictions (e.g. that the model predicts that the input is true), and tokens understood by the underlying models.

Definition 1 *A* ***free-form explanation*** *amounts to a finite, non-empty sequence of propositions. A* ***free-form explanation for a model's prediction, given an input****, is a free-form explanation that includes, among its propositions, some elements of the input and the prediction itself. We use the notation $\mathcal{P} \rightsquigarrow_f [m(\mathbf{X}) = \hat{y}]$ to indicate that $\mathcal{P}$ is a free-form explanation for prediction $\hat{y}$ by model m, given input $\mathbf{X}$.*

The inclusion of the prediction in a free-form explanation relates to "relevance" thereof to the prediction. This prediction is often implicit, as in the example presented in Table 1.

Note that we do not enforce that $\mathcal{P}$ is restricted to elements in the input $\mathbf{X}$ and indeed, in general, it could also include elements not in $\mathbf{X}$, as in the illustration in Table 1. Free-form explanations may take several concrete forms in practice. Wiegreffe and Marasovic (2021) distinguish between highlights, e.g. token-wise saliency maps, and free-form text. Neural attention is used to create saliency maps (Li et al., 2016). Instead, when free-form natural language explanations are used, a sequence-to-sequence model is typically employed to generate a text that serves as the rationale for the predictions. In this setting, Camburu et al. (2018) explore two paradigms: one which jointly generates explanation and prediction, and another which first generates the explanation and then the prediction. As a further example, Kumar and Talukdar (2020) looks to generate label-specific explanations for each possible label prediction. Furthermore, Kotonya and Toni (2020b)'s explanations as summaries could be seen as a form of free-form explanation. Most explanations of this type are self-generated, i.e. the model is expected to both predict outputs *and* explain its reasoning, e.g. by way of a prompt-based model (Narang et al., 2020; Marasović et al., 2021).

A popular Facebook post about the life and death of British mathematician Alan Turing is truthful.

Verdict: `Mostly True`
Explanation:
The popular Facebook post got most of the facts right (p_1). However, there's no evidence that Turing inspired the design of the Apple computer company's logo (p_2). Also, Turing's death in 1954 deserves further examination than what was provided in the post, which we included below (p_3).

Table 1: Example of free-form explanation, matching the abstract illustration in Figure 1(a). Here, the claim, prediction (Verdict (label)) and explanation (except for the p_i, which are our addition) are taken from the claim verification platform Snopes `https://www.snopes.com/fact-check/alan-turing-facebook-post/`.

However, there is an ongoing debate regarding whether explanations of this form are sufficient (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019; Pruthi et al., 2020), to which we contribute by characterizing other forms, in § 3.2-3.3 below.

3.2. Deductive Explanations

Like free-form explanations, deductive explanations also consist of propositions, but these propositions are "connected" by a (binary) *relation*. Again, since we aim to provide as general guidance as possible to designers of explanations, we do not place any stipulations on the nature of $\mathcal{R}$ and keep it abstract, but, in practice, the decision of what to admit as relation needs to be taken before deductive explanations are drawn from models. Concretely, for example, this relation could be support between propositions, or relevance between propositions, or a proposition being in relation with another could amount to the former being a reason for the latter or for the two to have something in common. Deductive explanations also have strong parallels with the reasoning obtained through chain-of-thought (Wei et al., 2022a), tree-of-thought (Yao et al., 2023), and other similar graph topological large language model prompting techniques.

Definition 2 *A* ***deductive explanation*** *amounts to a pair composed of a finite, non-empty sequence of propositions and a binary relation over the propositions. A* ***deductive explanation for a model prediction, given an input****, is a deductive explanation that includes, among the pair described above its components, some elements of the input as well as the prediction itself. We use the notation* $\langle \mathcal{P}, \mathcal{R} \rangle \rightsquigarrow_d [m(\mathbf{X}) = \hat{y}]$ *to indicate that* $\langle \mathcal{P}, \mathcal{R} \rangle$ *is a deductive explanation for prediction* $\hat{y}$ *by model* m*, given input* $\mathbf{X}$*.*

p_3: A daffodil plant can live for more than two years.

Verdict: `Verified`
Explanation $\langle \mathcal{P}, \mathcal{R} \rangle$, where:
$\mathcal{P} = \{p_1, p_2, p_3\}$, for:
p_1: Daffodil is the common name for plants of the narcissus genus, which are perennial.
p_2: A perennial plant has a minimum life span of two years.
$\mathcal{R} = \{(p_1, p_2), (p_2, p_3)\}$.

Table 2: Example of deductive explanation, matching the abstract illustration in Figure 1(b).

The example in Table 2 gives an illustration of deductive explanation: here and later, we represent $\mathcal{R}$ as a set of pairs, so, for example, $(p_1, p_2) \in \mathcal{R}$ indicates that p_1 and p_2 are related by $\mathcal{R}$. Here, $\mathcal{R}$ may be seen as a logical reasoning chain or as a linking of propositions by common entities. In general, several other possibilities could be considered for identifying $\mathcal{R}$, e.g. chronological ordering of evidence.

As in the case of free-form explanations, we enforce "relevance", ensuring the prediction is in $\mathcal{P}$ (but, again, we may have that the prediction is implicit, as in Table 2).

Also, we may impose a direction in $\mathcal{R}$ or not, if we want to capture bidirectionality as in the case of $\mathcal{R}$ representing that propositions have something in common. Furthermore, a compound layered deductive explanation could be acquired by considering multiple semantics for $\mathcal{R}$: we leave this as future work.

A clear form of deductive explanations in the literature is *chains of connected facts* (Inoue et al., 2020; Tafjord et al., 2021). The example we present in Table 2 amounts to a chain of facts because there is a sequence of propositions $p_1 \rightarrow p_2 \rightarrow p_3$ that leads from the proposition p_1 to the claim (p_3). Propositions p_1 and p_2 provide evidence for which the logical conclusion is p_3, thus this chain of facts justifies the claim p_3. Chains of facts are analogous to chain-of-thought prompting (Wei et al., 2022b; Lampinen et al., 2022a). In our view, the reasoning output produced by the chain-of-thought process amounts to a deductive explanation.

3.3. Argumentative Explanations

We now examine explanations that provide justifications for model predictions using *arguments*, as opposed to simple propositions. Intuitively, an argument consists of a conclusion that is supported by premises. In particular, we can choose premises and conclusions of arguments to be propositions, understood broadly as in the definitions of free-form and deductive explanations.

Since the kinds of arguments used in argumentative explanations are expressed in natural language, we do not place any stipulations on the logical connection between premises and the conclusion of an argument. In particular, arguments could be enthymemes (Razuvayevskaya and Teufel, 2017) with partially specified or even empty premises.

Arguments are the building blocks of debates. For this work, in the spirit of (Dung, 1995; Atkinson et al., 2017), we represent debates as argumentation frameworks, modeling the interactions between arguments as relations. Specifically, we focus on bipolar argumentation frameworks (Cayrol and Lagasquie-Schiex, 2005), where an argument can be attacked or supported by arguments, thus modeling both conflict and agreement (respectively) between arguments. In the spirit of (Cayrol and Lagasquie-Schiex, 2005), we leave the definition of what conflict or agreement may mean completely unspecified, assuming instead that they are captured by abstract relations. We will return to them later in Definition 4.

Definition 3 *An* ***argumentative explanation*** *is given by a 3-tuple which amounts to a finite, non-empty set of arguments, a binary attack relation over the set of arguments and a binary support relation over the set of arguments. An* ***argumentative explanation for a model prediction, given an input****, is an argumentative explanation that includes, among its arguments, attack and support relations, some elements of the input as well as the prediction itself. We use the notation* $\langle \mathcal{A}, \mathcal{R}_{\mathit{Sup}}, \mathcal{R}_{\mathit{Att}} \rangle \rightsquigarrow_a [m(\mathbf{X}) = \hat{y}]$ *to indicate that* $\langle \mathcal{A}, \mathcal{R}_{\mathit{Sup}}, \mathcal{R}_{\mathit{Att}} \rangle$ *is a an argumentative explanation for prediction* $\hat{y}$ *by model*

m, given input $\mathbf{X}$.

Definition 3 could be realized so that at least one argument exists that admits as its conclusion the output prediction $\hat{y}$ and that some arguments need to admit propositions in the input among their premises. If there is just one argument in the argumentative explanation then these restrictions impose that the argument 'connects' the input and the prediction, i.e. there is a way to reason from the inputs which forms a rationale for the prediction.

Definition 3 leaves the attack/support relations unspecified. These could be defined in several ways, including as follows, making use of generic notions of *contradiction* and *implication*, as in *textual entailment* (Dagan et al., 2005).

Definition 4 *Let $a_i, a_j \in \mathcal{A}$ where $\mathcal{A}$ is a set of arguments. Then:*

- *a_j **attacks** a_i (**by undercutting**) if the conclusion of a_j is in contradiction with some premise(s) in a_i;*

- *a_j **attacks** a_i (**by rebutting**) if the conclusion of a_j is in contradiction with the conclusion of a_i;*

- *a_j **supports** a_i (**by providing reasons**) if the conclusion of a_j implies some premise(s) of a_i;*

- *a_j **supports** a_i (**by accrual**) if the conclusion of a_j implies the conclusion of a_i.*

Note that this definition borrows some concepts from the literature on argumentation frameworks, specifically, the notion of accrual (Prakken, 2005), undercutting (Bex et al., 2003), and rebuttal (Kowalski and Toni, 1996). Note also that typically attack by rebutting and support by accrual will be symmetric (i.e. if a_j attacks by rebuttal or supports by accrual a_i, then a_i does so towards a_j). For an illustration of argumentative explanations and notions of attack/support, consider Table 3.

Here, arguments $\mathcal{A} = \{a_1, a_2, a_3\}$, and several argumentative explanations $\langle \mathcal{A}, \mathcal{R}_{\text{Sup}}, \mathcal{R}_{\text{Att}} \rangle$ are possible, including where:

1. $\mathcal{R}_{\text{Sup}} = \{(a_3, a_2)\}$ and $\mathcal{R}_{\text{Att}} = \emptyset$; here, argument a_3 is reinforcing argument a_2, we have adopted a view of support as providing a rationale to justify the conclusion of an argument; there are no attacking arguments in this particular explanation;

2. $\mathcal{R}_{\text{Sup}} = \{(a_3, a_2), (a_1, a_2), (a_2, a_1)\}$ and (again) $\mathcal{R}_{\text{Att}} = \emptyset$; here, arguments a_1 and a_2 corroborate one another (by accrual) in support of the output of the classifier. An abstract depiction of this argumentative explanation is shown in Figure 1(c), with edges representing support.

Consider the following claim for fact-checking:

The King of the United States of America lives in the White House.

Also, consider the following arguments in the context of some argumentative explanation for prediction `Refuted` in verdict to the claim:

a_1: the King of the USA does not live in the White House ($\hat{y}$) because the USA has no king, as it is a republic (p_1).
a_2: The White House is the official residence of the USA President (p_2), thus it can not be the official residence of a king or any other head of state ($\hat{y}$).
a_3: The head of state of the USA is the President (c_3) because the title "president" is typically given to the head of a republic (p_3).

These arguments can also be represented symbolically as pairs (consisting of premises and claims): $a_1 = (\{p_1\}, \hat{y}), a_2 = (\{p_2\}, \hat{y}), a_3 = (\{p_3\}, c_3)$.

Table 3: Concrete example of arguments in an argumentative explanation.

Note that in this example the explanation includes an argument (a_3) which is neither for the prediction nor for any alternative predictions. In some settings, argumentative explanations could be restricted to make sure that the conclusion of each argument in $\mathcal{A}$ must imply a prediction from the set of possible outputs for the model. For these types of argumentative explanations, which we may call *flat* if there are arguments for different outcomes than computed by the classifier, the attack relation would be non-empty and include some attacks by a rebuttal. This may be the case, for example, in (Wadden et al., 2020; Schuster et al., 2021). Whereas it is clear why we may want to include support, some considerations about the inclusion of attack in argumentative explanations are in order. We include an attack to reflect two scenarios:

1. First, the possibility of conflicting evidence the model found in the input, giving reasons for undercutting or rebutting other reasons;

2. Second, to represent the fact/foil relationship (Barnes, 1994), given that attacks can distinguish between inputs which contribute to the prediction from distractor inputs that do not; here, attacks are of the rebuttal variety (as they point to contradictory predictions).

In other words, attacks may be needed to explain a low-confidence prediction from a model.

4. Properties

We define a number of properties for our forms of rationalizing explanations. Our list of properties is not exhaustive, and we see each of the properties as a useful criterion for assessing the validity of explanations in rationalizing a model's prediction in NLP. We consider separately properties tailored to free-form, structured and argumentative explanations (§ 4.1, 4.2 and 4.3, respectively).

4.1. Free-Form Properties

We adapt the properties for free-form explanations introduced by Kotonya and Toni (2020b) for evaluating explainable summaries (a form of free-form explanation) for automated fact verification. Here, we propose a single property for any free-form explanations: *coherence*. We define the notion of coherence in terms of a notion of logical contradiction (which could amount to the implication of negation), in line with the definitions in (Kotonya and Toni, 2020b).

Definition 5 *A free-form explanation $\mathcal{P}$ satisfies **coherence** if there exists no contradictory subset of propositions in $\mathcal{P}$.*

Thus, coherence is a measure of the cohesiveness of propositions in a free-form explanation. For coherence to hold, in particular, any two propositions in an explanation must not contradict one another, i.e. there is no pairwise disagreement between propositions which make up the explanation. More generally, our definition excludes contradictions involving any number of propositions making up the explanation.

4.2. Deductive Properties

The coherence property for free-form explanations is still applicable to deductive explanations $\langle \mathcal{P}, \mathcal{R} \rangle$ on the set of propositions $\mathcal{P}$. In addition, we identify four bespoke properties for deductive explanations: *non-circularity*, *(weak* and *strong) relevance*, and *non-redundancy*, defined below.

Definition 6 *A deductive explanation $\langle \mathcal{P}, \mathcal{R} \rangle$ is **non-circular** if there does not exist a proposition p_i in $\mathcal{P}$ and a set of propositions $\mathcal{P}'$ from $\mathcal{P}$, $\mathcal{P}' = \{p'_1, \dots, p'_k\} \subseteq \mathcal{P}$, such that $\{(p'_1, p'_2), \dots, (p'_{k-1}, p'_k)\} \subseteq \mathcal{R}$ and $p'_1 = p'_k = p_i$.*

Thus, non-circularity amounts to acyclicity of the $\mathcal{R}$ component of deductive explanations (when seeing them as directed graphs with propositions as nodes and elements of the relation as edges, as in Figure 1). It seeks to avoid circular explanations, which are not sound in a rhetorical sense.

Definition 7 *A deductive explanation $\langle \mathcal{P}, \mathcal{R} \rangle$ such that $\langle \mathcal{P}, \mathcal{R} \rangle \rightsquigarrow_d [m(\mathbf{X}) = \hat{y}]$ is **strongly relevant** if all propositions p_i in $\mathcal{P}$ are such that $(p_i, \hat{y})$ is in $\mathcal{R}$.*

Thus, all propositions in a strongly relevant deductive explanation are directly connected to the model's prediction.

Definition 8 *A deductive explanation $\langle \mathcal{P}, \mathcal{R} \rangle$ such that $\langle \mathcal{P}, \mathcal{R} \rangle \rightsquigarrow_d [m(\mathbf{X}) = \hat{y}]$ is **weakly relevant** if, for all propositions p_i in $\mathcal{P}$, there exists a set of propositions $\mathcal{P}'$ in $\mathcal{P}$, $\mathcal{P}' = \{p'_1, \dots, p'_k\} \subseteq \mathcal{P}$, such that $\{(p'_1, p'_2), \dots, (p'_{k-1}, p'_k)\} \subseteq \mathcal{R}$, $p'_1 = p_i$ and $p'_k = \hat{y}$.*

Namely, for a deductive explanation to be weakly relevant, each proposition needs to be connected by some chain (path) to the prediction. Thus, in weakly relevant deductive explanations there are no unconnected propositions. Note that we could easily define additional versions of relevance, e.g. to enforce links to input propositions (we refrain from doing so for lack of space).

Non-redundancy, the last property we define for deductive explanations, requires that no superfluous propositions are contained in an explanation.

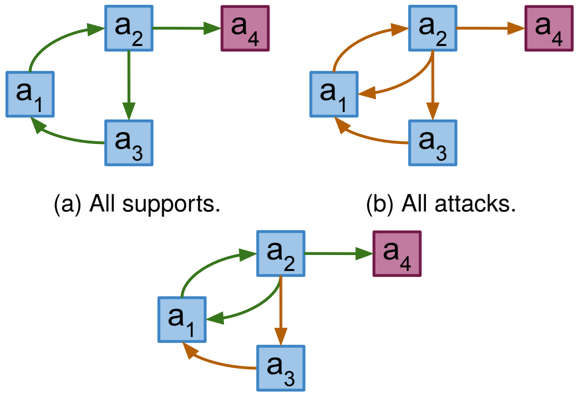
Definition 9 *A deductive explanation $\langle \mathcal{P}, \mathcal{R} \rangle$ such that $\langle \mathcal{P}, \mathcal{R} \rangle \rightsquigarrow_d [m(\mathbf{X}) = \hat{y}]$ is **non-redundant** iff for all propositions $p_i \in \mathcal{P} \setminus \{\hat{y}\}$, for $\mathcal{P}' = \mathcal{P} \setminus \{p_i\}$ and $\mathcal{R}' = \mathcal{R} \cap (\mathcal{P}' \times \mathcal{P})$, the pair $\langle \mathcal{P}', \mathcal{R}' \rangle$ is not a deductive explanation for $\hat{y}$ by m, given $\mathbf{X}$, i.e. $\langle \mathcal{P}', \mathcal{R}' \rangle \not\models m(\mathbf{X}) \rightarrow \hat{y}$.*

In other words, no proposition can be eliminated from a non-redundant explanation while still rationalizing the prediction for which it is intended. Note that, when eliminating a proposition from a deductive explanation, we also delete from the relation component all connections to and from deleted propositions. Note also that, in practice, by our definition of deductive explanation, $\langle \mathcal{P}', \mathcal{R}' \rangle \not\models m(\mathbf{X}) \rightarrow \hat{y}$ means that either $\hat{y}$ is not in $\mathcal{P}'$ or $\mathcal{P}'$ contains no elements of the input $\mathbf{X}$.

4.3. Argumentative Properties

The coherence property for free-form explanations, presented in § 4.1, could be enforced on the premises of arguments in argumentative explanations (e.g. by seeing arguments with premises $\mathcal{P}$ as free-form explanations $\mathcal{P}$). Furthermore, we could enforce properties similar in spirit to those for deductive explanations, presented in § 4.2, to individual arguments in argumentative explanations, (e.g. by seeing arguments with premises $\mathcal{P}$ and conclusion c as deductive explanation $\langle \mathcal{P}, \{(p, c) \mid p \in \mathcal{P}\} \rangle$). Here, we focus instead on additional properties of argumentative explanations. Specifically, we identify a number of properties regarding relations between arguments.

First, we can impose natural conditions on the attack and support relations, i.e. that there are no cycles therein:



(a) All supports.

(b) All attacks.

(c) Mixed supports and attacks.

Figure 2: Examples of dialectical circularity for three argumentative explanations. Attacks are shown in orange and supports are shown in green. Argument a_4 with conclusion $\hat{y}$ is purple.

Definition 10 *An argumentative explanation $\langle \mathcal{A}, \mathcal{R}_{\mathit{Sup}}, \mathcal{R}_{\mathit{Att}} \rangle$ is **dialectically non-circular** if there is no argument a_i in $\mathcal{A}$ and no set of arguments $\mathcal{A}'$ from $\mathcal{A}$, $\mathcal{A}' = \{a'_1, \ldots, a'_k\} \subseteq \mathcal{A}$, such that $a'_1 = a'_k = a_i$, and $\{(a'_1, a'_2), \ldots, (a'_{k-1}, a'_k)\} \subseteq (\mathcal{R}_{\mathit{Sup}} \cup \mathcal{R}_{\mathit{Att}})$.*

Violation of dialectical non-circularity would entail a cycle of supports and/or attacks. The specific case of violation of dialectical non-circularity, when some argument in the explanation is self-supporting, may amount to a situation of an argument that is not grounded in evidence. The specific case of violation of dialectical non-circularity by a cycle of supports (Figure 2(a)) may amount to an unsound debate in a rhetorical sense. The specific case of violation of dialectical non-circularity when some argument in the explanation is self-attacking may amount to a paradoxical situation of the argument being self-contradictory. The specific case of violation of dialectical non-circularity by a cycle of attacks (Figure 2(b)) may also amount to an unsound debate in a rhetorical sense. Mixed cases, as in Figure 2(c) are also challenging from a rhetorical perspective.

In addition, we can demand that argumentative explanations satisfy other dialectical properties, in the spirit of various argumentation frameworks from symbolic AI, notably abstract argumentation (Dung, 1995), bipolar argumentation (Cayrol and Lagasquie-Schiex, 2005) and quantified bipolar argumentation (Baroni et al., 2019). These frameworks rely upon notions of *acceptability* of sets of arguments (Dung, 1995) or *dialectical strength* for arguments (Baroni et al., 2019). In our setting, properties inspired by these notions can be used to point towards the explanations' credibility in the context of the confidence of the underlying model in the prediction. Intuitively, a credible (or strong) argument is supported by other (credible) argument(s). Conversely, a less credible (weakened) argument is attacked by (credible) arguments. In this spirit, we deem an argumentative explanation *dialectically faithful* if its credibility reflects the prediction confidence. We formalize this property using a generic notion of dialectical strength for arguments (Baroni et al., 2019):

Definition 11 *An argumentative explanation $\langle \mathcal{A}, \mathcal{R}_{\mathit{Sup}}, \mathcal{R}_{\mathit{Att}} \rangle$ such that $\langle \mathcal{A}, \mathcal{R}_{\mathit{Sup}}, \mathcal{R}_{\mathit{Att}} \rangle \rightsquigarrow_a [m(X) = \hat{y}]$ is **dialectically faithful** if*

- *whenever m gives $\hat{y}$ with top confidence, $\mathcal{R}_{\mathit{Att}}$ is such that there are no arguments in $\mathcal{A}$ attacking by rebutting any argument for $\hat{y}$;*
- *whenever m gives $\hat{y}$ with high confidence, the dialectical strength of the arguments in $\mathcal{A}$ with conclusion $\hat{y}$ is higher than the dialectical strength of the arguments attacking them (as per $\mathcal{R}_{\mathit{Att}}$);*
- *whenever m gives $\hat{y}$ with low confidence, $\mathcal{A}$ must either include only dialectically weak arguments with conclusion $\hat{y}$ or include some arguments attacking by rebutting some argument for $\hat{y}$ with higher dialectical strength than arguments in $\mathcal{A}$ with conclusion $\hat{y}$, if any.*

Intuitively, the argument for a prediction with high confidence should be supported by strong arguments (as in the case of the argumentative explanations outlined when discussing Table 3). Furthermore, consider the argumentative explanations in Figure 3. Here, the argument a_4 for the prediction with top confidence (left-most argumentative explanation) is not attacked, thus the explanation is dialectically faithful. Note that rebuttals imply contradiction between arguments' conclusions and are thus singled out in our definition of dialectical faithfulness. Also, in Figure 3, the argument a_4 for the prediction with high confidence (middle argumentative explanation), being again unattacked, trivially has a higher dialectical strength than its attackers, so, again, the explanation is dialectically faithful. Finally, the argument a_4 for the prediction with low confidence (right-most argumentative explanation) is supported by argument a_2 which is weakened by the attack from a_3; the latter also weakens the support from a_1 to a_2; thus, overall a_2 is a weak argument and the argumentative explanation can be deemed dialectically faithful.

In the special case when dialectical strength is "binary" (in that it sanctions an argument as winning or losing, e.g. as in Dung (1995)), we can refine dialectical faithfulness to define a notion of *acceptability* as follows:

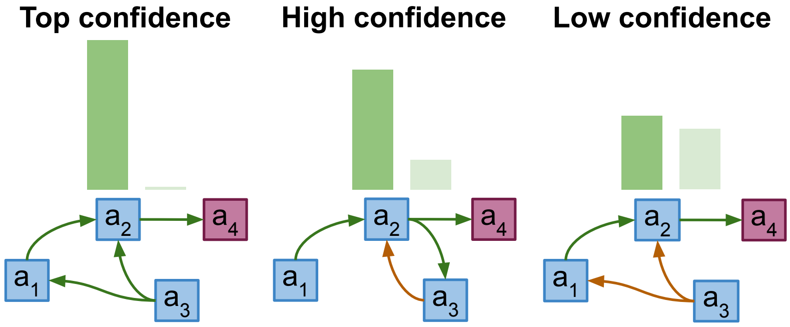


Figure 3: An illustration of argumentative explanations for top, high, and low confidence (binary) predictions. Attacks are shown in orange and supports are shown in green. Argument a_4 with conclusion $\hat{y}$ is purple.

Definition 12 *An argumentative explanation* $\langle \mathcal{A}, \mathcal{R}_{\mathit{Sup}}, \mathcal{R}_{\mathit{Att}} \rangle$ *such that* $\langle \mathcal{A}, \mathcal{R}_{\mathit{Sup}}, \mathcal{R}_{\mathit{Att}} \rangle \rightsquigarrow_a [m(\mathbf{X}) = \hat{y}]$ *is* ***acceptable*** *if*

- *whenever m gives $\hat{y}$ with top or high confidence, there exists no $\mathcal{A}' \subseteq \mathcal{A}$ containing all arguments in $\mathcal{A}$ with conclusion $\hat{y}$ such that for all arguments $a_i \in \mathcal{A}'$, if $a_j \in \mathcal{A}$ attacks a_i (i.e. $(a_j, a_i) \in \mathcal{R}_{Att}$), then there exists $a_k \in \mathcal{A}'$ attacking a_j (i.e. $(a_k, a_j) \in \mathcal{R}_{Att}$); in simpler terms, $\mathcal{A}'$ defends itself against all attacking arguments;*
- *whenever m gives $\hat{y}$ with bottom or low confidence, for $\mathcal{A}' \subseteq \mathcal{A}$ containing all arguments in $\mathcal{A}$ with conclusion $\hat{y}$, there exists some arguments $a_i \in \mathcal{A}'$ and $a_j \in \mathcal{A} \setminus \mathcal{A}'$ such that a_j attacks a_i (i.e. $(a_j, a_i) \in \mathcal{R}_{Att}$) but there exists no argument $a_k \in \mathcal{A}'$ such that a_k attacks a_j (i.e. $(a_k, a_j) \in \mathcal{R}_{Att}$); in simpler terms, $\mathcal{A}'$ cannot defend itself against all attacking arguments.*

The explanation presented in Figure 1(c) with the edges representing support satisfies acceptability (as there are no attacks). The left-most and middle explanations in Figure 3 are acceptable as there is no argument attacking a_4, the only argument with conclusion $\hat{y}$. Instead, the right-most explanation in Figure 3 is not acceptable, as there is no argument attacking a_4 in the explanation. Note that the definition of acceptable could be extended to allow for chains of support to provide a defense, inspired by (Cayrol and Lagasquie-Schiex, 2005).

5. Evaluation Metrics

We propose some metrics for evaluating empirically rationalizing explanations for NLP models, drawn from properties introduced in § 4.[2] We focus only on sample properties for lack of space, but metrics for other properties are also possible.

5.1. Free-Form Evaluation

We devise a metric for free-form explanation, Coh, relating to the property of coherence defined in § 4.1. For a free-form explanation $\mathbf{E} = \mathcal{P}$ such that $\mathcal{P} \rightsquigarrow_f m(\mathbf{X}) \rightarrow \hat{y}$, let $N = |\mathcal{P}|$ and N' be the number of subsets of $\mathcal{P}$. Then, violation of coherence can be measured by $\textsc{Coh}(\mathbf{E})$ as shown in Eq. 1, where $contr(x, x') = 1$ if x is in contradiction with x', and $contr(x, x') = 0$ otherwise.

$$\textsc{Coh}(\mathbf{E}) = \frac{1}{N'} \sum_{\mathcal{P}' \subseteq \mathcal{P}} (\neg contr(\mathcal{P}', \hat{y}) \cdot \neg contr(\mathcal{P}', \mathbf{X})) \quad (1)$$

5.2. Deductive Evaluation

We define the following metrics for deductive explanations: *weak relevance*, *strong relevance*, and *redundancy*. We employ the properties presented in § 4.2 for deductive explanations to devise these metrics. We start with the metric $\textsc{Rel}_{\text{WEAK}}$ which is based on the weak relevance property. Let $\mathbf{E} = \langle \mathcal{P}, \mathcal{R} \rangle$ be a deductive explanation such that $\langle \mathcal{P}, \mathcal{R} \rangle \rightsquigarrow_d m(\mathbf{X}) \rightarrow \hat{y}$, with $N = |\mathcal{P}|$. Then satisfaction of weak relevance can be measured by:

$$\textsc{Rel}_{\text{WEAK}}(\mathbf{E}) = \frac{1}{N} \sum_{p_i \in \mathcal{P}} path(p_i, \hat{y}) \quad (2)$$

where $path(p_i, \hat{y})$ holds if there exists a path in $\mathcal{R}$ (seen as a graph) which connects p_i to $\hat{y}$. Note that $\textsc{Rel}_{\text{WEAK}} = 0$ ($\textsc{Rel}_{\text{WEAK}} = 1$) means that none (all, respectively) of the propositions in the explanation are relevant to the prediction.

For the related metric of strong relevance, we specify that there must be a relation, i.e. a direct connection between each proposition in the explanation and the prediction $\hat{y}$. This metric is defined as follows:

[2] Implementations of the evaluation metrics discussed in this section can be found here: https://github.com/neemakot/Evaluating-Explanations

$$\textsc{Rel}_{\textsf{STRONG}}(\mathbf{E}) = \frac{1}{N}\sum_{p_i \in \mathcal{P}} is\_relation(p_i, \hat{y}) \tag{3}$$

If all propositions are directly connected to $\hat{y}$, the value for this metric is 1, whereas if none of the propositions is directly connected to $\hat{y}$, then $\textsc{Rel}_{\textsf{STRONG}} = 0$.

The final metric that we provide for the evaluation of deductive explanations computes a score for the non-redundancy of an explanation. We define a non-redundant explanation in § 4.2 as one for which all propositions have relevance to the explanation in the context of the prediction. That is to say, if one of the propositions is omitted, the explanation would no longer be a sufficient justification for the model's prediction. The computation for the non-redundancy-derived metric is thus:

$$\textsc{Red}(\mathbf{E}) = 1 - \frac{1}{N}\sum_{p_i \in \mathcal{P}} (p_i \in \mathbf{X}) \cdot (p_i \in \mathcal{G}_{\mathcal{R}}) \tag{4}$$

The non-redundancy metric checks that two conditions have been met. First, we must ensure that each proposition in the explanation is derived from the inputs, and second, it must be the case that there is some relation that connects each proposition to all others (either when the direction of edges is considered or not), i.e. if viewing the deductive explanation as a graph it should consist of a single connected component $\mathcal{G}_{\mathcal{R}}$. The best possible score for redundancy is zero, indicating that the explanation contains no redundant components. A score greater than zero indicates at least one redundant proposition, if not more, exists in the explanation.

5.3. Argumentative Evaluation

We define metrics corresponding to two argumentative explanation properties: acceptability and dialectical non-circularity (see § 4.3). For simplicity, we focus on argumentative explanations of a restricted kind, namely corresponding to sets of trees of depth two at most (where the root is of depth zero). Let $\mathbf{E} = \langle \mathcal{A}, \mathcal{R}_{\textsf{Sup}}, \mathcal{R}_{\textsf{Att}} \rangle$ be an argumentative explanation such that $\langle \mathcal{A}, \mathcal{R}_{\textsf{Sup}}, \mathcal{R}_{\textsf{Att}} \rangle \rightsquigarrow_a m(\mathbf{X}) \rightarrow \hat{y}$, and let $N = |\mathcal{A}|$. Then, the satisfaction of acceptability can be measured by

$$\textsc{Acc}(\mathbf{E}) = \frac{1}{N}\sum_{a_i = (\mathcal{P}, \hat{y}) \in \mathcal{A}} \Big(\frac{1}{|Atts(a_i)|}\sum_{(a_j, a_i) \in \mathcal{R}_{\textsf{Att}}} \delta(a_j)\Big) \tag{5}$$

where $Atts(a_i) = \{a_j \mid (a_j, a_i) \in \mathcal{R}_{\textsf{Att}}$ and $\delta(a_j)) = 1$ if there exists $(a_k, a_j) \in \mathcal{R}_{\textsf{Att}}$, and $\delta(a_j) = 0$ otherwise. If $\hat{y}$ is predicted with top or high confidence we expect $\textsc{Acc}(\mathbf{E}) = 1$ for the explanation to be acceptable. Instead, if $\hat{y}$ is predicted with bottom or low confidence, we expect $\textsc{Acc}(\mathbf{E}) \neq 1$.

The second metric which we devise for evaluating argumentative explanations is related to the property of dialectical non-circularity. We define this metric as follows:

$$\textsc{Cir}(\mathbf{E}) = \frac{1}{N}\sum_{a \in \mathcal{A}} \frac{1}{M} \sum_{\mathcal{A}' \in \mathcal{R}_{\textsf{Sup}}, \mathcal{R}_{\textsf{Att}}} head(\mathcal{A}', a) \cdot tail(\mathcal{A}', a) \tag{6}$$

Here, we check for each argument in the explanation if there exists an attack or support relation in the explanation such that the head and tail arguments in the relation are the same arguments, i.e. the argumentation framework is circular. For the circularity measure, a favorable explanation would have a low score, i.e. fewer circular arguments.

6. Conclusion

We have identified and defined three rationale-derived explanation classes, drawing on illustrations from the automated fact-checking task in NLP. We also offered several desirable properties, both generic and structure-specific, for these explanations. Finally, we provided some quantitative measures for explanation evaluation.

We understand that devising a framework that assumes that explanations modeled in the spirit of human reasoning will have some limitations, e.g. sociocultural differences in a population (one example being generational differences) may mean that an explanation that can be well understood by one population may not be as well received by another. In understanding this, our framework is modular and customizable, meaning it is flexible and can accommodate culturally and linguistically dependent preferences for explanations.

We would be interested to see the application of our metrics across a range of NLP tasks. Furthermore, there is also scope for expanding these metrics, either to account for further properties or to account for explanation structure at a finer level of granularity. Overall, we believe this work will help guide further research in explainable NLP and explanation evaluation.

7. Acknowledgments

We thank the anonymous reviewers for useful comments. F. Toni was partially funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101020934), by J.P. Morgan and the Royal Academy of Engineering under the Research Chairs and Senior Research Fellowships scheme, by the Royal Society, UK (IEC\R2\ 222045 - International Exchanges 2022) and by the UKRI INDICATE project.

8. Ethics Statement

In this paper, we present a framework for the evaluation of rationalizing explanations in the context of fact verification. We do not present empirical results, and, for that reason, we do not believe that serious ethical considerations arise from this work. However, we believe that this work presents a significant contribution towards improved AI ethics because explanations, and in particular means for assessing the quality of varied explanations, allow for greater model transparency and accountability.

9. Bibliographical References

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. 2023. Faithfulness tests for natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 283–294, Toronto, Canada. Association for Computational Linguistics.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020a. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Pepa Atanasova, Dustin Wright, and Isabelle Augenstein. 2020b. Generating label cohesive and well-formed adversarial claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3168–3177, Online. Association for Computational Linguistics.

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

Eric Barnes. 1994. Why p rather than q? the curiosities of fact and foil. *Philosophical Studies*, 73(1):35–53.

Pietro Baroni, Antonio Rago, and Francesca Toni. 2019. From fine-grained properties to broad principles for gradual argumentation: A principled spectrum. *International Journal of Approximate Reasoning*, 105:252–286.

Yonatan Belinkov, Sebastian Gehrmann, and Ellie Pavlick. 2020. Interpretability and analysis in neural NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 1–5, Online. Association for Computational Linguistics.

Floris Bex, Henry Prakken, Chris Reed, and Douglas Walton. 2003. Towards a formal account of reasoning about evidence: argumentation schemes and generalisations. *Artificial Intelligence and Law*, 11(2):125–165.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Claudette Cayrol and Marie-Christine Lagasquie-Schiex. 2005. On the acceptability of arguments in bipolar argumentation frameworks. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pages 378–389. Springer.

Qianglong Chen, Feng Ji, Xiangji Zeng, Feng-Lin Li, Ji Zhang, Haiqing Chen, and Yin Zhang. 2021. KACE: Generating knowledge aware contrastive explanations for natural language inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2516–2527, Online. Association for Computational Linguistics.

George Chrysostomou and Nikolaos Aletras. 2021. Improving the faithfulness of attention-based explanations with task-specific information for text classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 477–488, Online. Association for Computational Linguistics.

George Chrysostomou and Nikolaos Aletras. 2022. Flexible instance-specific rationalization of nlp models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10545–10553.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pages 177–190. Springer.

Anupam Datta, Matt Fredrikson, Klas Leino, Kaiji Lu, Shayak Sen, and Zifan Wang. 2021. Machine learning explainability and robustness: Connected at the hip. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, page 4035–4036, New York, NY, USA. Association for Computing Machinery.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76, Vancouver, Canada. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Gabriel Deza, Adelin Travers, Colin Rowat, and Nicolas Papernot. 2021. Interpretability in safety-critical financialtrading systems.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

John Dougrez-Lewis, Elena Kochkina, Miguel Arana-Catania, Maria Liakata, and Yulan He. 2022. PHEMEPlus: Enriching social media rumour verification with external evidence. In *Proceedings of the Fifth Fact Extraction and VERification Workshop (FEVER)*, pages 49–58, Dublin, Ireland. Association for Computational Linguistics.

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

Prashant Gohel, Priyanka Singh, and Manoranjan Mohanty. 2021. Explainable ai: current status and future directions.

Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Riccardo Guidotti. 2022. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, pages 1–55.

Naoya Inoue, Pontus Stenetorp, and Kentaro Inui. 2020. R4C: A benchmark for evaluating RC systems to get the right answer for the right reason. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6740–6750, Online. Association for Computational Linguistics.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Neema Kotonya and Francesca Toni. 2019. Gradual argumentation evaluation for stance aggregation in automated fake news detection. In *Proceedings of the 6th Workshop on Argument Mining*, pages 156–166, Florence, Italy. Association for Computational Linguistics.

Neema Kotonya and Francesca Toni. 2020a. Explainable automated fact-checking: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5430–5443, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Neema Kotonya and Francesca Toni. 2020b. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

Robert A Kowalski and Francesca Toni. 1996. Abstract argumentation. In *Logical Models of Legal Argumentation*, pages 119–140. Springer.

Amrith Krishna, Sebastian Riedel, and Andreas Vlachos. 2022. ProoFVer: Natural logic theorem proving for fact verification. *Transactions of the Association for Computational Linguistics*, 10:1013–1030.

Sawan Kumar and Partha Talukdar. 2020. NILE : Natural language inference with faithful natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8730–8742, Online. Association for Computational Linguistics.

Andrew Lampinen, Ishita Dasgupta, Stephanie Chan, Kory Mathewson, Mh Tessler, Antonia Creswell, James McClelland, Jane Wang, and Felix Hill. 2022a. Can language models learn from explanations in context? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 537–563, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Andrew K Lampinen, Nicholas Roy, Ishita Dasgupta, Stephanie Cy Chan, Allison Tam, James Mcclelland, Chen Yan, Adam Santoro, Neil C Rabinowitz, Jane Wang, and Felix Hill. 2022b. Tell me why! Explanations support learning relational and causal structure. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 11868–11890. PMLR.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016. Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691, San Diego, California. Association for Computational Linguistics.

Andreas Madsen, Siva Reddy, and Sarath Chandar. 2021. Post-hoc interpretability for neural nlp: A survey. *arXiv preprint arXiv:2108.04840*.

Ana Marasović, Iz Beltagy, Doug Downey, and Matthew E Peters. 2021. Few-shot self-rationalization with natural language prompts. *arXiv preprint arXiv:2111.08284*.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.*, 267:1–38.

Tim Miller. 2021. Contrastive explanation: a structural-model approach. *Knowl. Eng. Rev.*, 36:e14.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. Wt5?! training text-to-text models to explain their predictions. *CoRR*, abs/2004.14546.

Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. 2023. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Comput. Surv.*, 55(13s).

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240, Melbourne, Australia. Association for Computational Linguistics.

Henry Prakken. 2005. A study of accrual of arguments, with applications to evidential reasoning. In *Proceedings of the 10th international conference on Artificial intelligence and law*, pages 85–94.

Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton. 2020. Learning to deceive with attention-based explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4782–4793, Online. Association for Computational Linguistics.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Ashish Rana, Deepanshu Khanna, Tirthankar Ghosal, Muskaan Singh, Harpreet Singh, and Prashant Singh Rana. 2022. Rerrfact: Reduced evidence retrieval representations for scientific claim verification. In *Proceedings of the Workshop on Scientific Document Understanding co-located with 36th AAAI Conference on Artificial Inteligence, SDU@AAAI 2022, Virtual Event, March 1, 2022*, volume 3164 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Olesya Razuvayevskaya and Simone Teufel. 2017. Finding enthymemes in real-world texts: A feasibility study. *Argument & computation*, 8(2):113–129.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Alexis Ross, Ana Marasović, and Matthew Peters. 2021. Explaining NLP models via minimal contrastive editing (MiCE). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3840–3852, Online. Association for Computational Linguistics.

Swarnadeep Saha, Prateek Yadav, Lisa Bauer, and Mohit Bansal. 2021. ExplaGraphs: An explanation graph generation task for structured commonsense reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7716–7740, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

Jiasheng Si, Yingjie Zhu, and Deyu Zhou. 2023. Exploring faithful rationale for multi-hop fact verification via salience-aware graph learning. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 13573–13581. AAAI Press.

Liwei Song, Xinwei Yu, Hsuan-Tung Peng, and Karthik Narasimhan. 2021. Universal adversarial attacks with natural triggers for text classification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3724–3733, Online. Association for Computational Linguistics.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. 2021. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3621–3634, Online. Association for Computational Linguistics.

Mokanarangan Thayaparan, Marco Valentino, and André Freitas. 2021. Explainable inference over grounding-abstract chains for science questions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1–12, Online. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022a. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Sarah Wiegreffe and Ana Marasovic. 2021. Teach me to explain: A review of datasets for explainable NLP. *CoRR*, abs/2102.12060.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.